Un dictionnaire électronique pour apprenant de l'arabe (langue seconde) basé sur corpus

ZAAFRANI Riadh
Faculté des Sciences Juridiques, Economiques et de Gestion de Jendouba
Avenue de l'UMA – Jendouba – 8189 – Tunisie.
et Unité de recherches : Traitement automatique d'un corpus textuel arabe
Faculté des lettres – Université de La Manouba – Tunisie
Tél.: ++216 78 600 300 - Fax: ++216 78 601 176
Mél.: zaafrani.1@excite.com

Résumé – Abstract

Dans cet article, nous décrivons un dictionnaire électronique pour apprenant de l'arabe (langue seconde) basé sur corpus. Nous présentons d'abord les difficultés rencontrées par les utilisateurs des dictionnaires classiques avant de proposer des solutions à ces problèmes. Pour réaliser ce dictionnaire, nous avons élaboré des outils informatiques pour le traitement automatique de la langue arabe, dont notamment un analyseur des mots graphiques arabes et un étiqueteur morpho-syntaxique de textes. Nous montrerons l'utilité de ces outils et comment elles ont servi à élaborer le dictionnaire. Nous indiquerons enfin, comment nous avons résolu le problème d'accès lexical qui constitue le principal obstacle rencontré par les utilisateurs des dictionnaires arabes.

This contribution describes an Arabic electronic dictionary for foreign learners. We'll first list the difficulties encountered by classical dictionary users and then we'll give some solutions for these problems. To build this electronic dictionary, we have developed some Arabic language processing tools, among which are lexical analysis module and text morphosyntactic tagging module. So we'll describe these tools and show how they have been used for building the dictionary. We'll finally demonstrate how this dictionary can solve the problem of lexical access, that represents the main problem encountered by Arabic dictionary users.

Keywords – Mots Clés

apprentissage, langue seconde, langue naturelle, arabe, dictionnaire, lexique, vocabulaire, apprenant.

learning, second language, natural language, Arabic, dictionary, lexicon, vocabulary, learner.

1 Introduction

Le processus de compréhension de textes arabes présente d'énormes difficultés surtout pour les apprenants débutants. La principale difficulté consiste à repérer des unités de sens dans un flux d'information qui peut sembler assez flou. Le texte est, en effet, représenté dans une graphie d'où sont absentes les voyelles brèves et dont le découpage en « mots » n'apparaît pas toujours évident. L'apprenant se jette alors dans le dictionnaire, espérant que la signification d'un mot le mettra sur la voie. Mais en analysant les mots graphiques, il est parfois incapable de dégager la « bonne » racine qui constitue l'entrée des dictionnaires arabes, et même s'il le réussit, il se perd dans les différentes significations du mot à chercher, ce qui nuit véritablement au processus de lecture du texte. Le dictionnaire papier reste cependant, son seul recours face aux textes incompréhensibles en l'absence d'une aide humaine. Pour remédier à cet état de fait, nous tenterons dans cet article de jeter les fondements d'un dictionnaire électronique adapté aux apprenants de l'arabe et nous présenterons une réalisation d'un dictionnaire qui leur offre un accès simple et rapide aux sens et autres propriétés des mots lus dans les textes.

2 Les dictionnaires classiques

Selon Bogaards (Bogaards, 1995), plusieurs expériences ont montré que le dictionnaire classique ne semblait pas améliorer la compréhension des textes d'une manière significative. Pour cela, il avance plusieurs raisons :

- Les apprenants n'aiment pas utiliser un dictionnaire. Ils le considèrent comme une étape obligée et contraignante qui les détourne de leur lecture.
- Ils ne savent pas utiliser un dictionnaire. Ils ont des difficultés à repérer l'information pertinente et acceptent la moindre indication qui va dans le sens de leur hypothèse initiale de manière à abréger l'épreuve.
- Le dictionnaire nuit au processus de lecture : des expériences montrent que des étudiants utilisant un dictionnaire mettaient souvent plus de temps à terminer leur tâche, sans pour autant obtenir de meilleurs résultats.

Face à ce constat, Bogaards (Bogaards, 1995) en déduit qu'il faut, d'une part, avoir un niveau de connaissance avancé sur la langue pour pouvoir profiter des informations contenues dans les dictionnaires, et d'autre part, avoir une bonne dose de ténacité et de courage.

