Étude lexicale de textes arabes pour l'évaluation automatique de la complexité textuelle

MORSI Youcef Ihab,
BORDET Yves,
ATANASSOVA Iana
Centre de Recherche Lucien Tesnière
Université de Franche-Comté
Besançon, France

Résumé

Nous proposons une méthode pour évaluer la complexité du vocabulaire des textes en langue arabe en utilisant les résultats d'une analyse lexicale d'un corpus littéraire composé de 40 textes. Cette approche a été proposée par Yves Bordet dans le projet Doxilog appliqué à plusieurs langues, dont le français, l'anglais, l'espagnol, le russe et le chinois. La présence du vocabulaire arabe pérenne dans un texte en arabe est un élément clé permettant de déterminer la complexité du texte et notamment l'âge et le niveau d'études pour les apprenants natifs et le niveau de maîtrise de la langue pour les apprenants de langues étrangères.

الملخص

نقترح طريقة لتقييم مدى تعقيد مفردات النصوص العربية باستخدام نتائج تحليل المفردات من مجموع نصوص أدبية. تلعب النصوص الأدبية دورا هاما في اكتساب المفردات عند تعلم اللغة الأم أو لغة أجنبية أخرى. ترتكز طريقتنا على تحديد المفردات الدائمة الاستعمال فى اللغة العربية بالاستعانة بمجموعة من النصوص الأدبية متكونة من 40 نصا. اقترحت هذه المنهجية من قبل إيف بورديه في مشروع دوكسيلوق لتقييم مدى تعقيد النصوص وتم تطبيقها على عدة لغات، بما فيها الفرنسية والانكليزية والاسبانية والروسية والصينية. يعتبر وجود المفردات الدائمة الاستعمال فى نص عربى عنصر أساسى لتحديد مدى تعقيد النص , وخاصة سن و مستوى الدراسة للمتدربين الأصليين و مستوى إتقان اللغة للمتدربين الأجانب.

1 Introduction

Notre objectif est de proposer une méthode pour évaluer la complexité textuelle des textes arabes en nous basant sur l'analyse lexicale d'un corpus de textes littéraires. Il s'agit ici de définir le vocabulaire pérenne de la langue arabe, afin de

pouvoir proposer des critères pour l'évaluation de la complexité textuelle.

Ces dernières années, nombre de recherches ont été effectuées en matière d'études du lexique mais restent non utilisées dans l'enseignement et l'apprentissage des langues. Boulton (1998) parle d'un manque au niveau des approches basées sur une théorie de l'apprentissage du lexique. Il fait appel à la notion de « lexique mental ».

La notion de complexité textuelle a ses sources en philosophie, où la question consistait à savoir si certaines pensées ou idées sont en elles-mêmes complexes, si elles prennent une tournure complexe lors de leur expression ou s'il existe un lien entre les deux.Plus tard, les recherches en lisibilité se réapproprient le terme de complexité pour l'appliquer au domaine de la psycholinguistique. Des facteurs organisationnels comme la densité des idées et des concepts ou la structure et la présentation d'un texte sont au centre de ces études. Ce changement de cap dans les sciences cognitives est en partie lié à l'émergence de la linguistique textuelle. Comme le définit Barbaresi (2011), «L'analyse de la complexité se situe dans le cadre de l'assistance à la compréhension. Il s'agit ici de déterminer la lisibilité d'un texte pour des humains ou pour des machines, c'est-à-dire d'une part le niveau de maîtrise et de pratique de la langue requis et d'autre part le modèle formel et les instruments à utiliser. »

1.1 La notion de lisibilité

Les travaux sur la lisibilité textuelle consistent à appréhender les textes sous forme de variables linguistiques représentées en indices de difficulté. Les théories classiques s'appuient, essentiellement, sur des critères de types lexical et syntaxique.

Un mot connu par le lecteur se lit plus facilement qu'un mot jamais rencontré, un mot court se lit plus facilement qu'un mot long, une phrase complexe est plus difficile à lire qu'une phrase simple, un mot concret est plus facile à comprendre qu'un mot abstrait, etc. La lisibilité est souvent confondue avec l'intelligibilité et la compréhensibilité. Les formules de lisibilité, selon Sorin (1996), isolent le texte de son contexte de production et ne reflètent pas certains facteurs spécifiques tels que la motivation et l'intention de lecture. « La lisibilité irait de pair avec la cohésion, c'est-à-dire l'aspect linguistique du texte, alors que la compréhensibilité serait directement liée aux comportements psychologiques du lecteur et à la cohérence textuelle » (Sorin, 1996).

