Utilisation d'une approche basée sur la recherche cross-lingue d'information pour l'alignement de phrases à partir de textes bilingues Arabe-Français

Nasredine SEMMAR, Christian FLUHR
CEA, LIST, LIC2M
18 route du Panorama, BP6, FONTENAY AUX ROSES, F- 92265 France
{nasredine.semmar,christian.fluhr}@cea.fr

Résumé. L'alignement de phrases à partir de textes bilingues consiste à reconnaître les phrases qui sont traductions les unes des autres. Cet article présente une nouvelle approche pour aligner les phrases d'un corpus parallèle. Cette approche est basée sur la recherche cross-lingue d'information et consiste à construire une base de données des phrases du texte cible et considérer chaque phrase du texte source comme une requête à cette base. La recherche cross-lingue utilise un analyseur linguistique et un moteur de recherche. L'analyseur linguistique traite aussi bien les documents à indexer que les requêtes et produit un ensemble de lemmes normalisés, un ensemble d'entités nommées et un ensemble de mots composés avec leurs étiquettes morpho-syntaxiques. Le moteur de recherche construit les fichiers inversés des documents en se basant sur leur analyse linguistique et retrouve les documents pertinents à partir de leur indexes. L'aligneur de phrases a été évalué sur un corpus parallèle Arabe-Français et les résultats obtenus montrent que 97% des phrases ont été correctement alignées.

Abstract. Sentence alignment consists in identifying correspondences between sentences in one language and sentences in the other language. This paper describes a new approach to aligning sentences from a parallel corpora. This approach is based on cross-language information retrieval and consists in building a database of sentences of the target text and considering each sentence of the source text as a query to that database. Cross-language information retrieval uses a linguistic analyzer and a search engine. The linguistic analyzer processes both documents to be indexed and queries to produce a set of normalized lemmas, a set of named entities and a set of nominal compounds with their morpho-syntactic tags. The search engine builds the inverted files of the documents on the basis of their linguistic analysis and retrieves the relevant documents from the indexes. An evaluation of the sentence aligner was performed based on a Arabic to French parallel corpus and results show that 97% of sentences were correctly aligned.

Mots-clés : alignement de phrases, corpus parallèle, recherche cross-lingue d'information.
Keywords: sentence alignment, parallel corpora, cross-lingual information retrieval.

1 Introduction

L'alignement de textes bilingues dont l'un est une traduction de l'autre consiste à mettre en relation des unités linguistiques ou logiques qui se correspondent dans les deux textes. Ces unités peuvent être des paragraphes, des phrases, des syntagmes, des mots, etc. L'alignement de textes permet l'élaboration de lexiques et de bases de données phraséologiques multilingues nécessaires pour la traduction et la terminologie. Plusieurs techniques d'alignement de textes ont été proposées (Gale, Church, 1991) (Brown et al., 1991) (Debili, Samouda, 1992) (Gaussier, 1995) (Melamed, 1996) (Fluhr et al., 2000).

Dans cet article, nous présentons un aligneur de phrases à partir de corpus parallèles utilisant une approche basée sur la recherche d'information cross-lingue et combinant plusieurs sources d'information (dictionnaire bilingue, longueurs des phrases, numéros d'ordre des phrases dans le corpus parallèle). Cet aligneur a été développé initialement pour aligner les corpus parallèles Français-Anglais, il a été ensuite adapté pour aligner les corpus des couples de langues Arabe-Français et Arabe-Anglais.

Nous présentons dans la section 2 les principaux composants du moteur de recherche cross-lingue du LIC2M, en particulier, nous nous focalisons sur les modules de l'analyse linguistique. Dans la section 3, nous décrivons le prototype de notre aligneur de phrases. Nous discutons dans la section 4 les résultats obtenus en alignant le corpus MD (Monde Diplomatique) de la campagne ARCADE II. La section 5 conclut notre étude et présente nos travaux futurs.

