



**HAL**  
open science

## Portage linguistique de l'arabe et vers l'arabe

Najeh Hajlaoui, Christian Boitet

► **To cite this version:**

Najeh Hajlaoui, Christian Boitet. Portage linguistique de l'arabe et vers l'arabe. Infol@ngues III : Informatique et langues, Feb 2009, Tunisie. 10 p. hal-00390852

**HAL Id: hal-00390852**

**<https://hal.science/hal-00390852>**

Submitted on 2 Jun 2009

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Portage linguistique de l'arabe et vers l'arabe

Najeh HAJLAOUI<sup>1</sup>, Christian BOITET<sup>2</sup>

<sup>1,2</sup>Laboratoire LIG, GETALP, Université Joseph Fourier,  
385 rue de la Bibliothèque, BP n° 53, 38041 Grenoble, cedex 9, France

Najeh.Hajlaoui@imag.fr

Christian.Boitet@imag.fr

**Résumé :** nous nous intéressons à la multilinguisation, ou « portage linguistique » (plus simple que la localisation) des services de gestion de contenu traitant des énoncés spontanés en langue naturelle, souvent bruités, mais contraints par la situation, et constituant toujours un « sous-langage » plus ou moins restreint.

Un service de ce type (soit App) utilise une représentation du contenu spécifique (RC-App) sur laquelle travaille le noyau fonctionnel. Le plus souvent, cette représentation est produite à partir de la langue « native » L1 par un extracteur de contenu (EC-App). Nous avons dégagé trois méthodes de portage possibles, et les avons illustrées par le portage en français d'une partie de CATS, un système de traitement de petites annonces en SMS (en arabe) déployé à Amman, ainsi que sur IMRS, un système de recherche de morceaux de musique dont l'interface native est en japonais et dont seule la RC est accessible. Il s'agit de : (1) localisation « interne », i.e. adaptation à L2 de l'EC donnant EC-App-L2 ; (2) localisation « externe », i.e. adaptation d'un EC existant pour L2 au domaine et à la représentation de contenu de App (EC-X-L2-App); (3) traduction automatique des énoncés de L2 vers L1.

Le choix de la stratégie est contraint par la situation traductionnelle : types et niveau d'accès possibles (accès complet au code source, accès limité à la représentation interne, accès limité au dictionnaire, et aucun accès), ressources disponibles (dictionnaires, corpus), compétences langagières et linguistiques des personnes intervenant dans la multilinguisation des applications.

Les trois méthodes ont donné de bons résultats sur le portage d'arabe en français de la partie de CATS concernant l'occasion automobile. En localisation interne, la partie grammaticale a été très faiblement modifiée, ce qui prouve que, malgré la grande distance entre l'arabe et le français, ces deux sous-langages sont très proches l'un de l'autre, une nouvelle illustration de l'analyse de R. Kittredge. La localisation externe a été expérimentée sur CATS et sur IMRS en adaptant au nouveau domaine considéré l'extracteur de contenu du français écrit initialement par H. Blanchon pour le domaine du tourisme (projet CSTAR/Nespole!), puis en changeant de langue pour IMRS (anglais, arabe).

Enfin, le portage par TA statistique a également donné de très bonnes performances, et cela avec un corpus d'apprentissage très petit (moins de 10.000 mots) et un dictionnaire complet. Cela prouve que, dans le cas de sous-langages très petits, la TA statistique peut être de qualité suffisante en partant de corpus 100 à 500 fois moins grands que pour de la langue générale.

**Mots clés :** énoncés spontanés, langue naturelle, e-commerce, portage linguistique, sous-langage, extraction de contenu, traduction statistique, corpus.

## Introduction

Une conséquence majeure résultant de la mondialisation que nous vivons aujourd'hui est l'importance croissante du multilinguisme, accentuée par l'utilisation d'Internet et le multimédia. Face à ce phénomène, les développeurs d'applications informatiques ne peuvent pas encore fournir des applications utilisables dans la langue maternelle de tous les utilisateurs. Les services et les applications sont rarement disponibles dans de nombreuses langues, et encore moins traitent des textes naturels rédigés de façon spontanée et pouvant contenir des erreurs, des abréviations, etc.

Dans ce cadre, nous nous intéressons à un problème plus précis, celui de la multilinguisation des applications de e-commerce traitant des énoncés spontanés en langue naturelle (LN). Ces applications extraient le « contenu pertinent » des énoncés en langue naturelle, le représentent dans un langage approprié (CRL), et traitent ensuite les objets obtenus.