Fernbach (1990) définit la lisibilité comme étant « l'aptitude d'un texte à être lu rapidement, compris aisément et bien mémorisé ».En somme, comme l'indique Richaudeau (1969), un texte efficace est un texte qui permet une lecture. C'est ce facteur d'efficacité qui est mesuré par sa lisibilité.

Les premières méthodes et techniques de mesure de la lisibilité ont vu le jour avec les travaux de Lively et Pressey (1923). Ces travaux ont donné naissance à un grand nombre de formules basées exclusivement sur des indices lexicaux et syntaxiques (Flesch, 1948 ; Chall & Dale, 1995), ainsi que des variables typographiques. Sorin (1996) a recensé près de 200 formules de lisibilité en usage. La formule de Rudolf Flesch (Conquet, 1973), appelée aussi le Langage Simple, est jusqu'à aujourd'hui la plus utilisée. Elle se base sur une étude textuelle se reposant sur des critères de «facilité de lecture» et d'«intérêt humain». La « facilité de lecture » se calcule sur un échantillon de 100 mots en prenant le nombre de syllabes (W_1) et la moyenne des mots dans une phrase (S_1), avec la formule suivante :Facilité de lecture = $206,84 – 0,85W_1 – 1,02S_1$.

L' « intérêt humain » mesure ce qu'appelle Conquet la chaleur communicative produite par l'œuvre et se calcule sur un échantillon de 100 mots selon la formule suivante : Intérêt humain = $3,64W_2 + 0,32S_2$.La variable W_2 est le pourcentage des « mots personnels » qui sont les pronoms ou les mots du type « gens, peuple, maman ». La variable S_2 représente le pourcentage des phrases dites personnelles, relevant du style direct et s'adressant au lecteur.

La formule de Flesch (Conquet, 1973) reste une référence chez les chercheurs en complexité textuelle en ce qu'elle redéfinit le rapport entre l'écrivain et le lecteur. Les phrases courtes et les mots brefs apporteraient du rythme, les mots personnels et le style direct interrogent le lecteur et facilitent la transmission du message.

La formule de Dale-Chall (Conquet, 1973), quant à elle, se base sur la longueur moyenne des phrases et le pourcentage de mots difficiles appartenant à une liste type de 3 000 mots préétablie. Ces méthodologies quantitatives prennent comme outils des formules mathématiques pour calculer la fréquence et la disposition des mots ou encore la longueur des phrases. D'autres formules utilisent comme critères les syllabes.

Kandel et Moles (1958) et Landsheere (1963) ont adapté la formule de Flesch pour le français mais sans prendre en considération les spécificités linguistiques de la langue française. La première formule dédiée au français est celle d'Henry (1975), qui utilise un certain nombre de variables linguistiques. Cependant, les travaux en lisibilité sont restés en France peu développés.

1.2 La littérature dans l'enseignement des langues

Considérée comme un corpus idéal réunissant trois objectifs formatifs importants : esthétique, intellectuel et moral, la

littérature occupait une place primordiale dans les méthodologies traditionnelles. Elle était considérée comme le support d'une édification morale dans les années 1910-1920, le miroir de la société en 1930-1940 et enfin un support didactique dans les années 1950, notamment, avec les méthodes audio-orales. Vers les années 1960, la littérature se voit relégué au second plan dans l'enseignement des langues vivantes avec la méthodologie Structuro-Globale-Audio-visuelle. Avec l'arrivée de l'approche communicative en 1980, les textes littéraires regagnent leur place dans l'enseignement des langues. Cependant, ce retour reste assez « timide » car les textes littéraires font partie d'un ensemble englobant les textes authentiques. Ces dernières années, l'importance de l'oral dans l'apprentissage a, d'une certaine façon, oblitéré le rôle de la littérature.

1.3 La notion de langue pérenne

Yvon Bordet (2009) définit la langue pérenne comme étant « une langue utilisée dans un espace géographique déterminé dans lequel elle est officiellement reconnue. Sa grammaire, son dictionnaire et son orthographe sont définis avec précision. La littérature antérieure à cette définition est transcrite dans sa nouvelle version afin de faire pleinement partie de la langue pérenne. La langue pérenne est liée à sa propre littérature de manière intrinsèque.»

