Vers un modèle pour le recueil d'un corpus d'apprentissage d'une langue étrangère peu dotée de ressources.

Violetta Cavalli-Sforza et Mariam El Mezouar
Ecole de Science et Génie Université Al Akhawayn
Ifrane, Maroc

1. Introduction et motivation

L'enseignement et l'apprentissage d'une langue étrangère visent normalement, avec des variations selon la langue et les buts de l'apprenant, quatre compétences principales : lire, écrire, parler, écouter. Dans le cadre de l'apprentissage des langues assisté par ordinateur, la lecture des textes joue un rôle important pour deux raisons. Du côté pratique les systèmes de traitement automatique de la langue sont actuellement bien équipés dans ce sens (beaucoup plus que pour le traitement de la parole), et du côté pédagogique la lecture permet à l'apprenant de développer le vocabulaire de la langue cible et de comprendre les nuances des mots et leur bonne utilisation par rapport au contexte, ainsi que la relation entre la réalisation des mots et la structure syntaxique dans laquelle ils se trouvent.

Or, pour les langues qui ont été longuement enseignées en tant que langue étrangères—par exemple, l'anglais, le français et l'espagnol— il y a un grand choix de textes adaptés aux apprenants de différents niveaux, bien sûr avec certaines restrictions sur leur utilisation tels que les droits d'auteur. Par contre, puisque l'intérêt à enseigner une lange telle que l'arabe en tant que langue étrangère est relativement récent, dû aux bouleversements des évènements politiques et économiques globaux des dernières années, le choix de matériaux pédagogiques pour les apprenants est encore limité. De plus, même si cette langue a vu une large croissance sur le web pendant la dernière décennie, le web n'a qu'une variété étroite de

textes et pas toujours appropriée à l'apprenant, le genre infos et le genre blog étant les plus fréquents.

Malheureusement le premier offre un style et des sujets limités, tandis que le deuxième est souvent un mélange d'arabe standard et dialectal, ce qui est authentique mais peut poser un défi excessif pour l'apprenant et surtout ne lui présente pas un bon modèle de comment s'exprimer correctement par écrit. Maintenant si l'on passe de l'arabe à la langue amazighe, le choix de textes pour l'apprentissage devient encore plus limité du fait des différences régionales de cette langue et de son positionnement plus en tant que langue orale qu'écrite.

A présent nous sommes en train d'étudier le problème de comment collecter un corpus d'arabe standard pour l'apprenant de cette langue qui a un niveau moyen-bas, c'est-à-dire, qui possède les règles basiques de la morphologie et de la syntaxe mais qui a un vocabulaire encore assez étroit (autour de 2000 mots dont un millier supposément bien appris et les autres peut-être reconnus mais pas encore maitrisés) en se posant la question : qu'est-ce qui rend un texte approprié, c'est-à-dire lisible et compréhensible, pour un apprenant qui a certaines connaissances au niveau lexique et syntaxique ? Dans l'immédiat nous ciblons l'arabe à cause de notre expérience avec le traitement automatique de cette langue car, même si on peut la considérer une langue peu dotée, il y a quand même des ressources et des études précédentes pour l'arabe et d'autres langues sémitiques, particulièrement l'hébreu, qui soutiennent et encadrent respectivement notre recherche. Mais juste comme ces travaux existants nous ont fourni des indices, nous menons ce travail tout en pensant comment appliquer les résultats éventuels à l'amazighe, langue qui fait aussi partie de notre région et de nos intérêts.

La communication est organisée ainsi de la façon suivante. La première partie fournit le contexte pour notre travail. Nous commençons par une brève présentation d'une plateforme d'apprentissage, dans laquelle on va imbriquer ce travail de recueil de textes. Ensuite nous présentons un historique très succinct des études qui ont été faites sur la lisibilité des textes dans la langue d'enseignement pour l'arabe et d'autres langues[1]. Ceci sera suivi

d'une description des deux projets récents qui ont ciblé la lecture comme soutien à l'apprentissage d'une langue étrangère et qui ont directement inspiré et influencé ce travail. Une fois présenté le contexte du travail, nous passons à la description de la méthode suivie et des ressources utilisées avant de présenter les résultats, encore préliminaires, de notre investigation. Nous concluons avec quelques remarques sur le chemin ouvert devant nous pour l'arabe et comment cette expérience pourrait bénéficier à un recueil des textes pour l'apprenant de la langue amazighe.

2. Cadre du travail

2.1. Plate-forme d'apprentissage des langues

L'un des buts ultimes de ce travail est de développer un ensemble de critères, des lignes directrices, des outils de collecte et de sélectionner selon le besoin un ensemble de textes appropriés aux apprenants d'une seconde langue à différents niveaux de compétence dans une langue donnée (arabe, tamazight). Ces outils et les textes qui en résultent sont à utiliser dans une plate-forme d'apprentissage des langues qui, au moins au début, se concentre sur la lecture comme un élément essentiel de l'apprentissage. La plate-forme, qui pourrait être un accompagnement à un cours ou un outil indépendant d'étude, contient un programme de langue défini comme une séquence d'unités, chacune contenant des concepts linguistiques à acquérir avant de passer à l'unité suivante. Les concepts sont actuellement divisés en deux catégories générales: le vocabulaire et la grammaire (y compris les règles morphologiques et syntaxiques). Pour renforcer le contenu de chaque unité, les apprenants sont guidés par le système pour lire des textes qui contiennent les concepts linguistiques ciblés par l'unité. Des tests associés à chaque texte mesurent si l'apprenant a acquis les concepts souhaités. Selon les résultats des tests précédents, le système présente à l'apprenant des textes qui comprennent les concepts non encore assimilés et/ou sur lesquels l'apprenant n'a pas encore été évalué. En général, avant de passer à l'unité suivante, les apprenants devraient réussir tous les tests associés à l'unité sur laquelle ils ont travaillé.