2 Le moteur de recherche cross-lingue

Le moteur de recherche cross-lingue permet, à partir d'une requête en une seule langue, de fournir des réponses trouvées dans des documents qui sont dans d'autres langues. Le moteur de recherche cross-lingue du LIC2M est composé d'un analyseur linguistique, d'un analyseur statistique, d'un reformulateur et d'un comparateur (Semmar et al., 2005).

2.1 L'analyse linguistique

L'analyse linguistique des documents et de la requête est un composant important dans le système de recherche d'information cross-lingue du LIC2M. L'analyseur linguistique LIMA (LIc2m Multilingual Analyser) est composé d'un ensemble de modules dont le nombre et la nature varient selon la langue traitée et un ensemble de ressources linguistiques. Selon que l'on traite l'arabe, le français ou le chinois, le système sait modifier le traitement et utiliser les ressources adaptées. Nous présentons dans les sections suivantes les modules et les ressources utilisés dans l'analyseur linguistique LIMA en se focalisant sur les traitements spécifiques à la langue arabe (Grefenstette et al., 2005).

2.1.1 Modules de traitement linguistique

Certains des ces modules sont utilisés pour le traitement de la majorité des langues traitées par LIMA. D'autres, plus spécifiques, ne sont utilisés que pour certaines langues.

1. La tokenisation qui découpe le texte en mots (tokens).

2. La consultation du dictionnaire des formes qui permet éventuellement de récupérer des informations linguistiques concernant les mots à reconnaître. Pour ceux qui sont semi voyellés ou non voyellés, cette consultation du lexique permet de récupérer les formes voyellées correspondantes, c'est à dire leurs alternatives orthographiques lorsqu'elles existent. Dans le cas par exemple du mot non voyellé مدرسة la recherche dans le dictionnaire donne les deux alternatives orthographiques suivantes: مَدْرَسَة "Ecole" (Nom commun féminin singulier) et مُدَرِّسَة "Institutrice" (Nom commun féminin singulier).

3. Lorsque leur forme de surface le permet, les mots sont segmentés en proclitique-radical-enclitique ou en proclitique-radical ou en radical-enclitique ou en proclitique-enclitique. Ce module de segmentation n'est utilisé que pour l'Arabe et l'Espagnol. Par exemple, le mot مدرسه est candidat au découpage مُدَرِّس + ه "Instituteur" + "lui, à lui".

4. Les expressions idiomatiques sont ensuite reconnues et regroupées pour être considérées comme un seul mot dans le graphe d'analyse. Cette reconnaissance se fait à l'aide de règles de déclencheurs qui sont généralement des lemmes. Ces règles permettent par exemple de reconnaître les noms de mois arabes ذُو القِعْدَة et جُمَادَى الأولى comme des mots uniques.

5. Si, après ces étapes, un mot reste inconnu, le système lui attribue une/des catégorie(s) par défaut en s'appuyant généralement sur des informations révélées par sa forme de surface.

6. Après cette analyse morphologique, la majorité des mots restent ambigus notamment à cause du nombre élevé des voyellations possibles. Le rôle du désambiguïseur morpho-syntaxique est ensuite de réduire le nombre des ambiguïtés en utilisant des matrices de désambiguïsation. Ce sont des matrices de bi-grammes et tri-grammes obtenues à partir d'un corpus de 13 200 mots pour l'Arabe et de 25 000 mots pour le Français. Ce corpus est étiqueté et désambiguïsé manuellement. La précision du désambiguïseur morpho-syntaxique est d'environ 91% pour l'Arabe et de 94% pour le Français.

7. Une analyse syntaxique tente ensuite, par un jeu de règles écrites à la main, d'établir les relations de dépendance entre les mots dans un même syntagme et entre les syntagmes dans une même phrase. Par exemple, dans la chaîne nominale توزيع المياه الصالحة "Distribution des eaux potables", l'analyse syntaxique considère que cette chaîne est un mot composé des mots توزيع "Distribution" (nom commun), مياه "Eaux" (nom commun) et صالحة "Potables" (adjectif).