Nous présentons d'abord le type d'applications qui nous intéresse. Ensuite, nous présentons nos approches de portage linguistique qui permettent de rendre disponible de telles applications dans d'autres langues cibles. Enfin, nous illustrons ce travail par un portage de l'arabe vers le français d'une première application, CATS (Daoud 2006) et par un portage du japonais vers l'arabe d'une deuxième application, IMRS (Kumamoto 2007).

## **I. Applications visées**

Nous nous intéressons à un type particulier de service de gestion de contenu, dans lequel les conditions suivantes doivent être satisfaites : pas de traduction, interface « libre », besoin « naturel » de multilinguïsation.

### **I.1 Pas de la traduction**

Dans cette étude, on ne s'intéresse pas aux applications commerciales comme e-Bay qui ne font pas de réel traitement de contenu, même si elles sont nombreuses. Ainsi, nous ne visons pas les applications qui font un stockage de la version originale des données dans une base de données, et qui, pour le besoin multilingue, font appel à un ou des systèmes de TA (traduction automatique).

Dans d'autres applications, on trouve une génération multilingue de données qui sont déjà traduites en plusieurs langues, comme la documentation technique ou les manuels d'utilisation de matériels techniques ou de logiciels informatiques. Par exemple, au Canada, les bulletins météo doivent être disponibles dans les deux langues française et anglaise, et ils sont traduits grâce au système Météo (Chandioux, 1988).

Sur le Web, les offres d'emploi, les petites annonces de vente et d'achat sont disponibles par traduction en plusieurs langues. Par exemple, le site très connu e-Bay est disponible dans plus d'une vingtaine de pays, et le choix d'un pays affiche le site correspondant dans la langue du pays sélectionné. Si l'on élargit le critère de recherche en précisant le lieu de l'objet (en choisissant par exemple « *dans le monde entier* » au lieu de « *en France uniquement* »), un objet situé dans un autre pays peut être disponible par traduction dans la langue de navigation.

### **I.2 Caractéristique : traitement du contenu**

Nous nous intéressons dans cette étude aux applications qui donnent de la valeur ajoutée par le traitement du contenu. C'est la caractéristique principale que nous imposons dans le choix de notre application. Un traitement de contenu peut avoir plusieurs catégories (extraction de contenu, résumé, catégorisation, recherche d'information, etc).

Les principaux types d'applications que nous visons sont :

- Catégorisation de documents divers : dépêches de l'AFP « Agence France-Presse », brèves des différentes bourses, messages de clients à un serveur de SAV (Service Après Vente).
- Extraction d'informations pour nourrir ou consulter une base de données : petites annonces ciblés, FAQ (Foire Aux Questions) intelligentes, indexation ciblée à un domaine/métier, jusqu'aux hotlines automatisées, rapports d'incidents par des utilisateurs.
- Traitement du Web, par exemple pour des études d'opinion.
- Traitement du Web pour la catégorisation de courriels, le résumé automatique, etc.
- Recherche d'informations.
- Commerce : achat et vente de produits, assistance à la vente, etc.

### **I.3 Besoin d'utiliser les interfaces libres**

Malgré le progrès et la multiplication des méthodes de TALN, elles sont très peu ou presque pas utilisées dans le genre d'applications décrites précédemment. Ce sont les menus et les formulaires qui sont plutôt utilisés pour répondre à des besoins commerciaux, qu'il s'agisse de catégorisation, de commerce, etc.

La navigation et la recherche par mots-clés pilotées par des menus dans la plupart des sites commerciaux ont pourtant des limitations importantes : elles tendent à accabler et frustrer des utilisateurs avec des interactions prolongées et rigides. Or on sait que l'intérêt de l'utilisateur pour un site particulier diminue exponentiellement avec l'augmentation du nombre de clics de souris (Huberman, Pirolli et al. 1998). Par conséquent, le raccourcissement du chemin d'interaction pour obtenir des informations utiles devient important.

Beaucoup de sites de e-commerce essaient de résoudre ce problème en fournissant des possibilités de recherche par mots-clés. Cependant, les moteurs de recherche par mots-clés exigent habituellement que les utilisateurs connaissent le jargon spécifique du domaine. La recherche par mots-clés ne permet pas à des utilisateurs de décrire avec précision leurs intentions, et d'une manière primordiale, il leur manque les bons mots et les bonnes expressions de recherche.