Une version ambitieuse de la plate-forme qui vient d'être décrite est illustrée à la figure 1. Elle est encore en développement, mais plusieurs de ses composantes ont été largement complétées et sont ou seront bientôt testées.

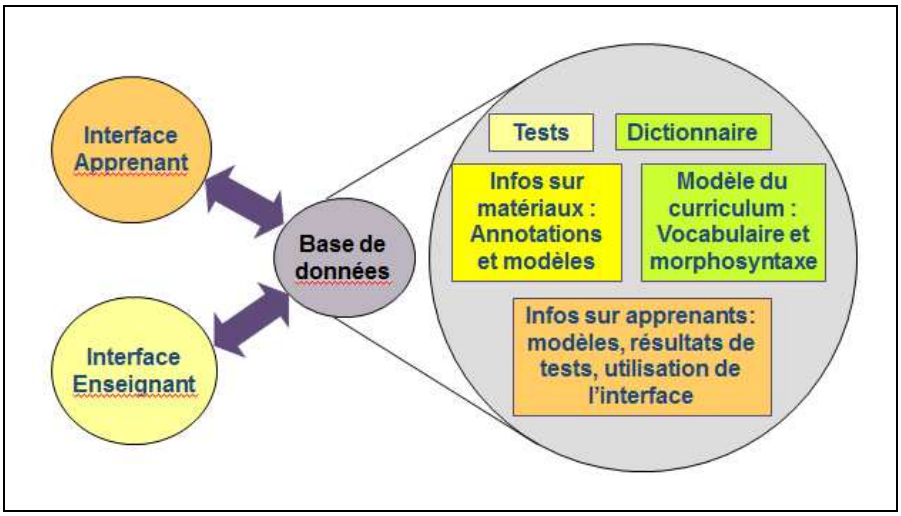

Figure 1: Schema de la plate-forme et son contenu

La plupart du contenu de la plate-forme d'apprentissage des langues est stocké dans une base de données MySQL. Son contenu est récupéré et rendu accessibles par les utilisateurs (enseignants ou étudiants) grâce à des interfaces spécialisées. Pour les enseignants : l'interface offre la possibilité de :

- Créer un programme d'études contenant les éléments de vocabulaire et de l'information morphologique et syntaxique;
- Ajouter, modifier ou supprimer des textes;
- Créer des tests individuels pour chaque texte et des tests unitaires ou modifier des tests existants.

Les étudiants auront, à travers leur interface, la possibilité de consulter les textes et les informations qui leurs sont associées, de prendre des notes sur ces derniers et de passer des tests.

Toutes ces fonctionnalités sont déjà développées et font actuellement l'objet des tests du système. Le travail qui reste à faire aujourd'hui concerne le modèle de l'apprenant, l'utilisation de matériaux autres que des textes (par exemple, audio ou vidéo ainsi que leur transcription sous forme de texte), l'intégration des concepts de syntaxe, en plus de concepts de morphologie (qui sont déjà présents mais non utilisé actuellement) et l'annotation des textes avec de tels concepts.

Le travail décrit dans le présent document a pour but de permettre à la plate-forme d'apprentissage d'enrichir (semi-)automatiquement la collection de textes au-delà de quelques-uns qui peuvent être ajoutés manuellement par un enseignant, afin de fournir un riche choix de textes de pratique et d'essais.

2.2. Lisibilité : Comment la définir

La caractérisation de la lisibilité d'un texte a attiré l'attention des pédagogues depuis un moment, surtout dans le monde anglophone. Par exemple, (Klare, 1963) l'a définie en termes d'une mesure de la facilité à comprendre due au style d'écriture, une définition qui se concentre sur les caractéristiques du texte. Par contre, (McLaughlin, 1969) met l'emphase sur le lecteur en basant la définition sur le degré d'attirance et compréhensibilité d'un texte pour une certaine classe de lecteur, tandis que (Dale & Chall, 1949) nous offrent une perspective compréhensive en définissant la lisibilité d'un texte comme l'ensemble d'éléments qui influencent le succès qu'un groupe de lecteurs a avec le texte – le niveau de compréhension, la rapidité de lecture et l'intérêt porté au texte. Le but ultime est d'établir la lisibilité d'un texte et finalement de déterminer s'il est approprié pour un certain groupe de lecteurs. Par le passé ce groupe consistait principalement en élèves des écoles primaires et secondaires mais plus récemment, avec la diffusion des informations d'intérêt général sur le web (la santé en est un exemple important), un public plus vaste et pas toujours très instruit est couramment concerné.