8. Une reconnaissance des entités nommées est ensuite activée. Cette étape de l'analyse utilise des fichiers de listes ainsi que de règles de déclencheurs pour reconnaître des entités telles que les noms de personnes, d'organisations, de produits et de lieux, les dates ainsi que les unités de mesure. Ainsi, un énoncé comme الأوّل مِن شَهْر مَارس "Le premier du mois de Mars" est reconnu comme une date et الشَّرْق الأوْسَط "Le Moyen-Orient" est reconnu comme un nom de lieu.

2.1.2 Ressources linguistiques

Pour le traitement de l'arabe, le système dispose de ressources lexicales et grammaticales suivantes:

- Un dictionnaire de formes qui contient toutes les formes fléchies et dérivées simples des mots en arabe. Ce dictionnaire est obtenu par un fléchisseur automatique développé au sein du LIC2M (Debili, Zouari, 1985). Cet outil produit 3 164 000 entrées à partir de 14 000 lemmes (noms, adjectives et verbes). Le dictionnaire final contient également les listes fermées comme les pronoms, les prépositions, les nombres, etc. Les mots du dictionnaire ont deux sortes d'entrées: des formes complètement voyellées ou complètement dévoyellées. Seules les entrées voyellées possèdent des informations linguistiques (catégorie, genre, nombre, etc.). Les entrées non voyellées, qui sont ambiguës par nature, ne possèdent que des pointeurs vers les entrées voyellées correspondantes et donc leur informations linguistiques comme il a été montré dans l'exemple de مدرسة plus haut. Le contenu du dictionnaire ne permet pas seulement d'attribuer des voyellations aux mots non voyellés mais aussi de proposer les différentes alternatives concernant certaines lettres comme c'est la cas pour آ إ أ qui sont trois manières différentes d'écrire la lettre ا et pour ىء et ؤء qui sont deux alternatives à l'écriture pour les lettres ئ et ئ.

- Un dictionnaire de proclitiques ainsi qu'un dictionnaire d'enclitiques simples et composés. La même structure est attribuée à ces entrées, c'est à dire une forme voyellée et une forme non voyellée correspondante. Par exemple, le proclitique أفبِ est décomposé en trois parties أ+فَ+بِ.

- Des dictionnaires bilingues pour toutes les paires de langues traitées par le système sont également disponibles. Ces dictionnaires permettent la reformulation bilingue dans le cadre de la recherche d'information cross-lingue. Le dictionnaire bilingue Arabe-Français est composé de 120 000 entrées validées manuellement.

2.2 Analyse statistique

L'analyse statistique consiste à attribuer un poids aux mots simples et aux mots composés sur l'ensemble des documents indexés, selon le "degré d'information" qu'ils contiennent. Ce poids est lié à l'hétérogénéité de répartition du terme dans la base de documents. Il sera maximum si le terme est complètement discriminant, c'est-à-dire s'il apparaît dans un seul document, et minimum s'il n'est pas discriminant et apparaît dans tous les documents (Andreewsky et al., 1981).

2.3 Reformulation de la requête

Dans certains cas, l'analyse linguistique et l'analyse statistique expliquées ci-dessus ne suffisent pas à établir un lien entre la requête et les documents pertinents. Dans ce cas, il est nécessaire d'ajouter un élément sémantique au processus sur la base de la requête originale afin d'inférer ce que recherche l'utilisateur. Il s'agit donc d'étendre la requête posée en utilisant d'autres formulations de l'idée qui y est exprimée pour que soient retrouvés les documents susceptibles d'être pertinents. Cette reformulation peut aussi bien être dans la

même langue (synonymes, hyponymes, etc.) que dans des langues différentes, et pour ce faire, le système du LIC2M utilise des dictionnaires de reformulation monolingue et bilingue.