Les utilisateurs frustrés et incapables de suivre des étapes de remplissage de formulaires et de recherche sont obligés d'abandonner ce mode d'interaction pour passer à une communication directe avec le propriétaire du service, soit en téléphonant soit en se déplaçant. Cela affaiblit le caractère autonome du service proposé censé fonctionner sans l'intervention d'une personne, et coûte très cher en temps pour répondre aux clients.

En conséquence, l'utilisation d'interfaces « ouvertes » dans lesquelles l'utilisateur peut exprimer sa requête (recherche, mécontentement, description d'un problème technique...) d'une façon très spontanée et non contrôlée peut dépasser les limites et les difficultés rencontrées dans l'utilisation des mots-clés et des formulaires.

### **I.4 Importance des services multilingues**

Un service multilingue possède plusieurs avantages. Essentiellement, il peut être déployé dans de nombreux pays. Mais on rencontre alors des obstacles. Par exemple, la société CISCO (<http://www.cisco.fr/>) utilisait les systèmes de TA « tout terrain » de Systran pour traduire (mal) sa documentation, notamment en CJK (chinois, japonais, et coréen). Ses centres d'appel étaient plus que surchargés, et coûtaient tellement cher qu'elle a passé (vers 2001) un contrat avec Systran pour construire et suivre des versions spécialisées. Même si la qualité linguistique n'est toujours pas fameuse, l'utilité de cette spécialisation est indéniable, car la baisse de charge des centres d'appel a permis d'économiser largement plus que le coût de ce contrat.

Le site numéro 1 des petites annonces en France (<http://www.paruvendu.fr/>) est disponible en une seule langue (français). Cela empêche un grand nombre d'étrangers ne parlant pas le français d'utiliser ce service assez connu pour répondre à leurs besoins quotidiens.

Les deux exemples précédents montrent bien la nécessité de services multilingues : non seulement la multilinguisation est nécessaire dans la plupart des pays multilingues (Canada, Inde, US), mais elle l'est aussi dans des pays monolingues (France) à cause du tourisme et de la nouvelle mobilité.

## II. Portage linguistique

La multilinguisation ou le « portage linguistique » dont nous parlons ici n'est pas nécessairement une « localisation ». Une localisation implique une adaptation à un autre contexte culturel. Par contre, un portage linguistique doit seulement permettre l'accès dans une autre langue à un service de e-commerce, tel qu'il est et où il est.

Notre but est de trouver, pour multilingualiser de telles applications, des solutions simples, peu coûteuses en termes de ressources et de temps, efficaces, applicables sur le terrain, adaptées à la situation traductionnelle, et en réutilisant ce qui est disponible comme outils, ressources et capacités langagières et linguistiques.

### II.1 Exemple de type d'applications : extraction de contenu

L'extraction de contenu est rarement fondée sur une analyse complète des énoncés : on utilise le plus souvent des « grammaires locales » et un dictionnaire mettant en relation les termes du domaine et les symboles (concepts, attributs, relations) du CRL. L'architecture de ce type d'application est donc la suivante :

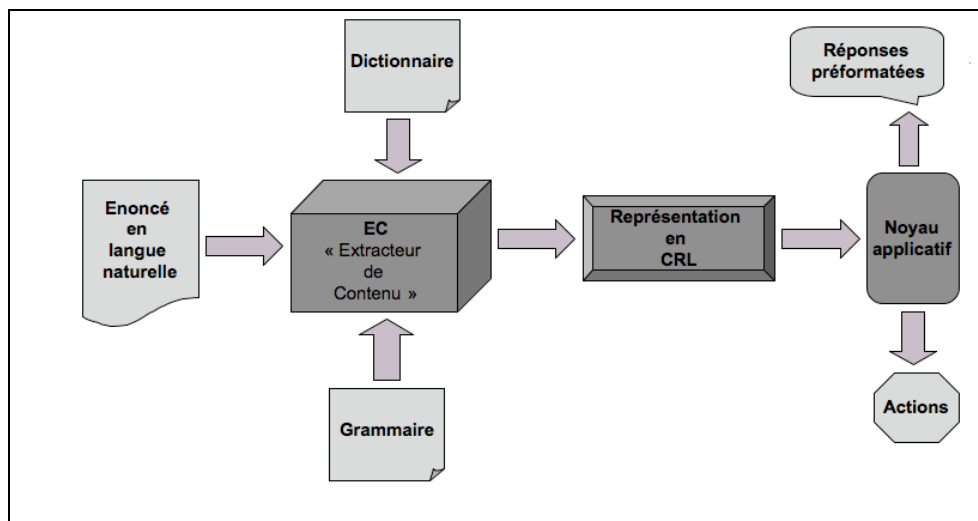


Figure 1 : architecture générale des applications de traitement du contenu d'énoncés en LN