Dans le projet Doxilog, Bordet (2009) développe une méthode d'analyse de corpus permettant d'obtenir la liste du vocabulaire du français fondamental pérenne. Le corpus est constitué de 40 textes littéraires, dont 20 en prose et 20 en vers. Ces textes sont choisis de façon à être représentatifs de la langue pérenne:

- Auteurs connus et reconnus dans tout l'espace francophone ;
- Œuvres traduits dans au moins une langue internationale;

- Textes étudiés dans le domaine éducatif.

Pour la langue française, les résultats démontrent que dans les textes scientifiques près de 84% des mots appartiennent à ce vocabulaire du français fondamental pérenne, 86% pour les articles de journaux, 90% pour les textes littéraires (Bordet, 2009).

Notre travail consiste à appliquer cette méthode àl'arabe, en considérant les spécificités de la langue arabe littéraire pourd'établir la liste du vocabulaire de l'arabepérenne (VAP).

1.4 De l'arabe à une langue pérenne

Dès lors que nous nous intéressons à la langue arabe, nous sommes confronté à étudier et évaluer la situation de Diglossie de cette langue dans tout l'espace arabophone. Ferguson (1959) définit la Diglossie en distinguant deux (ou plusieurs) systèmes linguistiques présents dans une communauté linguistique. Chaque système a son propre statut et remplit une fonction précise. Ces deux systèmes pour l'arabe, dans certains contextes, sont l'arabe littéraire (ou classique) et l'arabe dialectale.

Avant même l'arrivée de l'islam, les arabes parlaient des dialectes différents. Les poètes et les orateurs utilisaient une langue commune qui était comprise par tous.« Cette langue commune est très proche parente des dialectes mais elle se distingue d'eux par les finesses du vocabulaire, par les inflexions et les articulations de la syntaxe. » (Abd El Jalil, 1946)

Cette langue écrite a connu divers modification et s'est imposée avec l'avènement du texte sacré, le Coran.

De cet arabe, arabe littéraire ancien antérieur à l'islam, peu de documents écrits nous sont parvenus. Toutefois « quelques inscriptions anciennes nous sont connues, soit directement,

comme les graffiti relevés sur le mur du temple de Ramm dans le Sinaï (300 après J.-C.), une inscription chrétienne (accompagnant deux textes dédicatoires en grec et syriaque) à Zebed au sud-est d'Alep (512), une autre à Harran au sud-est de Damas (568) et un graffiti à Umm al-Djimal au sud de Bassora, soit indirectement, comme l'inscription de l'église de Hind à Hira (560) qui nous est rapportée par des historiens musulmans. » (Cohen, 2010)

Un grand répertoire littéraire est, cependant, connu en arabe littéraire ancien. Cette riche littérature de l'époque antéislamique à caractère oral reste une référence chez les poètes d'expression arabe.

C'est au XIXe siècle, à la suite de la conquête de Napoléon, que commence la période de la Renaissance qui donne naissance à l'arabe moderne. L'usage actuel de la langue est assez différent sur le plan syntaxique. L'influence étrangère en est la principale raison mais ne demeure pas la seule : les auteurs contemporains sont le plus souvent, en ce siècle moderne, bilingues. À cet arabe littéraire, s'ajoute l'arabe dialectal utilisé quotidiennement, en communication orale, et qui est assez souvent très loin de l'arabe écrit.

2 Méthodologie

Nous utilisons la méthode proposée par Yves Bordet dans le projet Doxilog pour établir la liste du vocabulaire pérenne. Notre objectif consiste à appliquer cette approche à la langue arabe. Des assimilations et des adaptations ont dû être établies afin d'adapter l'approche à la langue arabe, étant donné que le français et l'arabe sont des langues très différentes, même si elles sont toutes les deux alphabétiques.

Nos hypothèses sont les suivantes:

- le vocabulaire de l'arabe pérenne (désormais VAP) ainsi

établi constitue autour de 80% des mots dans un texte littéraire arabe ;
- la présence du vocabulaire de l'arabe pérenne dans un texte varie en fonction de l'accessibilité du texte.