2.4 Calcul de la proximité sémantique

Le comparateur sert à calculer la proximité sémantique entre la requête et les documents indexés à partir des mots communs (mots de l'intersection requête/documents). Ce comparateur consiste, d'une part, à identifier les meilleures intersections requête/documents, et d'autre part, à regrouper les intersections identiques et leur attribuer un poids. Le résultat est présenté sous forme d'une liste de classes d'intersections dans un ordre croissant de poids. Les documents de la base sont indexés et stockés dans des fichiers inversés. On construit un index pour chacune des langues des documents constituant le corpus et on applique l'analyse linguistique pour les documents à indexer et pour les requêtes.

2.5 Résultats de la recherche cross-lingue

Le système de recherche cross-lingue d'information du LIC2M utilise le modèle booléen pondéré. Lorsque la requête est effectuée, les documents sont renvoyés groupés par classes, qui représentent la répartition des mots de la requête dans les bases de données. Chaque classe contient une liste de documents classés par ordre de pertinence. Par exemple, le moteur de recherche retourne 12 classes pour la requête إدارة موارد المياه "gestion des ressources en eau" (Tableau 1).

Classe	Termes de la requête	Nombre de termes de la requête
1	إدارة_موارد_مياه	1
2	موارد_مياه ,إدارة_موارد	2
3	مياه ,إدارة_وارد	2
4	إدارة ,موارد_مياه	2
5	إدارة_موارد	1
6	موارد_مياه	1
7	إدارة ,موارد ,مياه	3
8	إدارة ,مياه	2
9	إدارة ,موارد	2
10	موارد ,مياه	2
11	مياه	1
12	موارد	1

Tableau 1 : Classes retrouvées pertinentes pour la requête إدارة موارد المياه

3 Alignement de phrases basé sur la recherche cross-lingue

L'alignement de phrases à partir de textes bilingues basé sur la recherche cross-lingue d'information consiste à construire une base de données des phrases du texte cible ($Corpus_{Fr}$) et considérer chaque phrase du texte source ($Corpus_{Ar}$) comme une requête à cette base de données (Figure 1).

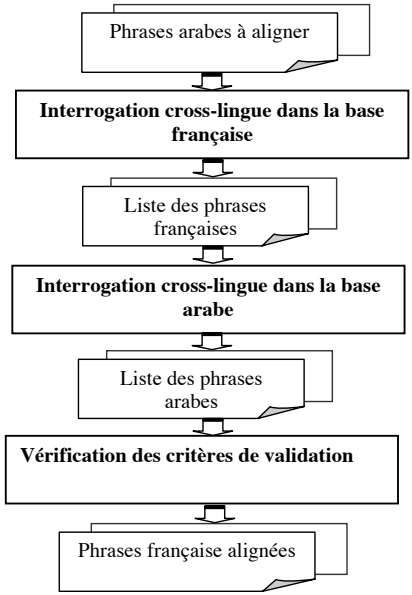

Figure 1 : Etapes de l'alignement de phrases

La validation de l'alignement est basée sur trois critères:

- Position de la phrase dans le document : L'alignement est validé si le rang (numéro d'ordre) de la phrase à aligner se situe dans une fenêtre de tolérance de 10 (rangs) par rapport à la dernière phrase alignée. Cette valeur a été établie expérimentalement.

- Le nombre de termes communs entre la phrase source et la phrase cible (intersection sémantique) doit représenter plus de 50% du nombre des termes de la phrase source.

- Le rapport entre la taille (exprimée en nombre de caractères) de la phrase cible et la taille de la phrase source doit être supérieur à 1.1. La valeur de ce rapport a été fixée expérimentalement. Elle repose sur l'idée qu'*une phrase aura tendance à être traduite par une phrase longue si elle est longue, et par une phrase courte si elle est courte*.