La multilinguisation de telles applications est en fait un problème difficile, ce qui explique que très peu de services la proposent. La difficulté du problème dépend de deux facteurs relatifs à la situation traductionnelle :

- le niveau d'accès aux ressources des applications, pour lequel quatre situations se présentent : accès complet au code source, accès à la représentation interne des données manipulées uniquement, accès aux dictionnaires uniquement, et aucun accès.
- le niveau de compétence langagière et linguistique des intervenants pour la multilinguisation des applications. Il peut être défini par rapport à la langue source ou par rapport aux compétences linguistiques de l'équipe qui veut faire la localisation (spécialiste ou non de TALN).

### II.2 Approches possibles

Le choix de l'application, l'étude de la situation traductionnelle, l'étude du sous-langage traité, le choix du couple de langue et le savoir faire de l'équipe de portage dans le domaine sont les facteurs permettant de déterminer la ou les méthodes de portage possibles pour chaque application.

Trois méthodes sont possibles : portage interne, portage externe et portage par traduction.

### II.2.1 Portage interne

La première méthode consiste à adapter à la nouvelle langue L2 l'extracteur de contenu de l'application prévu pour la langue originale L1. La Figure 2 résume cette méthode de « portage interne ».

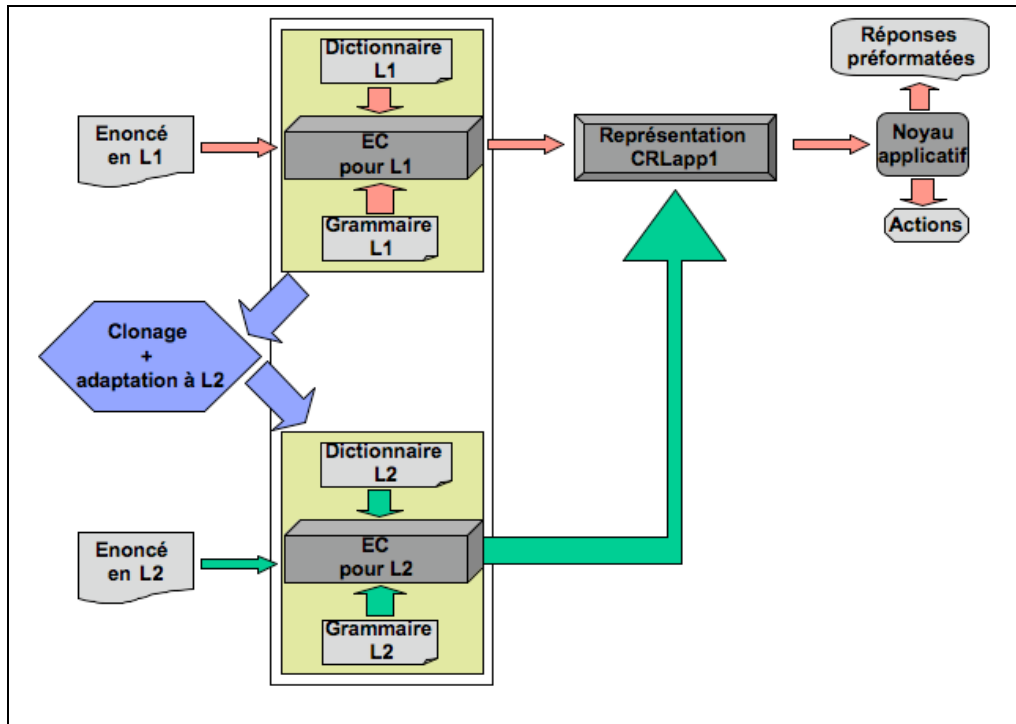


Figure 2 : méthode de portage interne

### II.2.2 Portage externe

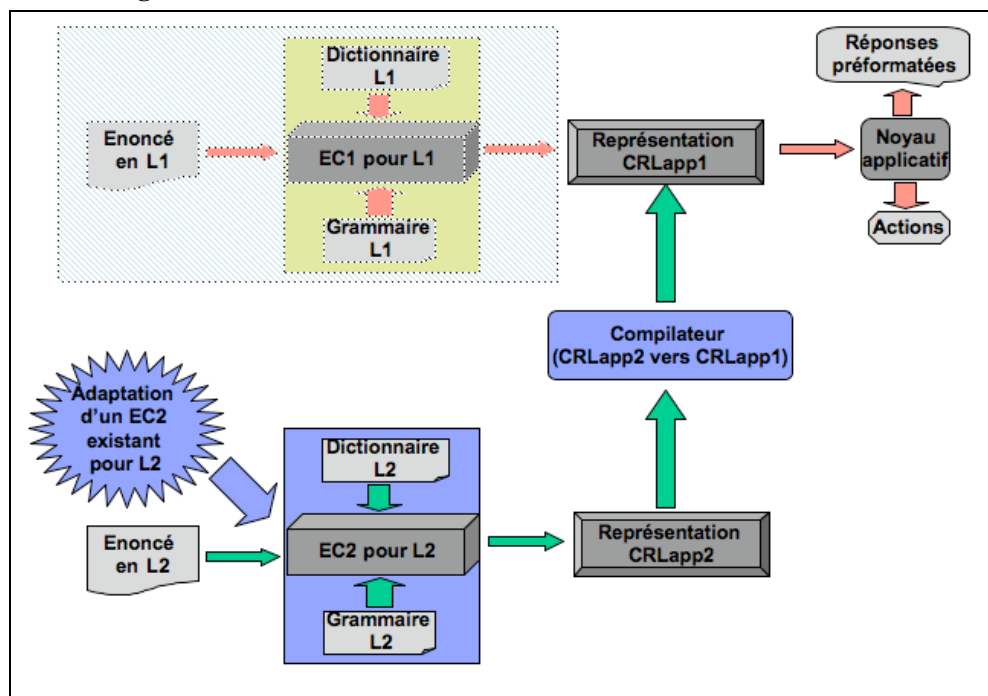


Figure 3 : méthode de portage externe

Une deuxième méthode consiste à adapter un extracteur de contenu existant, traitant la même langue cible L2, mais pour un autre domaine et/ou une autre tâche, et à transformer le résultat obtenu en une représentation de contenu de l'extracteur de contenu original. La Figure 2 résume cette méthode de « portage externe ».

### II.2.3 Portage par traduction

La troisième méthode consiste à traduire les énoncés de la nouvelle langue L2 vers la langue originale de l'application L1 et à utiliser l'extracteur de contenu de l'application originale. La Figure 4 décrit l'architecture de cette méthode de « portage par TA ».

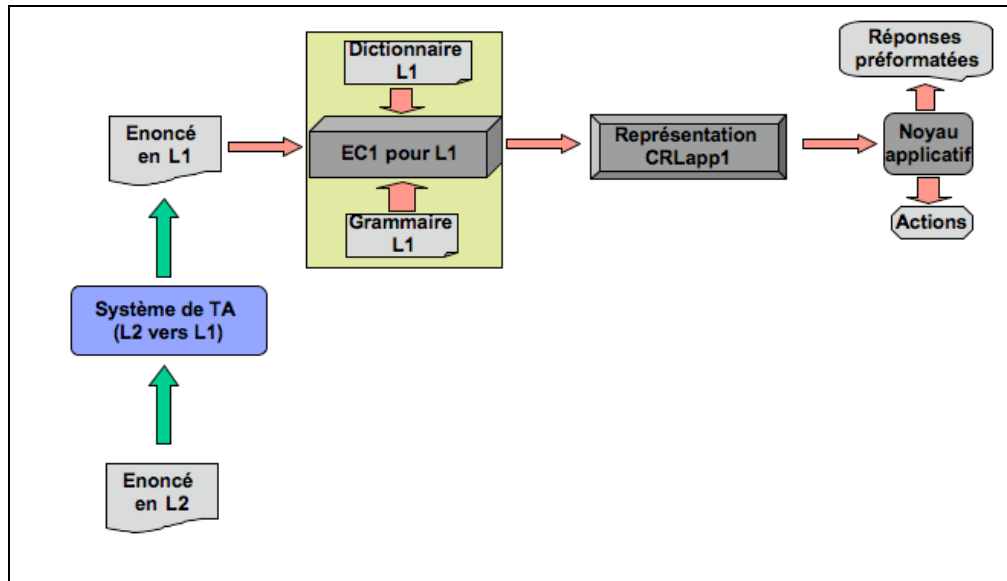


Figure 4 : méthode de portage par TA

La difficulté commune à toutes ces méthodes de portage est de trouver ou de construire les données de développement (corpus *fonctionnellement équivalent* dans la nouvelle langue et dictionnaire spécifique) .