Comme l'*intérêt* du lecteur pour un thème n'était pas aussi facile à déterminer qu'aujourd'hui avec les nouvelles technologies adaptatives, ni ne constituait nécessairement un critère important pour les créateurs des curriculums scolaires, la plupart des études de lisibilité ont caractérisé le texte sur la base de critères tels que:

- Les **traits lexicaux**, par exemple la longueur des mots, leur fréquence dans l'utilisation de la langue, la charge de vocabulaire d'un texte, la présence des mots monosémiques plutôt que polysémiques ;
- Les **traits de la phrase**, par exemple la présence des propositions explicatives ou entre parenthèses, la longueur moyenne de la phrase, la complexité de la structure grammaticale de la phrase, la densité des propositions, la présence des métaphores et comparaisons ;
- la **familiarité avec le sujet traité par le texte**.

Sur la base d'un sous-ensemble de traits mentionnés ci-dessus, différentes formules (à peu près 200 formules existent couramment) ont été développées dont le but est soit de calculer une note représentant la difficulté de compréhension d'un texte (souvent sur une échelle de 0 à 100) soit de prédire le niveau scolaire auquel un étudiant sera capable de comprendre l'écrit. Ces formules ont ciblé surtout l'anglais et les niveaux scolaires américains, mais aussi certaines langues latines et germaniques, le chinois (Pang, 2006), le japonais (Tateisi, 1988), et l'hébreu (Ben-Simon et Cohen, 2011), parmi d'autres, l'hébreu étant assez proche de l'arabe et dans la même famille des langues sémitiques. Pour l'arabe on trouve très peu d'études sur le sujet de la lisibilité des textes. Les travaux que nous avons trouvés sont brièvement décrits ci-dessous.

Il faut aussi mentionner que, au-delà des formules statistiques, il y a d'autres approches plus qualitatives de la détermination de la lisibilité d'un texte. Elles incluent l'opinion des experts pédagogues, l'utilisation de «cloze tests» (textes à trous) qui demande de compléter un texte où des espace -mots ont été laissés vides et les questions de compréhension. Ces approches requièrent normalement l'intervention d'un expert même si aujourd'hui, avec les nouvelles technologies du

traitement du langage naturel (TALN), la génération de questions de compréhension a été partiellement automatisée (Feeney et Heilman, 2008 ; Mostow et al., 2004).

2.3. Quelques études quantitatives sur la lisibilité de l'arabe.

Nous résumons dans les prochains paragraphes quelques formules de lisibilité utilisées pour l'anglais et l'arabe et le peu d'autres études trouvées sur la lisibilité des textes arabes.

Une des meilleures formules connues pour l'anglais, le Fleish-Kincaid Score, par exemple, est une combinaison linéaire de la longueur moyenne des mots (mesurée en syllabes par mot) et de la longueur moyenne de la phrase (mesurée en mots par phrase) (DuBay, 2004). La formule Dale-Chall se fonde sur la durée moyenne de la phrase, et le rapport des mots «difficiles» au total des mots, les mots difficiles étant ceux qui n'appartiennent pas à une liste de mots supposément connus par un élève de quatrième année dans le système scolaire américain (DuBay, 2004).

La formule SMOG (Mc Laughlin, 1969) utilise une formulation algébrique un peu plus complexe impliquant le nombre de mots polysyllabes (mots avec plus de 3 syllabes) et le nombre de phrases dans le texte. Les valeurs obtenues à partir de ces formules de lisibilité sont utilisées pour lier un texte à un niveau de qualité, soit directement comme dans le cas de la formule Flesh-Kincaid Grade Level (Flesh-Kincaid Niveau Scolaire, une variante du Fleish-Kincaid Score) ou la formule SMOG, soit en passant par un tableau, comme dans le cas de la formule Dale-Chall. Par exemple, un score de 4.9 ou moins pour Dale-Chall indique que le texte est approprié pour la 4e année et les années précédentes, tandis qu'un score de 9 à 9,9 affecte un texte au niveau universitaire et un score de 10 et plus est réservé pour les textes adaptés aux niveaux au-dessus de la licence. Ces formules ont une précision variable et ne peuvent pas être appliquées directement aux langues en dehors de l'anglais.

Pour l'arabe, deux formules ont été trouvées dans la littérature: la formule Dawood (Dawood, 1977), qui comprend cinq traits du texte

(longueur moyenne des mots, longueur moyenne de la phrase, la fréquence des mots, le pourcentage de phrases nominales, et le pourcentage de noms définis)[2] et la formule Al-Heeti (Al-Ajlan et al., 2008 ; Al-Khalifa et Al-Ajlan, 2010), qui ne prend en considération que la longueur moyenne des mots. La sélection de ces caractéristiques spécifiques n'a pas été complètement justifiée dans la littérature et ces formules n'obtiennent pas de très bons résultats.