A partir d'un corpus de textes littéraires en vers et en prose, nous considérons tous les lexèmes qui apparaissent dans ces textes. Pour qu'un mot soit considéré comme appartenant au VAP, une des deux conditions suivantes doit être satisfaite :

Condition 1 : Un mot (lexème) qui apparaît dans au moins deux textes est considéré comme pérenne et intégré dans la liste du VAP. Un mot qui apparaît plusieurs fois dans le même texte et pas dans d'autres textes du corpus n'est pas considéré comme étant pérenne.

Condition 2 : Certaines catégories spécifiques de mots appartiennent au vocabulaire pérenne et font exception à la règle ci-dessus. Nous appellerons ces mots *« mots pérennes spécifiques »*. Ils incluent : les chiffres arabes, les nombres ordinaux et cardinaux, les mois et les jours de la semaine, les pronoms personnels et les pronoms possessifs.

Les phénomènes de voyellisation sont pris en compte pour déterminer les lexèmes différents de la manière suivante. Les mots ayant la même racine mais pas la même classe morphosyntaxique dans les textes, sont considérés comme différents. Pour illustrer ce point, prenons l'exemple du mot *Kataba (*كتب*= a écrit)* qui apparaît dans le texte 1 et *koutoub (*كتب*= des livres)* dans le texte 2. Ces mots ont la même racine et apparaissent dans deux textes différents (condition 1 réalisée) mais n'ont pas la même voyellisation. Cette dernière représente leur unité distinctive. Les mots non voyellés compliquent le traitement du fait que leurs classes morphosyntaxiques doivent être déterminées à partir du contexte. C'est la raison pour laquelle nous avons opté pour une

analyse manuelle des textes pour la constitution de la liste du VAP.

Lors de l'analyse du corpus, deux catégories de mots sont à relever. La première catégorie est celle des mots pérennes qui regroupe les mots réalisant la première condition, à savoir mots qui apparaissent dans deux textes différents. La deuxième catégorie comprend les mots pérennes spécifiques, ainsi que les noms propres.

En langue arabe, les articles, les prépositions et les pronoms, ne sont pas toujours des mots séparés. Ces particules viennent s'ajouter aux adjectifs, noms, verbes pour former un seul et même mot. Ce phénomène d'agglutination, qui n'existe pas dans la langue française ou les autres langues latines, doit être pris en compte lors du traitement automatique de l'arabe.

L'agglutination engendre des ambiguïtés morphologiques, notamment dans les couples conjonction + pronom et conjonction + verbe.Ces spécificités de l'arabe nous ont conduit à faire des choix sur l'intégration de ces particules ou pas dans la liste du VAP.

2.1 nstitution du corpus

Nous avons constitué un corpus de 40 textes : 20 en prose et 20 en vers. Ce corpus a été conçu pour être représentatif de l'arabe littéraire : les textes couvrent plusieurs époques (Omeyyade, Abbasside, Contemporaine), et les auteurs sont originaires de différents pays (Irak, Égypte, Tunisie, Algérie, Liban, Arabie Saoudite, Syrie, Palestine, Ouzbékistan). Les tableaux 1 et 2 montrent les sources utilisées. Les textes sont d'environ 500 mots pour la prose et 100 mots pour la poésie.

Tableau 1 : Contenu du corpus en prose

Id	Texte (extraits)	Période[1]
1	Tahar Ben Jelloun « Cette aveuglante absence de lumière »(من رواية تلك العتمة الباهرة)	C
2	El Baghdadi (من كتاب خزانة الأدب)	M
3	Abdel Hamid El Kateb « Lettres de la littérature d'El Kateb »(رسالة أدب الكتّاب)	O
4	AhlemMostghanemi«Mémoires de la chair» (من رواية ذاكرة الجسد)	C
5	El Jahid«Les avares»(البخلاء)	A
6	Abdel Rahmane El Kawakibi«Traits de la répression et le combat contre l'esclavage»(الاستبداد و التخلُّص منه)	C
7	Mustapha LutfiI Al Manfalouti«Les larmes» (اَلعَبَرَات)	N
8	Mustapha Sadiq Al Rafi'i(وحي القلم)	N
9	GhassanKanfani«Une brève décision»(قرار موجز)	C
10	Ibn El Muqaffa«Clila y Dimna »(باب الحمامة المطوقة)	O
11	Ibn El Muqaffa« Mille et une nuits- le marchand et le génie »(حكاية التاجر مع العفريت)	O
12	Ibn Khaldoun«L'introduction»(المقدمة)	M
13	Elias Khoury«Récoltes (conséquences) dela tempête»(حصاد العاصفة)	C
14	Khaled Aouiss«Je suis un traître oh mon beau pays» (أنني خائن أيها الوطن الجميل)	C
15	Khalil Gibran«Le Prophète»(النبي)	C
16	Al Khawarizmi«Lettres de Khawarizmi» (من رسائل الخوارزمي)	A
17	May Ziade«La fraternité»(الإخاء)	C
18	MikailNaima«El Baydar»(المَذاهِبُ والمُتَمَذهِبُون البيادر)	N
19	TahaHussain«Les jours»(من كتاب الأيام)	N
20	WacinyLaredj(أنثى السراب)	C