Le processus d'alignement se déroule en quatre étapes :

1. Alignement 1-1 Exact Match: L'objectif est d'obtenir un alignement avec une précision maximale en utilisant les trois critères de validation.
2. Alignement 1-2: Cet alignement consiste à trouver pour la phrase à aligner deux phrases en langue cible en utilisant comme référence le rang de la phrase précédente déjà alignée. Nous utilisons pour valider cet alignement les deux premiers critères.
3. Alignement 2-1: L'objectif de cet alignement est de trouver pour les deux phrases en langue source suivant une phrase déjà alignée une phrase en langue cible en utilisant comme référence le rang de la phrase précédente déjà alignée. Cet alignement est validé par les deux premiers critères.
4. Alignement 1-1 Fuzzy Match: Cette étape consiste à aligner la phrase source avec la première phrase cible de la première classe retournée par le moteur de recherche cross-lingue. Cet alignement n'utilise aucun critère de validation.

Nous décrivons ci-après l'algorithme de l'aligneur 1-1 Exact Match qui constitue la base des autres aligneurs. Cet algorithme utilise les fonctions de l'API du moteur de recherche:

```
- PerformCrosslinguageSearch(Requête, Corpus, Langue source, Langue cible):
retourne l'ensemble des classes retrouvées pertinentes pour la question "Requête"
dans la base de données textuelles "Corpus". Chaque classe est composée d'un
ensemble de phrases dans la langue cible.

- GetNumberOfCommonWords(Classe): retourne le nombre de termes communs entre la
phrase source et la phrase cible (intersection sémantique).

- GetNumberOfWords(Phrase): retourne le nombre de mots pleins d'une phrase.

- GetNumberOfCharacters(Phrase): retourne le nombre de caractères d'une phrase.

Fonction GetExactMatchOneToOneAlignments(Corpus_Ar, Corpus_Fr)
 Pour chaque phrase arabe P_jAr (de rang j) ∈ Corpus_Ar faire
  C_Fr = PerformCrosslinguageSearch(P_jAr, Corpus_Fr, Ar, Fr)
  R = 0; Initialisation du rang de la dernière phrase alignée.
  Pour chaque classe C_lFr ∈ C_Fr faire
   Pour chaque phrase française P_mFr (de rang m) ∈ C_lFr faire
    C_Ar = PerformCrosslinguageSearch(P_mFr, Corpus_Ar, Fr, Ar)
    Pour chaque classe C_qAr ∈ C_Ar faire
     Pour chaque phrase arabe P_qAr ∈ C_qAr faire
      Si P_qAr = P_jAr alors
       NM_Fr = GetNumberOfCommonWords(C_lFr); NM_Ar = GetNumberOfWords(P_jAr);
       NC_Ar = GetNumberOfCharacters(P_jAr); NC_Fr = GetNumberOfCharacters(P_mFr)
       Si (NM_Fr >= NM_Ar/2) et (R – 5 <= m <= R + 5) et (NC_Fr = (1.1)* NC_Ar) alors
        La phrase P_mFr est l'alignement de la phrase P_jAr; R = m
       Fin Si
      Fin Si
     Fin Pour
    Fin Pour
   Fin Pour
  Fin Pour
 Fin Pour
Fin Fonction
```

Par exemple, pour aligner la phrase arabe [4/30] (Phrase de rang 4 dans une base de données contenant 30 phrases) " في ايطاليا ادت طبيعة الاشياء الى اقناع غالبية الناخبين في طريقة غير مرئية بأن زمن الاحزاب التقليدية قد بلغ نهايته", l'aligneur 1-1 Exact Match procède comme suit:

- La phrase arabe est considérée comme une requête dans la base de données des phrases françaises en utilisant le moteur de recherche cross-lingue. Les phrases retrouvées pertinentes des deux premières classes sont illustrées dans Tableau 2.