Dans une autre étude, (Al-Khalifa et Al-Ajlan, 2010) ont utilisé des techniques d'apprentissage-machine pour trouver la relation entre les caractéristiques des textes et les niveaux scolaires. Les traits choisis en tant qu'entrées du classificateur incluaient la longueur moyenne de la phrase (mesurée en nombre moyen de mots par phrase), la longueur moyenne des mots (mesurée en nombre moyen de caractères par mot), le nombre moyen de syllabes par mot, la fréquence des mots et les scores de perplexité du modèle bigramme. Les auteurs concluent que la durée moyenne de la phrase est le meilleur facteur pour déterminer la lisibilité du texte arabe avec un taux de précision de 66,67%, tandis que la meilleure combinaison de caractéristiques est la longueur moyenne de la phrase, le modèle bigramme et la fréquence des termes. Les auteurs remarquent aussi que la précision de la prédiction du modèle est excellente sur des textes simples mais significativement plus faible sur les textes du niveau moyen ou haut, les conduisant à douter comme dans les travaux précédents (Ajlan et al., 2008) de la validité de leur hypothèse de départ que les textes du curriculum saoudien sont correctement répartis entre les trois niveaux de difficulté. Ils soulignent la nécessité d'un corpus arabe correctement étiqueté avec des niveaux de lisibilité pour obtenir une plus grande confiance dans les études de mesures de lisibilité automatiques.

Dans une étude similaire visant à déterminer les facteurs qui influent sur la lisibilité d'un texte et son appartenance à un des 10 niveaux du cursus scolaire jordanien, (Al-Tamimi et al., 2013) ont utilisé l'analyse factorielle sur un ensemble de traits qui comprenait la longueur des mots, la fréquence des mots, la charge de vocabulaire, le nombre de mots difficiles, la durée moyenne de la phrase, la

complexité des phrases, la clarté de l'idée principale du texte, l'utilisation de métaphores et la complexité de la structure grammaticale. L'ensemble de traits initial a été réduit par la suite pour enlever une certaine redondance en utilisant l'analyse en composantes principales pour déterminer les caractéristiques les plus saillantes : le nombre de caractères du texte, la longueur moyenne des mots, et la longueur moyenne des phrases. Ces trois traits ont été utilisés pour créer la formule de base AARI et une formule dérivée pour associer un texte à un niveau scolaire. La meilleure performance (avec une précision de plus de 83%) a été obtenue lors de l'attribution d'un texte à l'un des trois groupes (du 1er au 3e année, 4e à 5e année et 6e à 10e année). Par contre, la précision est tombée à moins de 50% lors de l'attribution a chaque année individuellement.

Une troisième étude (Daud et al., 2013), qui a adopté une approche beaucoup plus simple pour évaluer la difficulté des textes, repose sur l'addition du score des mots dans un texte en le divisant par le nombre de mots dans le texte. Le score de chaque mot a été élaboré à partir de la fréquence des mots dans le «King Abdulaziz City for Science and Technology Arabic Corpus» (KACSTAC), un corpus général où les textes sont tirés de magazines, livres, journaux, revues, thèses, circulaires du gouvernement, curriculums scolaires, services d'infos et le web. Les auteurs ont supposé que les mots les plus fréquents sont les plus faciles. Le score d'un texte est obtenu en additionnant le classement inversé des mots dans le corpus de sorte que des scores plus faibles sont affectés à des textes plus faciles. Malheureusement, ce travail semble être plutôt préliminaire et n'a pas fourni de résultats précis sur l'efficacité de l'estimation de lisibilité.

Pour conclure, les travaux pour l'arabe ne sont ni nombreux ni toujours concluants. Une déduction générale que nous pouvons en tirer est que les formules développées pour les langues comme l'anglais ou le suédois ne marchent pas pour l'arabe. Les coefficients qui ont été estimés n'ont pas la bonne valeur et l'arabe est assez différent dans sa structure pour que les traits utilisés pour les autres langues ne soient pas nécessairement appropriés ni suffisants pour

caractériser la lisibilité des textes arabes. Les méthodes d'apprentissage automatique suggèrent que les meilleurs traits incluent: la fréquence des mots, le nombre de caractères dans le texte, le nombre moyen de caractères par mot, le nombre moyen de mots dans la phrase. Malheureusement, la fréquence des mots est une mesure relative à l'expérience de l'apprenant, qui n'est pas la même pour un élève qui est né et étudie dans un pays arabe et pour celui qui apprend l'arabe en tant que langue étrangère. De plus, le nombre de caractères par mot et le nombre de caractères dans le texte ne semblent pas prendre en considération que les processus de formation des mots en arabe sont bien différents de ceux des langues comme les langues latines et germaniques. Donc le calcul du nombre des caractères est peut-être une mesure sous-optimale de la complexité des mots en arabe.