[1]Périodes : A - Abbasside ; C – Contemporaine ; M – Mamelouke ; N – Nahda ; O - Omeyyade

Tableau 2: Contenu du corpus en vers

Id	Texte(extraits)	Période[2]
21	Abu Kacem El Chebbi«Hymnetitanesque » (نشيد الجبار)	C
22	Abu Nouwas«Diwane d'Abu Nouwas» (دَعْ عَنْكَ لَوْمي فإن اللّوْمَ إغْرَاءُ)	A
23	Abu Tammam«Diwane d'Abu Tammam» (سَيْفُ أَصدق إنْبَاءً من الكتب)	A
24	Ahmed Chawqi(بسامُكَ من سقراطَ في الخطب أَخْطَبُ)	C
25	BacharIbnBurd(جِهّزْ طلل في النّصَب النّواءُ)	A
26	AantaraIbnChadad«DiwaneAantaraIbnChadad» (يا صاحبي لا تَبْكِ رَبعاً قد خلا)	A
27	El Akhatal«Diwane el Akhtal» (ومحبوسةٍ في الحيّ ضامنةِ القرى)	O
28	El Buhturi«Diwane» (أنزاعاً في الحب بعد نزوع)	A
29	El Farazdaq«Diwane»(إذا كُفّت صلُ أُقْعوانٌ، فما لَهُ)	O
30	Abu Lala Al Maari(إذا كُفّت صلُ أُقْعوانٌ، فما لَهُ)	A
31	Elia Abu Madi «Vivons!»(فلنعش)	C
32	El Mutanabbi«Diwane»(ألا كل مَاشِيَةِ الخَيْزَلَى)	A
33	Hafiz Ibrahim «Oh mon maître et mon idole» (يا سيدي وإمامي)	C
34	Omar IbnAbi Rabia (حَدِّثْ حَديثَ فتاةِ حَيٍّ مرّة)	O
35	Jarir(حيوا أمامةَ واذكروا عهداً مضى)	O
36	MikailNaima«Mon frère»(أخي)	N
37	MoufdiZakaria«L'Iliade de l'Algérie»(إلياذة الجزائر)	C
38	Ibrahim Naji(أنـوار)	C
39	NazarKabani(هجم النفط مثل ذئب علينا)	C
40	Imrou'lQuays«Arrêtez vos montures, vous deux, et pleurons au souvenir d'un campement» (فا نبك من ذِكرى حبيب)	P

2.2 Méthodologie de l'analyse

Dans un premier temps, nous avons identifié les lexèmes appartenant au VAP à partir des textes du corpus, ainsi que les mots pérennes spécifiques. Cette première expérimentation

[2] Périodes : A - Abbasside ; C – Contemporaine ; N – Nahda ; O – Omeyyade ; P – Préislamique

nous a permis de constituer la liste du VAP, selon la méthodologie présentée plus haut.

Une première annotation a été faite manuellement sur le corpus, lors de laquelle les mots pérennes ont été notés par le symbole « degré » (°) , les mots pérennes spécifiques et les noms propres par un plus (+) et les mots non pérennes par une étoile (*). Un exemple d'un extrait annoté est présenté sur la figure 1.