3 Les dictionnaires pédagogiques

Le dictionnaire pédagogique ou dictionnaire pour apprenant est un monolingue destiné aux personnes apprenant une langue étrangère. Il se différencie sur un certain nombre de points du monolingue pour natif. Davantage de précisions sont données sur les entrées du dictionnaire, dont nous indiquons les principaux apports :

*** Le vocabulaire définitoire :** Le point le plus significatif et le plus visible est l'utilisation d'un vocabulaire définitoire contrôlé pour décrire les entrées dans les définitions. Il est nécessaire en effet de définir et d'expliquer des mots inconnus avec des mots simples, que l'apprenant est susceptible de connaître. Ce procédé a cependant l'inconvénient, dans certains cas, d'alourdir de façon notable les définitions.

*** Les exemples :** L'utilisation d'un corpus textuel pour l'élaboration d'un dictionnaire, permet d'obtenir des exemples authentiques illustrant l'entrée définie. Ces exemples contiennent

cependant un nombre important de mots hors du vocabulaire contrôlé des définitions.
*** Les illustrations :** L'illustration est très utile dans les dictionnaires pour apprenants, puisqu'elle supplée la définition lorsque les moyens purement linguistiques sont insuffisants pour expliquer un mot ou produisent une définition trop lourde.
*** Les renvois :** Les dictionnaires pédagogiques utilisent des renvois analogiques vers d'autres mots lorsqu'ils estiment qu'une comparaison est nécessaire pour bien saisir les nuances. Les mots auxquels on renvoie sont des synonymes, antonymes, mots composés ou présentant des similarités morphologiques.
*** Les informations grammaticales :** Les informations grammaticales occupent un espace relativement important dans les dictionnaires pour apprenants. Les apprenants de langue étrangère ont en effet besoin, plus que les natifs, d'être dûment renseignés sur ce point afin de s'exprimer correctement.
* **Les fréquences et les registres :** Deux éléments d'informations peuvent assister l'apprenant en production : la fréquence et le registre. Les dictionnaires basés sur corpus permettent d'associer des indications sur la fréquence de chaque entrée dans le corpus. L'apprenant est ainsi informé de l'utilisation réelle d'une entrée et peut en mesurer, par exemple, son côté démodé ou au contraire dans le vent. Le COBUILD par exemple (Collins, 1999) applique systématiquement une échelle de 0 à 5 diamants (pour les plus fréquents). Le COBUILD aborde aussi les problèmes de pragmatique en appliquant des registres spécialisés aux entrées du dictionnaire tels que journalisme, légal, littéraire, démodé, parlé, écrit, etc.
*** Limites des dictionnaires pédagogiques :** Malgré toutes les améliorations apportées aux dictionnaires classiques, quelques problèmes persistent encore et rendent parfois les dictionnaires pédagogiques inefficaces. Le problème d'accès lexical reste particulièrement présent (surtout pour les dictionnaires arabes) et augmente la durée de consultation du dictionnaire, ce qui nuit au processus de lecture des textes.

4 Les dictionnaires électroniques

Les dictionnaires électroniques ont apporté un nouveau mode d'usage et quelques améliorations :
1) L'accès lexical : Le passage d'une forme fléchie telle qu'elle est dans le texte à la forme canonique dans un dictionnaire n'est pas évident dans une langue comme l'arabe dont le processus d'analyse morphologique est difficilement maîtrisable. Dès lors, les possibilités de traitement automatique de la langue et de recherche d'un mot à partir d'un index constituent les grands atouts du dictionnaire électronique (Selva, 1999). Pour accéder aux entrées d'un dictionnaire électronique, deux solutions sont généralement proposées :

- une première solution qui consiste à lister l'ensemble des entrées du dictionnaire à partir desquelles la forme fléchie pourrait être dérivée;
- et une seconde solution plus sophistiquée, qui fait fonctionner un analyseur morphologique et permet ainsi de mieux cerner l'entrée appropriée dans le dictionnaire.

2) L'interactivité : Le second avantage des dictionnaires électroniques est l'interactivité qui facilite la navigation entre les différentes informations du dictionnaire et améliore l'efficacité de la consultation. L'utilisateur peut influer sur la nature et la quantité des informations qui lui sont présentées et s'affranchit des limites du papier. Le support informatique n'est en effet plus tributaire du manque de place des versions papier. Il doit par conséquent permettre d'agir sur la présentation en laissant la possibilité à l'utilisateur d'afficher tel ou tel élément d'information. On peut choisir de visualiser les définitions, ou les exemples, ou les règles

grammaticales, ou les traductions etc., ou bien des combinaisons de plusieurs d'entre eux. Les dictionnaires électroniques permettent aussi d'avoir des renvois entre différentes entrées du dictionnaire, des liens directs avec d'autres applications, des possibilités de recherche complexe et des interactions avec les utilisateurs.