- Les réponses de l'interrogation cross-lingue montrent que la phrase française [4/36] est un bon candidat pour l'alignement. Pour confirmer cet alignement, nous utilisons cette phrase comme une requête à la base de données des phrases arabes. Les phrases retrouvées pertinentes pour cette phrase sont groupées dans deux classes dans Tableau 3.

Classe	Nombre de phrases retrouvées	Phrases retrouvées
1	1	[4/36] En Italie, l'ordre des choses a persuadé de manière invisible une majorité d'électeurs que le temps des partis traditionnels était terminé
2	3	[32/36] Au point que, dès avant ces élections, un hebdomadaire britannique, rappelant les accusations portées par la justice italienne contre M. Berlusconi, estimait qu'un tel dirigeant n'était pas digne de gouverner l'Italie, car il constituait un danger pour la démocratie et une menace pour l'Etat de droit [34/36] Après le pitoyable effondrement des partis traditionnels, la société italienne, si cultivée, assiste assez impassible (seul le monde du cinéma est entré en résistance) à l'actuelle dégradation d'un système politique de plus en plus confus, extravagant, ridicule et dangereux [36/36] Toute la question est de savoir dans quelle mesure ce modèle italien si préoccupant risque de s'étendre demain à d'autres pays d'Europe

Tableau 2 : Phrases retrouvées pertinentes pour la phrase arabe à aligner [4/30]

Classe	Nombre de phrases retrouvées	Phrases retrouvées
1	1	[4/30] في ايطاليا ادت طبيعة الاشياء الى اقناع غالبية الناخبين في طريقة غير مرئية بأن زمن الاحزاب التقليدية قد بلغ نهايته
2	3	[26/30] يشكل هؤلاء الرجال اكثر ثلاثية مثيرة للسخرية والتقزز في اوروبا، الى درجة ان احدى المجلات الاسبوعية البريطانية اعتبرت في معرض استعادتها للاتهامات القضائية الموجهة الى السيد برلوسكوني قبل هذه الانتخابات ان مسؤولا من هذا النوع "ليس جديرا بحكم ايطاليا" وانه يمثل "خطرا على الديموقراطية" وعلى ""دولة القانون [28/30] وقد تبينت صحة هذه التوقعات المتشائمة، فبعد الانهيار المثير للشفقة للاحزاب التقليدية، شهد المجتمع الايطالي المعروف بثقافته ومن دون ان يبدي حراكا (باستثناء قطاع السينما الذي لجأ الى المقاومة) التدهور الراهن لنظام سياسي يعاني المزيد من الغموض والشطط والسخف والخطورة [30/30] وكل المسألة تكمن في معرفة الى اي مدى يمكن هذا النموذج الايطالي المثير للقلق ان ينتشر غداً في بلدان اوروبية اخرى

Tableau 3 : Phrases retrouvées pertinentes pour la phrase française [4/36]

La première phrase proposée par l'interrogation cross-lingue correspond à la phrase initiale à aligner et plus de 50% des mots sont communs entre les deux phrases. De plus, le rapport

entre la phrase française et la phrase arabe est supérieur à 1.1 et les positions des deux phrases dans les deux bases de données sont identiques. Par conséquent, l'aligneur 1-1 Exact Match considère la phrase française [4/36] comme la traduction de la phrase arabe [4/30].

4 Résultats et Discussions

Pour mener nos expérimentations et être en mesure de calculer la performance de notre aligneur de phrases, nous avons utilisé le corpus MD (Monde Diplomatique) de la campagne ARCADE II (Chiao et al., 2006). Le corpus contient 5 textes arabes (244 phrases) alignés avec 5 textes français (283 phrases).

Pour évaluer l'aligneur au niveau de la phrase, nous avons utilisé les mesures suivantes :

$$\text{Précision} = \frac{|A \cap A_r|}{|A|} \text{ et Rappel} = \frac{|A \cap A_r|}{|A_r|}$$

A correspond à l'ensemble des alignements fournis par l'aligneur et A_r correspond à l'ensemble des alignements corrects.