Exemple du texte N° 22 : Abu Nouwas, دَعْ عَنْكَ لَوْمي فإن اللَّوْمَ إغْرَاءُ (Poésie)

دع° عنك° لَومي* فإن° اللَّوْمَ* إغْرَاءُ* و° دَاوني* بالتي° كانت° هي°

الدَاءُ* صفراء * لا° تنزل° لأحزان° سَاحَتها* لو° مَسَها* حجر° مَسَّتْهُ* سَوَاءُ* مِنْ° كف° ذات°

حِرٍ* في° زِيّ* ذي° ذكر° لها° مُحِبّانِ* لُوطيٌّ* وَزَنَاءُ* قامت° بِإبْريقِها* و° الليل° مُعْتَكِرْ *

فَلاحَ* من° وجهها° في° البيت° لألاءُ* فأرسلَتْ* من° فَم* الإبْريق* صافيَة* كأنما° أخذها °

بالعين° إغفاءُ* رقَّتْ* عَنِ° الماء° حتى° ما° يلائمُها* لَطافَة* وَجَفا* عَنْ° شكلها° الماءُ°

فلو° مَزَجْتَ* بها° نوا° لَمَازَجَها* حتى° تَوَلدَ* أنوار° و° أضواء° دارت° على° فِتْيَةٍ* دانَّ *

الزمان° لهم° فَما° يصيبهم° إلا° بما° شاؤوا* لتِلكَ° أبكي° و ° لا° أبكي° لمنزلة* كانت °

تحل* بها° هندٌ+ و° أسماءُ+ ااا

Figure 1. Exemple d'un texte annoté

Les résultats de cette annotation ont servi comme entrée d'un module de traitement automatique qui a permis de créer les listes des trois différentes catégories de mots. Ces listes contiennent les formes fléchies.Elles ont été traitées manuellement afin d'engendrer la forme de base de chaque mot. Des vérifications contextuelles ont été faites pour déterminer la catégorie des mots de la liste. Cette étape est très importante car elle nous a permis de déterminer par exemple si :

- le mot : (ذهب) est un verbe et signifie « est parti » ou un substantif et signifie « or ».

- le mot : (علم) est un verbe et signifie « a su » ou un substantif et signifie « savoir/ science ».

Lors de cette étape, les verbes ont été présentés sous leur forme de base trilitère pour la plupart d'entre eux, et les substantifs avec le déterminant (الـ) . Cette distinction morphologique est très importante lorsque le mot est isolé de son contexte.

3 Résultats de l'analyse

En considérant tous les lexèmes qui apparaissent dans au moins deux textes différents du corpus, nous avons obtenu la liste du VAP. Ce vocabulaire contient autour de 1280lexèmes différents. La taille de la liste obtenue est comparable avec les tailles des listes pour les autres langues alphabétiques traitées dans le projet Doxilog : français (Bordet et Bouhadiba, 2014), anglais, espagnol (Bordet, 2015), russe.

Notre analyse a abouti sur les résultats présentés dans le tableau 3.

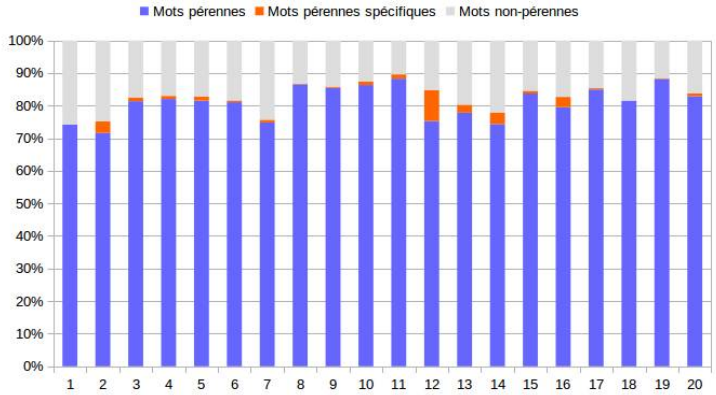
Tableau 3: Occurrences de mots pérennes dans le corpus

ID	Nb de mots	Occurrences de mots pérennes	Occurrences de mots non pérennes	Occurrences de mots pérennes spécifiques	Pourcentage des mots pérennes
Corpus en prose					
1	458	340	118	0	74,24%
2	419	300	104	15	75,18%
3	628	511	110	7	82,48%
4	529	434	90	5	82,99%
5	612	499	105	8	82,84%
6	570	462	105	3	81,58%
7	612	458	149	5	75,65%
8	565	489	75	1	86,73%
9	536	458	76	2	85,82%
10	534	461	67	6	87,45%
11	683	603	71	9	89,60%
12	736	554	112	70	84,78%