2.4. Les projets REAP et ARET

Les travaux de lisibilité décrits ci-dessus concernent surtout la lecture dans la première langue, la langue maternelle, même si l'arabe standard n'est véritablement la langue maternelle de personne. Dans notre recherche nous essayons d'évaluer le texte par rapport aux connaissances linguistiques de l'apprenant de l'arabe en tant que langue étrangère. L'inspiration pour notre travail sur ce qui rend un texte approprié à un apprenant provient d'abord d'un projet précèdent, ARET (Arabic Reading Enhancement Tools) (Maamouri et al., 2012) qui a mis en ligne les textes des trois volumes d'une série de livres très fréquemment utilisée pour l'apprentissage de l'arabe aux adultes anglophones—*Al-Kitaab fii Ta^{c}allum al-^{c}Arabiyya, A Textbook for Beginning Arabic* (Brustad et al., 2004, 2007, 2001). Les outils résultant de ce projet-là sont hébergés au Linguistic Data Consortium de l'Université de Pennsylvanie à Philadelphie et fournissent à l'apprenant plusieurs outils, parmi eux un dictionnaire en ligne et un étiquetage morphologique complet des textes avec analyseur et lexique SAMA (Maamouri et al, 2010), la génération automatique des questions à réponses multiples basée sur l'analyse morphologique, et la synthèse de la parole avec la technologie développée par la société RDI (Egypte)[3] pour écouter la prononciation des mots et des textes.

Un des auteurs de cette communication, en participant au projet ARET, avait fait beaucoup de travail sur le contenu lexical et morphosyntaxique des textes pour développer un curriculum d'apprentissage de ces deux volets de la langue, avec l'intention de l'utiliser pour le développement d'un système intelligent de fouille de textes basé sur le modèle du projet REAP[4] (Reader-Specific Lexical Practice for Improved Reading Comprehension). Ce deuxième projet ciblait la lecture en tant que pratique du lexique adaptée à chacun pour l'amélioration de la compréhension des textes (Brown and Eskenazi, 2004). Le but était de sélectionner le texte approprié en s'appuyant sur une modélisation du curriculum (vocabulaire à savoir à chaque niveau), du texte (en utilisant les techniques de la recherche d'informations) et de l'apprenant (en utilisant un modèle à base d'histogrammes d'abord et ensuite de réseaux de Bayes). Appliqué d'abord à l'anglais, et successivement étendu au français et au portugais, REAP fournit la conception de base de notre plate-forme en voie de développement qui pourtant devra concevoir différemment le côté morphosyntaxique du curriculum, bien plus compliqué pour l'arabe.

1. Méthode et ressources

Dans cette partie de la communication nous tournons notre attention sur le travail dont l'aboutissement serait de pouvoir choisir un texte adapté aux connaissances de l'apprenant. Nous commençons par cibler «l'apprenant parfait», celui qui a tout appris au bon moment et n'oublie rien. On peut résumer notre approche de la façon suivante :

Nous prenons comme point de départ pour l'ensemble de textes et le vocabulaire des 25 premiers chapitres de la série du manuel Al-Kitaab (Brustad et al., 2004, 2007). Ces textes et leur vocabulaire définissent notre curriculum ou programme d'apprentissage. Nous traitons chaque chapitre comme une étape de compétences, étape qui est définie par le vocabulaire ciblé explicitement dans le chapitre. En se concentrant sur chaque étape, qu'on va nommer l'étape «actuelle», nous essayons de caractériser les textes associés à cette étape en termes de mots «connus», c'est-à-dire les mots ciblés dans la totalité

des étapes précédentes (d'où l'idée de l'apprenant parfait), les nouveaux mots «ciblés» dans l'étape actuelle, et les mots encore «inconnus» de l'apprenant, qui seront introduits à un stade ultérieur dans le programme ou peut-être pas du tout. L'objectif est de découvrir s'il existe une relation claire entre les pourcentages de mots connus, ciblés et inconnus dans les textes et des limites pour chaque catégorie pendant que nous progressons à travers le curriculum (les étapes d'apprentissage) et si cette relation peut être utilisée pour affecter de manière fiable un nouveau texte à un niveau d'apprentissage spécifique. À son tour, si cette relation s'avère utile pour déterminer le bon placement d'un texte dans un programme, elle peut en outre être utilisée pour rechercher des textes comme un moyen d'enrichir dynamiquement la collection existante pour les différentes étapes de qualification.

1.1. Ressources

Notre point de départ est une base de données de 67 textes de différentes longueurs (36-920 mots, moyenne de 219) provenant de Al-Kitaab, et spécifiquement de 20 chapitres du premier volume et 5 des 10 chapitres du deuxième. A partir de ces textes nous formulons une hypothèse (un modèle) des relations entre vocabulaire et texte qui rend le texte approprié à l'étape (le chapitre) où il est présenté à l'apprenant.

L'hypothèse a été testée d'abord sur un corpus de 23 textes ciblés aux cinq dernières étapes du curriculum pris en considération. Il s'agit de transcriptions de vidéos ou d'écrits de longueur comparable à ceux de Al-Kitaab choisis par un enseignant d'arabe expérimenté. On va aussi essayer l'hypothèse sur deux autres recueils des textes, comme par exemple, des textes qui font partie du curriculum de l'école syrienne, de la primaire au lycée (Syrian, 2013) et d'autres textes collectés manuellement à partir de journaux disponibles sur le web.

Pour le curriculum, nous nous sommes concentrés sur le vocabulaire, en ignorant pour le moment les traits morphosyntaxiques et en utilisant au début quatre différents niveaux de compétence:

- Haute : mots cibles spécifiquement dans l'étape actuelle et que l'étudiant doit apprendre;
- Moyenne: mots signalés dans certaines sections du texte (ex. grammaire, avant un texte);
- Basse: mots traduits dans le texte pour expliquer mais que l'apprenant ne doit pas nécessairement apprendre;
- Autre: mots que l'étudiant pourrait comprendre sur la base de leur relation avec des mots qu'il devrait connaitre en utilisant ses connaissances de morphologie.