Les résultats d'alignement sont illustrés dans Tableau 4 et montrent que la précision est autour de 97% et le rappel est autour de 93%. Ces résultats ne prennent pas en compte les alignements partiellement corrects (Alignement 1-1 Fuzzy Match).

Texte parallèle	Précision	Rappel
1	0,969	0,941
2	0,962	0,928
3	0,985	0,957
4	0,983	0,952
5	0,966	0,878

Tableau 4 : Résultats d'alignement au niveau de la phrase du corpus MD

Par ailleurs, l'analyse de ces résultats montre, d'une part, que l'alignement est correct même si le corpus n'est pas parfaitement parallèle, et d'autre part, que la précision dépend fortement des mots discriminants présents dans les phrases source et cible.

5 Conclusion et Perspectives

Nous avons proposé une nouvelle approche pour aligner les phrases d'un corpus parallèle. Cette approche est basée sur une recherche cross-lingue d'information et combine plusieurs sources d'information (dictionnaire bilingue, longueurs des phrases, numéros d'ordre des phrases dans le corpus parallèle). Les résultats que nous avons obtenus montrent des valeurs correctes pour la précision et le rappel même lorsque le corpus n'est que partiellement parallèle. Nos travaux vont maintenant s'étendre, d'une part, à l'utilisation des structures

syntaxiques du corpus parallèle pour améliorer la performance de l'alignement de phrases, et d'autre part, au développement d'un outil d'aide à la traduction basé sur les textes bilingues alignés.

Références

A. ANDREEWSKY, J. P. BINQUET, F. DEBILI, C. FLUHR, B. POUDEROUX. (1981). Le traitement linguistique et statistique des textes et son application dans la documentation juridique. Actes du *Sixième Symposium sur l'Informatique Juridique en Europe*, Thessaloniki, Grèce.

BROWN P., LAI L., MERCIER L. (1991). Aligning Sentences in Parallel Corpora. Actes de *ACL-1991*.

CHIAO Y. C., KRAIF O., LAURENT D., NGUYEN T., SEMMAR N., STUCK F., VÉRONIS J., ZAGHOUANI W. (2006). Evaluation of multilingual text alignment systems: the ARCADE II project. Actes de *LREC-2006*.

DEBILI F. SAMMOUDA E. (1992). Appariement des Phrases des Textes Bilingues. Actes du *14th International Conference on Computational Linguistics*.

DEBILI F., ZOUARI L. (1985). Analyse morphologique de l'arabe écrit voyellé ou non fondée sur la construction automatique d'un dictionnaire arabe. Actes de *Cognitiva-1985*.

FLUHR C., BISSON F., ELKATEB F. (2000). Parallel text alignment using cross-lingual information retrieval techniques. *Parallel text processing*. Kluwer, Boston.

GALE W.A. CHURCH K. W. (1991). A program for aligning sentences in bilingual corpora. Actes du *29th Annual Meeting of Association for Computational Linguistics*.

GAUSSIER E. (1995). *Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues*. Ph.D. Thesis, Paris VII University.

GREFENSTETTE G., SEMMAR N., ELKATEB-GARA F. (2005). Modifying a Natural Language Processing System for European Languages to Treat Arabic in Information Processing and Information Retrieval Applications. Actes de *ACL-2005,* 31-38.

MELAMED I. D. (1996). A Geometric Approach to Mapping Bitext Correspondence. Actes de *Conference on Empirical Methods in Natural Language Processing*.

SEMMAR N., ELKATEB-GARA F., LAIB M., FLUHR C. (2005). A Cross-language information retrieval system based on linguistic and statistical approaches. Actes du *Deuxième Congrès International sur l'Ingénierie de l'Arabe et l'Ingénierie de la Langue*.