13	609	474	120	15	80,30%
14	760	565	168	27	77,89%
15	719	602	111	6	84,56%
16	381	303	66	12	82,68%
17	505	429	74	2	85,35%
18	586	478	108	0	81,57%
19	557	491	65	1	88,33%
20	689	571	112	6	83,74%
Corpus en vers					
21	111	78	33	0	70,27%
22	101	67	32	2	68,32%
23	101	72	29	0	71,29%
24	113	78	32	3	71,68%
25	120	78	40	2	66,67%
26	104	73	30	1	71,15%
27	99	58	41	0	58,59%
28	88	54	34	0	61,36%
29	102	59	38	5	62,75%
30	108	84	23	1	78,70%
31	128	99	29	0	77,34%
32	96	61	34	1	64,58%
33	81	56	24	1	70,37%
34	91	54	37	0	59,34%
35	140	90	49	1	65,00%
36	100	68	32	0	68,00%
37	99	69	28	2	71,72%
38	115	83	32	0	72,17%
39	102	66	30	6	70,59%
40	135	83	51	1	62,22%

Nous observons que dans la grande majorité des textes en prose (textes 1-20) le pourcentage de mots pérennes varie entre 75% et 89% (voir figure 2).

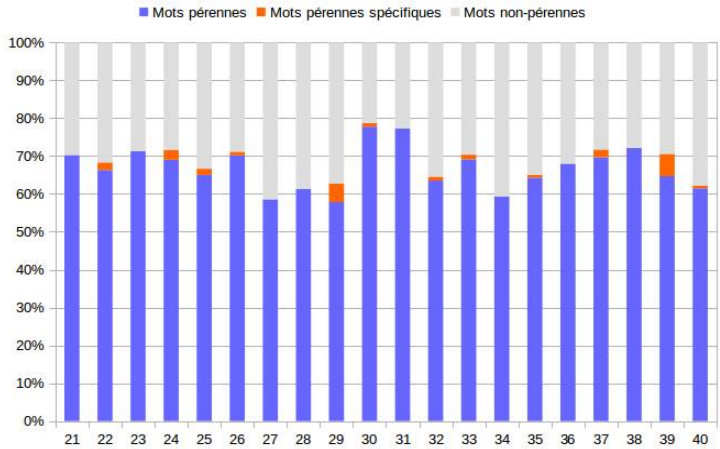
Figure 2 : Pourcentage de mots pérennes : corpus en prose



Ce pourcentage est beaucoup moindre pour les textes en vers (textes 21-40) : entre 58% et 78%, comme le montre la figure 3.

Le fait que le pourcentage des mots pérennes est moindre dans les textes en vers est dû à la spécificité de la poésie arabe. La poésie arabe vise une écriture recherchée, où l'auteur choisitles mots en prenant en considération la métrique des syllabes et le rythme des phrases. L'auteur s'appuie sur des éléments esthétiques, comme les accents ou les pauses, ou encore fait un travail sur les rimes pour toucher la sensibilité de ses lecteurs. En prose, la longueur des phrases ou le choix des mots dépendent moins de leur sonorité mais beaucoup plus de la fidélité des idées auxquels ils s'attachent. La description, l'illustration et l'argumentation sont par exemple des outils couramment utilisés en prose, ce qui favorise l'utilisation de mots fréquents dans la langue.

Figure 3: Pourcentage de mots pérennes dans les textes en vers



La langue arabe est une langue agglutinante. Les mots grammaticaux sont le plus souvent « collés » à la racine des noms et des verbes.

En langue française, près de 50 % des mots dans un texte sont de type grammatical, comme observé par Yves Bordet, et les pourcentages des mots pérennes observés pour le français sont en moyenne 95% pour les textes en prose et 94% pour les textes en vers. Les résultats que nous avons obtenus nous laissent croire que le pourcentage des mots pérennes dans les textes littéraires arabes ne dépasserait pas les 90 %.