## III. Portage linguistique de l'arabe et vers l'arabe

### III.1 Portage à partir de l'arabe

Nous avons expérimenté les trois précédentes approches de portage sur une première application, CATS « Classifieds Ads Transaction System », un système de petites annonces d'achat et de vente basé sur l'utilisation des SMS en arabe (Daoud 2006), et déployé en Jordanie par la société FastLink. La Figure 5 présente des exemples de SMS arabes traités par la version originale du système.

SMS envoyé	CRL-CATS obtenue	SQL
<p>مطلوب ميتسوبيش لانسر</p> <p>« Recherche une Mitsubishi Lancer »</p>	<pre>wan(saloon:00, wanted:00) mak(saloon:00, MITSUBISHI(country&lt;japan):06 mod(saloon:00, Lancer(country&lt;japan, make&lt;MITSUBISHI):0G)</pre>	<pre>select MsgContact from Cars where make = 'mitsubishi' and model = 'lancer' and maincat = 'saloon'</pre>
<p>للبيع لانسر 99 سعر 5000 دينار</p> <p>« A vendre une Lancer 99 prix 5000 dinars »</p>	<pre>sal(saloon:00, sale:00) mod(saloon:00, Lancer(country&lt;japan, make&lt;MITSUBISHI):06) yea(saloon:00, 99:0I) pri(saloon :00, 5000:0L)</pre>	<pre>Insert into cars (maincat, model, year, price, country, make) Values('saloon', 'lancer', ', '99', '5000', 'japan', 'mitsubishi')</pre>

Figure 5 : exemples de SMS traités par la version arabe de CATS

Pour le portage interne, l'auteur (D. Daoud) nous a donné accès à l'EC-CATS (Extracteur de Contenu de CATS) qu'il avait développé pour l'arabe, ce qui nous a permis d'appliquer cette méthode en adaptant cet extracteur au français (Hajlaoui and Boitet 2007).

Pour le portage externe, nous avons effectué une adaptation et une greffe de l'extracteur de contenu du français du projet Nespole!<sup>1</sup> « *NEgotiating through SPOken Language in E-commerce* » écrit par (Blanchon 2004) pour un système de TA de dialogues oraux concernant des réservations et des renseignements touristiques. Afin d'aboutir au même format que le format CRL-CATS généré par la version originale de CATS, nous avons construit un traducteur de format qui permet de transformer le format IF-CATS (Interchange Format) généré par cet EC vers le format original CRL-CATS (Hajlaoui, Daoud and Boitet 2008).

Pour le portage par traduction automatique, nous avons construit un système de TA statistique qui traduit des énoncés de la langue L2 (français) vers la langue L1 (arabe). L'arabe est donc langue cible pour la traduction et langue source pour le portage de CATS. Le résultat de la TA est soumis à l'extracteur de contenu original de l'application CATS. Nous ne disposions que d'un très petit corpus parallèle. Cependant, bien que la TA statistique soit réputée ne fonctionner assez bien que si l'on dispose de très grands corpus parallèles, le système que nous avons construit avec Pharaoh a produit des résultats satisfaisants, au sens où les descripteurs de contenu obtenus sont assez proches de ceux obtenus à partir des SMS correspondants en arabe. Il semble donc qu'on puisse se passer de très grands corpus pour utiliser efficacement la TA statistique sur des "sous-langages" très restreints : les traductions ne sont pas très "fluides", mais elles sont "adéquates", et ce même si les deux "langues-mères" des deux sous-langages considérés sont assez distantes.

Le Tableau 1 récapitule les pourcentages de portage de ces trois méthodes, calculés par une évaluation informationnelle sur le corpus limité « CorpusEvalFr200SMS ».

<sup>1</sup> <http://nespole.itc.it> : site Web du projet Nespole !, visité en 2006.



Portage par	Par rapport à la version originale			Par rapport à la version de référence		
	Minimum	Moyenne	Maximum	Minimum	Moyenne	Maximum
adaptation interne	95%	98%	100%	83%	91%	99%
adaptation externe	46%	77%	99%	41%	72%	99%
TA	85%	93%	98%	75%	86%	96%

Tableau 1 : récapitulatif des résultats d'évaluation informationnelle

Le Tableau 2 résume les résultats de l'évaluation générale sur tout le corpus « CorpusEvalFr1100SMS », obtenus par un calcul de distance entre les résultats d'extraction de contenu obtenus par chaque méthode de portage et ceux obtenus dans la version originale.