Finalement, car le même mot peut apparaitre dans le curriculum plusieurs fois avec différents niveaux de compétence, comment établir quand le mot a été véritablement ciblé ? Après un bon nombre d'expériences pour examiner la sensibilité de notre analyse aux différents regroupements des catégories de compétence et différents traitement du moment d'introduction d'un mot, nous avons choisi de considérer seulement la catégorie de haute compétence, étant donné que l'étiquetage d'autres catégories n'était pas trop fiable au-delà des 20 premiers chapitres de Al-Kitaab. En conséquence, un mot est supposé être ciblé dans l'étape qui correspond à sa première introduction de haute compétence et il est supposé être connu dans tous les étapes suivantes.

1.2. Outils

Si la classification de mots par chapitre et l'étiquetage en termes de traits morphosyntaxiques ont été effectués manuellement, par contre pour pouvoir relier les mots présents dans les textes aux mots dans le lexique du curriculum (et de l'apprenant, par conséquence) nous avons utilisé le système MADA+TOKAN (Habash et al., 2009). Ce système accepte comme entrée une phrase en arabe codé en UTF-8, sépare les mots, détermine la partie du discours et les autres étiquettes morphosyntaxiques, et choisit même le sens du mot en utilisant l'analyseur morphologique et dictionnaire de SAMA (Maamouri et al., 2010) et d'autres outils et ressources qui lui

permettent de choisir les analyses les plus probables. Bien que cet étiquetage ne donne pas des résultats uniques, il a un haut niveau de précision (autour de 95-96% pour les variables qui nous intéressent tel que la partie du discours et le choix du sens). Spécifiquement pour les textes de Al-Kitaab, nous avons aussi vérifié que 80% de mots n'avaient pas plus de trois analyses. Une investigation sur une centaine de mots choisis de façon aléatoire parmi les mots contenus dans les textes des 25 chapitres a aussi démontré que MADA choisissait la bonne analyse dans 94% des cas. Donc nous avons décidé de baser les analyses suivantes sur le premier choix de MADA, tout en effectuant certaines analyses en parallèle avec le deuxième et troisième choix, ce qui nous a permis de conclure que nos résultats n'étaient pas vraiment influencés par cette incertitude.

Description de la procédure

Tous les textes ont été traités de la même façon pour les présenter à MADA. D'abord ils ont été segmentés en phrases, une phrase par ligne pour respecter les besoins de MADA, à l'aide d'outils automatiques et d'une vérification manuelle. La classification de la totalité des mots dans un texte en catégories «connu», «ciblé», «inconnu» a été faite après l'élimination de signes de ponctuation et des chiffres. La catégorie «inconnu» inclue soit les mots trouvés dans le curriculum lexical qui ne sont ni connus ni ciblés à l'étape considérée soit les mots qui n'ont pas été trouvés ni dans le curriculum ni par MADA dans le dictionnaire de SAMA.

Nous avons effectué le calcul pour chaque texte mais, car les textes peuvent varier beaucoup en longueur et contenu, les résultats sont plus valables au niveau du chapitre, c'est-à-dire en considérant l'ensemble de textes pour un chapitre (étape).

3. Résultats

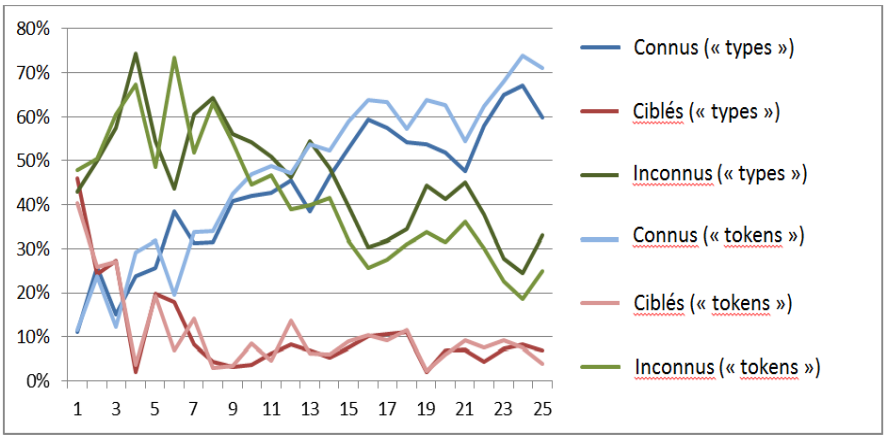

Figure 2: Tendances des mots connus, ciblés et inconnus

La figure 2 montre les tendances des mots connus, ciblés et inconnus le long du curriculum de 25 chapitres. Nous avons calculé le pourcentage de chaque catégorie en utilisant le nombre des mots sans compter les répétitions («tokens») et aussi après avoir éliminé les répétitions («types»). Il est raisonnable de supposer que dans les textes pour apprenants on va essayer de répéter plusieurs fois les mots ciblés par l'étape, mais le calcul basé sur les mots uniques donne une meilleure idée de la difficulté du texte en termes de mots connus et inconnus. Les deux tendances apparaissent être consistantes, mais les tendances des «types» semblent adoucir celles des «tokens».