5 Réalisation d'un dictionnaire prototype

Dans le cadre du projet européen DIINAR-MBC (DIctionnaire INformatisé de l'ARabe, Multilingue et Basé sur Corpus) (Dichy et al., 1998), nous avons conçu et réalisé un dictionnaire prototype. Ce prototype est constitué d'environ 8000 unités lexicales (verbes, noms et adjectifs). Il reprend, les informations morphologiques et syntaxiques antérieurement développées et stockées dans DIINAR.1 (Braham et al., 2002) et ajoute les définitions et les équivalents en français et en anglais. Sur la base de ce prototype, nous avons établi un dictionnaire électronique pour apprenant que nous exploitons dans le cadre d'un environnement d'apprentissage (Zaafrani, 2002b). Nous avons d'abord réalisé des outils de traitement automatique de l'arabe qui vont assister l'expert humain dans le choix du vocabulaire définitoire et des exemples du corpus textuel. Nous avons essayé ensuite de répondre aux problèmes soulevés concernant l'usage des différents types de dictionnaires.

5.1 Elaboration du dictionnaire

a) Choix du vocabulaire définitoire : Le choix du vocabulaire définitoire relève d'un outil que nous avons élaboré pour le calcul des fréquences (Zaafrani, 2002a). Pour le prototype de DIINAR-MBC, il s'est établi sur un large corpus de textes scolaires tunisiens bruts d'environ 10 millions de mots. Il s'agit d'un outil permettant, à partir d'un texte brut donné, de générer dans un fichier les fréquences de tous les mots rencontrés. Ce programme se base sur les résultats retournés par l'analyseur des mots graphiques (Zaafrani, 2002a), qui sont souvent équivoques. Les résultats obtenus présentent ainsi quelques imprécisions et font l'objet d'un travail de révision d'un expert humain.

b) Choix des exemples : Afin de rechercher rapidement les exemples d'un mot donné dans un corpus textuel, et de fournir tous les contextes des occurrences trouvées, nous avons développé un concordanceur adapté à la langue arabe (Zaafrani, 2002a). Ce concordanceur ne se contente pas de rechercher une forme graphique dans un texte, mais assure une fonction d'analyse qui ramène d'abord le mot recherché à son lemme, puis d'effectuer la recherche à partir de ce lemme, ce qui permet d'obtenir les contextes des différentes formes fléchies du mot en question (figure 1).

Figure 1 : Recherche de concordances

5.2 Principaux apports

a) Accès lexical : Le schéma ci-dessous (figure 2) résume les différentes étapes permettant le passage d'un mot graphique arabe à sa signification dans le dictionnaire. Ce schéma est divisé en deux colonnes : la colonne gauche correspond à un accès classique alors que la colonne droite correspond à une méthode d'accès optimisée que nous employons dans le cadre d'un environnement d'apprentissage (Zaafrani, 2002b). Il existe en effet 3 modes d'accès différents au dictionnaire qui correspondent à 3 situations différentes :

- L'apprenant cherche le sens d'un mot dont il ignore la forme canonique (i.e. l'entrée du dictionnaire). Dans ce cas, il recourt à l'analyseur morphologique qui lui proposera un ensemble de solutions possibles parmi lesquelles il sélectionnera la forme canonique du mot en question.
- L'apprenant cherche le sens d'un mot dont il connaît la forme canonique. Dans ce cas, le système lui proposera les différentes significations de la forme parmi lesquelles il sélectionnera la bonne signification du mot à partir des autres indications (fréquence, illustration, etc.).
- L'apprenant cherche le sens d'un mot à partir d'un texte préalablement étiqueté. Une simple sélection du mot le fait directement passer directement à sa signification (figure 3).