Conclusion

L'objectif de notre recherche étaitde mettre en place la liste du vocabulaire de l'arabe pérenne pour définir des critères de mesure de la complexité des textes littéraires arabes. Nous avons établi un corpus de 40 textes littéraires en arabe. Ce corpus comprend des textes qui ont traversés différentes

époques et étudiés dans différents pays arabophones. Nous avons utilisé la méthodologie développée par le Dr Yves BORDET pour analyser la langue arabe et établir la liste du vocabulaire de l'arabe pérenne.

Nos résultats montrent que l'arabe possède un vocabulaire pérenne qui se compose d'une liste d'autour de 1280lexèmes. Le phénomène d'agglutination de la langue arabe doit être pris en compte lors de l'analyse lexicale du corpus. D'autres caractéristiques de l'arabe, par exemple la vocalisation des mots, jouent un rôle important dans le traitement automatique de cette langue.

L'analyse du corpus a montré que la présence du vocabulaire de l'arabe pérenne dans les textes littéraires analysés est autour de 80% pour les textes en prose et autour de 70% pour les textes en vers.

La prochaine étape de ce travail consistera en la mise en place de l'algorithme d'analyse des textes arabes afin d'identifier le pourcentage des mots pérennes dans un texte. A partir de ces données, des indicateurs seront définis afin de déterminer l'âge et le niveau équivalent au CECR auxquels le vocabulaire d'un texte est accessible pour les apprenants. Ce travail nécessite, entre autres, d'identifier les formes de base des mots dans un texte afin de les comparer aux mots de la liste du vocabulaire de l'arabe pérenne. Ceci peut être fait de deux manières : par la mise en place d'analyses morphosyntaxiques des textes en arabe, ou bien par la génération de l'ensemble des formes fléchies des lexèmes de la liste. Dans les deux cas, vu la forte polysémie des formes lexicales en arabe, il est nécessaire d'examiner le contexte de chaque occurrence afin de la désambiguïser.

Références

ABD-EL-JALIL, J-M. (1947) *Brève histoire de la littérature arabe*. Paris. Maisonneuve.

ALOULOU, C. (2003) Analyse syntaxique de l'arabe: Le système MASPAR. In Actes RECITAL-2003, pp. 419-429. Batz-sur-Mer.

BARBARESI, A. (2011) *La complexité linguistique : méthode d'analyse*. In TALN 2011. N°2, pp. 229-234. Montpellier.

BORDET, Y. (2015) *An Example of the "Cultural Approach" of Language: The Doxilog Project*. VII Conferencia CientíficaInternacional. Universidad de Holguín. Cuba. ISBN 978-959-16-2472-7.

BORDET, Y. et BOUHADIBA, F. (2014) *Noulisons : un logiciel référencier*, Revue Maghrébine des Langues. Université d'Oran. Algérie. No 9, pp. 13-31.

BORDET, Y. (2009) *Français littéraire et français fondamental, une étude lexicale*. Thèse de doctorat. Université de Franche Comté, Besançon.

BOULTON, A. (1998) *L'acquisition du lexique en langue étrangère*. In actes du 26ème congrès d'UPLEGESS, pp. 77-87.

COHEN, D. (2010) *ARABE (MONDE) Langue*. *Encyclopædia Universalis*.

CONQUET, A. et RICHAUDEAU, F. (1973) *Cinq méthodes de mesure de la lisibilité*. Communication et langages. N°17.

CONQUET, A. (1972) *La Lisibilité et inspiration*. Nous, gens de la Bible. Paris.

FERGUSON, C.A. (1959) *Diglossia*. Word. Vol.15.

FERNBACH, N. (1990) *La lisibilité dans la rédaction juridique au Québec*. Centre canadien d'information juridique. Ottawa.

HENRY, G.(1975) *Comment mesurer la lisibilité*. Bruxelles. Labor.

LABASSE, B. (1999) *La lisibilité rédactionnelle : fondements et perspectives*. Communication et Langages. N° 121, pp. 86-103.

PRÉFONTAINE, C. et LECAVALIER, J. (1996) *Analyse de l'intelligibilité des textes prescriptifs*. Revue québécoise de linguistique. vol. 25 .n° 1, pp. 99-144.

RICHAUDEAU, F. (1969) *La Lisibilité*. C.E.P.L.-Denoël : Paris.

SORIN, N. (1996) *De la lisibilité linguistique à une lisibilité sémiotique*. Revue québécoise de Linguistique. vol. 25. n° 1, pp. 61-98.