Les résultats montrent clairement qu'en avançant dans le curriculum le nombre des mots connus augmente régulièrement, avec des oscillations, mais aussi qu'il y a encore un bon pourcentage (entre le 25% et le 45% pour les «tokens») qui peut être encore inconnu vers la fin de ce curriculum qui arrive seulement à un niveau de connaissance de la langue moyenne-basse. Donc même des textes qui posent encore des défis considérables au niveau de connaissances lexicales sont considérés appropriés pour les apprenants s'ils sont accompagnés par des soutiens tel que l'enseignant (humain ou ordinateur) capables de fournir des explications supplémentaires

quand il le faut. On peut aussi observer que le taux de mots ciblés reste moyennement constant autour de 8% et ne monte pas beaucoup au-delà de ce nombre si on élimine les premiers chapitres où il y a une forte charge de vocabulaire ciblé et inconnu et pas beaucoup de mots connus.

Ces résultats, bien que préliminaires, ont été utilisés pour prédire, à l'aide d'un arbre de décision probabiliste entrainé sur les textes et vocabulaire de Al-Kitaab, le placement dans le curriculum des 21 textes choisis spécialement par l'enseignant d'arabe comme appropriés pour les chapitres 21 à 25 de Al-Kitaab, et les résultats ont confirmé l'hypothèse implicite dans les tendances illustrées ci-dessus. D'autres analyses et expériences sont encore en cours avec d'autres textes et méthodes.

4. Conclusions et travaux futurs

Nous avons décrit le début d'une investigation qui essaye de relier le contenu des textes choisis pour les apprenants de l'arabe aux connaissances linguistiques que cet apprenant est supposé avoir acquises à différentes étapes d'un curriculum et nous avons constaté certaines régularités dans les tendances des mots connus, ciblés et inconnus. Les résultats ont été utilisés pour prédire, avec un certain succès, le bon placement des textes choisis par un enseignant pour les cinq derniers niveaux du curriculum considéré.

Nos résultats sont encore préliminaires, ils doivent encore être testés avec d'autres textes et améliorés avec d'autres techniques d'apprentissage automatique, mais ils nous donnent l'espoir de pouvoir développer une hypothèse qui nous aidera à évaluer d'autres textes disponibles sur le web pour construire une base de données riche et facile à enrichir dans le temps. Reste à établir aussi comment utiliser concrètement ces résultats pour chercher les textes avant de les évaluer, car si les pourcentages des mots dans les trois catégories nous aident à décider si un texte peut être considéré approprié, c'est-à-dire *ni trop difficile ni trop facile* pour une étape spécifique, ils ne nous aident pas à décider quels mots parmi ceux ciblés utiliser pour la recherche des textes ni comment les combiner car il n'est pas probable

de pouvoir trouver des textes qui utilisent tous les mots ciblés à la fois. Ce travail fera surement partie de nos investigations futures. Il faut aussi rappeler que cette étude jusqu'à maintenant s'est focalisée sur le volet lexique des textes, tout en ignorant le volet morphosyntaxique, mais les prochaines études cibleront certains aspects de la morphologie et de la syntaxe locale.

Pour conclure, on peut espérer que certains de nos résultats pourront être appliqués au tamazight, surtout au niveau lexical car la morphosyntaxe est assez différente entre le tamazight et l'arabe. Il est important de remarquer quand même que, avant de pouvoir faire cette expérience, il va falloir collecter un corpus où chaque texte est étiqueté avec un niveau de difficulté, soit explicitement soit implicitement par le biais de la séquence des textes dans le programme d'études, comme dans notre cas. Il va aussi falloir développer un curriculum lexical et morphosyntaxique qui suit les textes, manuellement ou à l'aide d'outils semi-automatiques qui pour le tamazight sont malheureusement encore peu abondants.

Remerciements

Ce travail a commencé à voir le jour pendant le déroulement du projet ARET, développé chez le Linguistic Data Consortium (LDC) de Université de Pennsylvanie et financé par le Département d'Education des Etats Unis d'Amérique à travers le programme d'Etudes de Recherche Internationale (International Research Study— IRS Grant No. P017A050040-07-05. Nous remercions nos collègues de LDC qui nous ont donné la permission d'utiliser certaines de leurs ressources linguistiques.

Références

Al-Ajlan, A. A., Al-Khalifa, H. S., and Al-Salman, A. S. (2008), «Towards the Development of an Automatic Readability Measurements for Arabic Language», *in Proceedings of the Third International Conference on Digital Information Management*, November 13-16, p. 506-511.

Al-Khalifa, H. S., and Al-Ajlan, A. A. (2010), «Automatic Readability Measurements of the Arabic Text: An Exploratory Study», The Arabian Journal for Science and Engineering, Volume 35, Number 2C, p. 103-124.