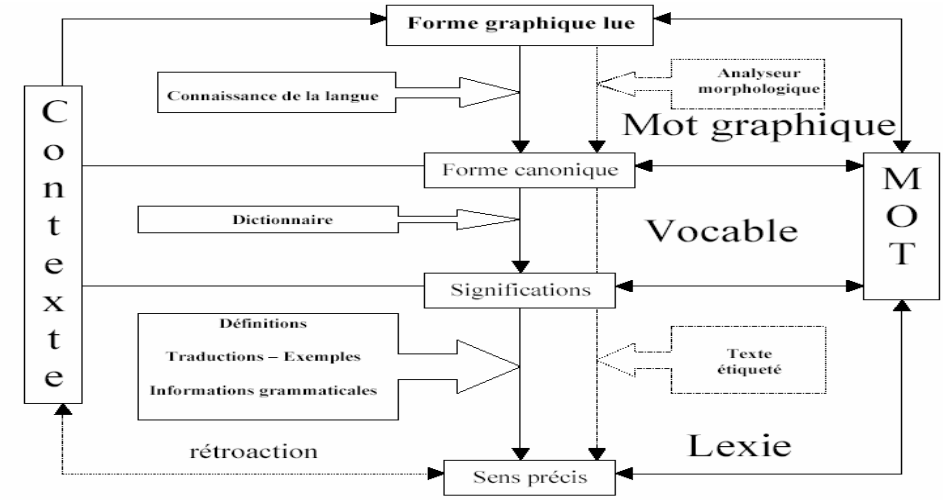

Figure 2 : Modèle d'accès lexical au dictionnaire

C'est ce dernier mode d'accès qui sera le plus utilisé dans l'environnement d'apprentissage, puisqu'on utilisera des textes complètement étiquetés. Pour étiqueter les textes nous avons développé un outil qui permet de passer d'un texte brut, exempt d'informations linguistiques, à une séquence d'**unités élémentaires** lexicales (les lemmes) assorties d'étiquettes morphologiques, syntaxiques et sémantiques (Zaafrani, 2002a). Cet étiqueteur est semi-automatique parce qu'il fait intervenir l'expertise humaine lorsque les résultats d'analyse sont équivoques.

b) Ergonomie des interfaces de consultation : Un autre apport du dictionnaire réalisé est le travail effectué sur l'ergonomie des interfaces de consultation. En effet, les informations sont hiérarchisées selon leur importance. Les utilisateurs qui ne cherchent qu'à consulter le sens d'un mot ne seront pas ainsi détournés de leur lecture. Par contre, les utilisateurs qui cherchent à approfondir leurs connaissances grammaticales peuvent naviguer entre les différents liens qui sont proposés sur l'interface (figure 3). Les renvois sont facilement repérables et la nature des relations sont très clairs.

Figure 3 : Accès au dictionnaire à partir d'un texte

6 Conclusion et perspectives

Nous avons défini dans cet article les bases d'un dictionnaire électronique pour apprenant. A partir de ces critères, nous avons réalisé un premier prototype pour la langue arabe basé sur un corpus de textes scolaires. Le dictionnaire obtenu, est très riche en informations grammaticales et en exemples, et constitue un bon outil d'aide aux apprenants lors des tâches de compréhension et de production. Mais, étant donné que c'est le sens et les définitions des mots qui sont les premières préoccupations de l'utilisateur, ce prototype fait actuellement l'objet d'un énorme travail d'enrichissement et de révision de la part d'experts linguistes afin de maximiser le taux du vocabulaire définitoire contrôlé dans les définitions. C'est en effet, le vocabulaire utilisé qui va déterminer le niveau et le domaine des utilisateurs du dictionnaire. A l'avenir, on envisage de réaliser différentes versions du dictionnaire, chacune répondant à un profil particulier d'utilisateurs.

Références

Bogaards P. (1995), Dictionnaires et compréhension écrite, *Cahiers de Lexicologie 67*, 1995-2, pp. 37-53.

Braham A., Dichy J, Ghazeli S., Hassoun M. (2002), La base de connaissance linguistique DIINAR.1, *Proceedings of the International Symposium on : the processing of arabic*, Université de Manouba, Tunisie, pages 45-56, 18 – 20 Avril 2002.

Collins (1999) : *Cobuild Home Page*, consulté en Février 2004 : http://titania.cobuild.collins.co.uk

Dichy J., Hassoun M. (1998), Some aspects of the DIINAR-MBC research programme, *Proceedings of 6th ICEMCO, International Conference and Exhibition on Multi-lingual Computing*, Cambridge, England, April 16-19.

Selva T. (1999), *Ressources et activités pédagogiques dans un environnement informatique d'aide à l'apprentissage lexical du français langue seconde*, Thèse d'Université, Université de Franche-Comté, Besançon, 210 pages, Octobre 1999.

Zaafrani R. (2002a), *Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère*. Thèse de doctorat, Université Lyon 2, Janvier 2002.

Zaafrani R. (2002b), Le système « AL-MucaLLiM » : une tentative pour un apprentissage assisté par ordinateur de l'arabe *Proceedings of the International Symposium on : the processing of arabic*, Université de Manouba, Tunisie, pages 88-93, 18 – 20 Avril 2002.