	Portage interne	Portage externe	Portage par SMT (TA stat)
F-mesures	0,718	0,540	0,901

Tableau 2 : récapitulatif des résultats d'évaluation générale

### III.2 Portage vers l'arabe

Nous avons expérimenté un portage linguistique vers l'arabe sur l'application IMRS (Impression-based Music-Retrieval System (Kumamoto 2007)), un système de recherche de musique basé sur un dialogue en langue naturelle en japonais. Le service visé consiste à permettre à un utilisateur de rechercher des morceaux de musique à partir d'impressions subjectives (dansant, calme, rythme plus rapide, etc.).

Les énoncés traités par le système sont des phrases ou fragments de phrase spontanés qui sont transformés en vecteurs de requête. Chaque vecteur est composé de 10 composantes. La valeur d'une composante est un nombre réel entre 1 et 7. Le symbole particulier « *nil* » veut dire « *don't care* ».

Par exemple, l'axe « *Happy – Sad* » est caractérisé par sept valeurs symboliques : « *very happy* », « *happy* », « *a little happy* », « *medium* », « *a little sad* », « *sad* », et « *very sad* », qui correspondent respectivement aux valeurs numériques 7.0, 6.0, 5.0, 4.0, 3.0, 2.0, et 1.0. Une requête pour trouver un morceau de musique qui donne une heureuse impression (*happy*) correspondra ainsi au vecteur : (*nil nil nil nil nil nil nil 6,0 nil nil*).

Adapter d'abord au français (puis à l'anglais et à l'arabe) le système IMRS, était une tâche différente du portage de CATS. En effet,

- nous ne pouvions pas adapter l'extracteur de contenu de l'application, car nous n'y avons pas accès, et, si nous l'avions eu, nous aurions dû faire faire le travail par quelqu'un connaissant le japonais et le français,
- nous n'avions pas de corpus parallèle français-japonais adapté, et ne pouvions pas en créer.

Il nous restait donc comme seule possibilité de créer un extracteur de contenu produisant un vecteur requête à partir d'un énoncé en français. Pour créer cet extracteur, nous avons appliqué la méthode de portage externe en adaptant le même extracteur de contenu du français que pour CATS à ce nouveau domaine, et en transformant la représentation obtenue (IFmusique) vers un vecteur de dimension 10.

Nous avons adopté les mêmes étapes d'analyse que pour le second démonstrateur du projet Nespole! (Blanchon 2004) et la même démarche d'adaptation que celle utilisée pour CATS (Hajlaoui, Daoud and Boitet 2008). Comme pour CATS, le travail d'adaptation le plus important a été effectué au niveau de l'étape d'instanciation des arguments liés au domaine. Ainsi, de la même façon, nous avons procédé à la recherche des arguments un par un en consommant la chaîne en entrée. Par la suite, nous avons utilisé une sortie standard pour tous les arguments, qui contient directement un vecteur requête de dimension 10. Si l'argument est trouvé, sa valeur est affichée, sinon elle est remplacée par la chaîne `nil` qui veut dire « *don't care* ».

Nous avons expérimenté sur une centaine d'énoncés (créés manuellement) et voici quelques exemples de résultats pour l'arabe, le français et l'anglais :

```

Exemple_Ar 1 : أريد قطعة موسيقية جد هادئة //je veux un morceau de musique
très calme
IFMusique_Ar 1:{c:give-information+service(service=music, musique-
spec=(nil 7,0 nil nil nil nil nil nil nil nil))}

Exemple_Ar 2 : أريد قطعة موسيقية قليلة الضجيج //je veux un morceau de
musique un peu bruite
IFMusique_Ar 2:{c:give-information+service(service=music, musique-
spec=(3,0 nil nil nil nil nil nil nil nil))}

Exemple_Fr 1:je veux un morceau de musique calme et très solennel
IFMusique_Fr 1:{c:give-information+disposition+service
(disposition=(desire, who=i), service=music,
musique-spec= (nil 6,0 nil nil 7,0 nil nil nil nil nil))}

Exemple_Fr 2:je veux un morceau de musique assez fort et clair
IFMusique_Fr 2:{c:give-information+disposition+service
(disposition=(desire, who=i), service= music,
musique-spec=(3,0 nil nil 6,0 nil nil nil nil nil nil))}

Exemple_En 1:I want a calm and very solemn music
IFMusique_En 1:{c:give-
information+disposition+service(service=music,
music-spec=(nil 6,0 nil nil 7,0 nil nil nil nil nil))}

Exemple_En 2:I want a little noisy and bright music
IFMusique_En 2:{c:give-information+disposition+service
(service=music, music-spec=(3,0 nil nil 6,0 nil nil nil nil nil
nil))}