Al-Tamimi, A-K., Jaradat, M., Aljarrah N., and Ghanem, S. (2013), «AARI: Automatic Arabic Readability Index», *The International Arab Journal of Information Technology*, [Accepted March 12, 2013].

http://www.ccis2k.org/iajit/PDF/vol.11,no.4/5200.pdf, August 2013.

Ben-Simon, A. and Cohen, Y. (2011), «The Hebrew Language Project: Automated Essay Scoring & Readability Analysis», International Atomic Energy Agency.
http://www.iaea.info/documents/paper_4e1237ae.pdf, August 2013.

Brown, J., and Eskenazi, M. (2004), «Retrieval of authentic documents for reader-specific lexical practice», *in Proceedings of InSTIL/ICALL Symposium*, June 17-19, Venice.

Brustad, K., Al-Batal, M., and Al-Tonsi A. (2004), *Al-Kitaab fii Ta^{c}allum al-^{c}Arabiyya, A Textbook for Beginning Arabic: Part One*, Second Edition, Washington D.C., Georgetown University Press.

Brustad, K., Al-Batal, M., and Al-Tonsi A. (2007), *Al-Kitaab fii Ta^{c}allum al-^{c}Arabiyya, A Textbook for Beginning Arabic: Part Two*, Second Edition, Washington D.C., Georgetown University Press.

Brustad, K., Al-Batal, M., and Al-Tonsi A. (2001), *Al-Kitaab fii Ta^{c}allum al-^{c}Arabiyya, A Textbook for Beginning Arabic: Part Three*, First Edition, Washington D.C., Georgetown University Press.

Daud, N. M., Hassan H., and Abdul Aziz, N. (2013), «A Corpus-Based Readability Formula for Estimate of Arabic Texts Reading Difficulty», *World Applied Sciences Journal*, Volume 21, p. 168-173.

DuBay, W. H. (2004). *The Principles of Readability*. Institute of Education Sciences, http://eric.ed.gov/?id=ED490073, August 2013.

Dawood, B. A-K. (1977). *The Relationship between Readability and Selected Language Variables*. Thesis submitted to the College of Education

Board in the Universit of Baghdad in partial fulfillment of the requirements for the degree of Master of Arts in Education and Psychology.

http://dspace.ju.edu.jo/xmlui/bitstream/handle/123456789/12718/JUA0305740.pdf, August 2013.

Feeney, C. et Heilman, M. (2008), «Automatically Generating and Validating Reading-Check Questions», Ninth International Conference on Intelligent Tutoring Systems.

Habash, N., Rambow, O., and Roth, R. (2009), «MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization», *in Proceedings of the 2nd International Conference on Arabic Language Resources and Tool*s (MEDAR), April 22-23, Cairo, 2009.

Maamouri, M. (2009), «Arabic Literacy», *in Encyclopedia of Arabic Language and Linguistics*, Lemma 11,16, Volume 2, Brill.

Mohamed Maamouri, David Graff, Basma Bouziri, Sondos Krouna, Ann Bies and Seth Kulick (2010), *Standard Arabic Morphological Analyzer (SAMA) Version 3.1*, Linguistic Data Consortium, Catalog No.: LDC2010L01.

Maamouri, M., Zaghouani, W., Cavalli-Sforza, V., Graff, D., and Ciul, M. (2012), «Developing ARET: An NLP-based Educational Tool Set for Arabic Reading Enhancement», *in Proceedings of The 7th Workshop on Innovative Use of NLP for Building Educational Application*s (NAACL-HLT-2012), June 7, Montreal, p. 127-135.

Mc Laughlin, H. G. (1969), «SMOG Grading - a New Readability Formula», *Journal of Reading*, Volume 12, Number 8, p. 639-646.

Mostow, J., Beck, J., Bey, J., Cuneo, A., Sison, J., Tobey, B., and Valeri, J. (2004), «Using Automated Questions to Assess Reading Comprehension, Vocabulary, and Effects of Tutorial Interventions», *Technology, Instruction, Cognition and Learning*, Volume 2, p. 103-140.

Pang, L. T. (2006), *Chinese Readability Analysis and its Applications on the Internet*, Thesis submitted in partial fullfilment of the requirements for the degree of Master of Philosophy in Computer Science and Engineering, The Chinese University of Hong Kong.

Syrian (2013), «المناهج الجديدة»,
http://me.syrianeducation.org.sy/ebook/classes.html, March 2013.

Tateisi, Y., Ono, Y., and Yamada, H. (1988), «A Computer Readability Formula of Japanese Texts for Machine Scoring», *in Proceedings of COLING 1988*, Volume 2, August 22-27, Budapest, p. 649-654.

1- La langue d'enseignement correspond normalement à la langue maternelle pour la plupart des études, mais ceci n'est pas le cas de l'arabe standard (Maamouri, 2009).
2- La formule Dawood comprend ces cinq traits selon (Al-Khalifa et Al-Ajlan, 2010) et (Al-Tamimi et al., 2013), mais seulement trois selon le travail originel de Dawood (Dawood, 1977) : longueur moyenne des mots, longueur moyenne de la phrase et répétition moyenne des mots.
3- <http://www.rdi-eg.com/Technologies/speech.htm>
4 - <http://reap.cs.cmu.edu/index.html>