```

Figure 6 : exemple de résultats sur IMRS pour l'arabe, le français et l'anglais

Ce travail a nécessité beaucoup moins de temps que celui passé pour la première adaptation de l'IF du domaine du tourisme à celui de l'automobile. Nous avons passé environ une semaine pour chaque langue. Cela revient essentiellement à la maîtrise préalable du code de l'extracteur de contenu de Nespole! (environ 38000 lignes en Tcl/Tk), et aussi à la taille du vocabulaire du domaine de la musique, qui est très petit devant celui de l'automobile.

Malheureusement, nous n'avons pas pu avoir accès au système complet, et n'avons donc pas pu évaluer les résultats d'extraction trouvés.

Ce travail montre aussi que rien n'empêche de partir d'un même extracteur d'une langue donnée (le français ici) pour l'adapter à la fois à un nouveau

domaine, et à de nouveaux sous-langages, du français (« natif »), de l'anglais et de l'arabe.

## Conclusion

Nous avons proposé et expérimenté trois méthodes de portage linguistique d'applications de gestion de contenu de et vers une langue très importante comme l'arabe.

L'apport théorique principal de ces travaux consiste en la démonstration ou la vérification des deux hypothèses suivantes :

- Bien que le traitement des énoncés spontanés soit difficile, le passage d'un sous-langage à son correspondant dans une autre langue est parfois très simple. C'est une confirmation de l'hypothèse de Kittredge (proximité grammaticale de deux sous-langages analogues, même si leurs deux langues-mères sont distantes).
- La traduction statistique peut « bien marcher » pour l'extraction de contenu en utilisant un corpus de développement très petit mais formé de paires d'énoncés corrects (SL<sub>1</sub>, SL<sub>2</sub>), et un dictionnaire bilingue complet.

## Références

- Blanchon, H. (2004). Comment définir, mesurer et améliorer la qualité, l'utilisabilité et l'utilité des systèmes de TAO de l'écrit et de l'oral - Une bataille contre le bruit, l'ambiguïté, et le manque de contexte. Habilitation à diriger les recherches. Université Joseph Fourier. Grenoble, France. Décembre 2004. 356 p.
- Chandioux, J. (1988). *10 ans de METEO (MD)*. Traduction Assistée par Ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires. Paris. Observatoire Francophone des Industries de la Langue (OFIL). A. Abbou, éd. pp. 169-173.
- Daoud, D. M. (2006). It is necessary and possible to build (multilingual) NL-based restricted e-commerce systems with mixed sublanguage and content-oriented methods. Thèse. Université Joseph Fourier. Grenoble, France. September 23, 2006. 296 p.
- Hajlaoui, N. (2008). *Multilinguisation de systèmes de e-commerce traitant des énoncés spontanés en langue naturelle*. Thèse. Université Joseph Fourier. Grenoble, France. Septembre 2008. 318 p.
- Hajlaoui, N. and C. Boitet (2007). *Portage linguistique d'applications de gestion de contenu*. Proc. TOTH 2007 Conférence sur Terminologie & Ontologie : Théories et Applications. Annecy, France. 1er juin 2007. 13 p.
- Hajlaoui, N., D. M. Daoud, et al. (2008). *Methods for porting NL-based restricted e-commerce systems into other languages*. Proc. LREC 2008. Marrakech, Morocco. June 26-31, 2008. 7 p.
- Huberman, B., P. Pirolli, et al. (1998). *Strong Regularities in World Wide Web Surfing*. Science. in Science. Vol. 280. pp. 95-97.
- Kittredge, R. (1982). *Variation and Homogeneity of Sublanguages*. in Sublanguage - Studies of Language in Restricted Semantic Domains. Walter de Gruyter. Berlin / New York. 20 p.
- Kumamoto, T. (2007). *A Natural Language Dialogue System for Impression-based Music-Retrieval*. Proc. CICLING-07 (Computational Linguistics and Intelligent Text Processing). Mexico. February 12-24, 2007. 12 p.