TA statistique à petits corpus pour de petits sous-langages

Najeh HAJLAOUI, Christian BOITET
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 avons appliqué un système de TA statistique au "portage linguistique" de l'arabe au français de CATS, un système traitant le contenu de brefs messages spontanés en langue naturelle (SMS). Il s'agit d'un "sous-langage" très restreint. Nous ne disposions que d'un très petit corpus parallèle, augmenté d'un dictionnaire bilingue assez complet lié à l'application choisie (petites annonces en occasion automobile). 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.

Mots-clés : sous-langage, langue générale, langue standard, énoncés spontanés, traduction statistique, extraction de contenu.

Introduction

Les chercheurs du groupe TAUM à l'UdM (Université de Montréal) furent les premiers à se rendre compte de la relative facilité de construction de certains systèmes de TALN, et de leur grande qualité, quand on pouvait les limiter à des « sous-langages ». Après avoir connu une « bonne surprise » avec la traduction de bulletins météo (système

TAUM-météo)[1], le groupe TAUM a cherché longtemps (en vain d'ailleurs[2]) d'autres sous-langages aussi « faciles » pour la méthode employée (programmation « experte » reposant sur une étude précise du sous-langage en question et sur la mise en œuvre d'heuristiques adaptées).

Cela conduisit les linguistes du groupe TAUM (surtout R. Kittredge et J. Lehrberger) à approfondir la notion de sous-langage, introduite par Z. Harris en 1968, pour la rendre opérationnelle. Beaucoup de chercheurs les ont suivis dans cette voie, et ont montré l'importance de la notion de sous-langage dans le traitement du texte d'un langage naturel amélioré ou simplifié par l'utilisation de restrictions lexicales, syntaxiques ou sémantiques spécifiques (Kittredge and Lehrberger 1982a), (Grishman and Kittredge 1986), (Slocum 1986), (Biber 1993), (Sekine 1994). Dans ce dernier article, intitulé « A New direction for Sublanguage NLP », Satoshi Sekine montre de façon convaincante que la restriction (explicite ou implicite) à des sous-langages « assez restreints » conduit en général au succès : on arrive à construire des systèmes très performants avec un investissement très raisonnable en temps humain de spécialistes et en ressources de calcul (temps, place). Il cite lui aussi le cas du système TAUM-METEO.

Nous présentons dans la première partie quelques définitions possibles du terme *sous-langage* et un exemple de sous-langage réel. Dans la deuxième partie, nous décrivons quelques méthodes de portage

---

[1] Ce système fut construit par le groupe TAUM de l'UDM en 1975-76 (Isabelle 1984), (Chandioux 1988). Il fut mis en service opérationnel à Environnement Canada le 24 mai 1977 par la société J. Chandioux Conseil. C'est un système de traduction automatique qui marche extrêmement bien pour le sous-langage des bulletins météo (mais pas pour ceux des situations ou des avertissements météo !). Il traduit environ 20 M mots/an d'anglais en français et 10 M mots/an dans l'autre sens, avec une qualité liée à la tâche de plus de 97 % (moins de 3 opérations d'édition pour 100 mots traduits).

[2] NTT a trouvé une application de ce type, la traduction en anglais des brèves ("flash reports") du Nikkei (bourse de Tokyo), et développé pour cela le système ALTFlash, totalement automatique, de grande qualité, et « bimoteur » (système à patrons avec en secours une version spécialisée du système général ALT/JE).

linguistique d'applications traitant des énoncés spontanés en langue naturelle dont nous détaillons, dans la dernière partie, le portage par TA statistique, et son efficacité, au moins dans un cas de sous-langage très petit et restreint à une tâche, même si on ne dispose que d'un dictionnaire bilingue assez complet et d'un petit corpus parallèle.

1. Sous-langage naturel

1.1. Selon Zellig Harris

Plusieurs définitions pour le terme « sous-langage » ont été données. Il semble que la première a été proposée par Zellig Harris (Harris 1968) : « *Certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it.* »

Autrement dit,
« Un sous-ensemble strict d'une langue peut être fermé pour un sous-ensemble des opérations définies dans la langue, et ainsi en constituer un sous-langage. »

Cette définition semble à première vue incorrecte, car les phrases d'un « sous-langage » ne sont souvent pas des phrases (correctes) de la « langue standard », dont on suppose que parle un linguiste, et alors on ne pourrait pas parler de « sous-ensemble » au sens usuel.

Par exemple, il est acceptable dans un article de biochimie de dire « *The polypeptides were washed in hydrochloric acid* », mais pas « *hydrochloric acid was washed in polypeptides* ».

Comme Z. Harris savait parfaitement ce qu'est un sous-ensemble d'un ensemble, nous sommes conduits à admettre qu'il entendait par le terme « langue » une extension du terme « langue standard ». Nous utiliserons donc le terme « langue standard » pour désigner l'ensemble des énoncés d'une communauté linguistique formés d'une façon « correcte » par rapport à la grammaire et au vocabulaire usuels, tels qu'enseignés dans les cours de langue, et nous appellerons « langue générale » l'union d'une langue standard et de toutes ses variantes (jargons, langues de spécialité, parlers régionaux, langages « techniques », et langages « sécrétés » par des contextes socioprofessionnels).

Dans la définition précédente, assez générale, Harris ne dit pas de quelles opérations il parle. Mais il propose ensuite une définition « inductive » plus précise : un sous-langage SL est le plus petit ensemble contenant une base B et fermé (stable) par un ensemble de règles R.

SL = <B, R>, où

- la base B est un ensemble « noyau » d'énoncés ou schémas d'énoncés observés ;
- les règles R sont des règles de transformation comme la passivation, l'extraposition, l'interrogation, la mise au passif, à l'impersonnel, à l'interrogatif, ou simplement à un autre temps ou un autre mode, etc.

Un énoncé du sous-langage est donc dans le "noyau", ou bien il résulte d'un énoncé du sous-langage par une transformation de R. Par exemple, si « *The enzyme activated the process.* » est dans le sous-langage, et si la passivation est une des transformations permises, « *The process was activated by the enzyme.* » le sera aussi.

Cette définition est difficilement utilisable en pratique, car elle ne fournit pas de moyen opérationnel pour identifier le noyau et les règles caractérisant un sous-langage observé.

1.2. Définition selon l'usage

Une deuxième définition a été donnée par Bross et autres (Bross, Shapiro et al. 1972) :

« *Informally, we can* define *a sublanguage as the language used by a particular community of speakers, say, those concerned with a particular subject matter or those engaged in a specialized occupation.* »

Autrement dit, un sous-langage est l'ensemble des énoncés susceptibles d'être prononcés par une communauté (de communication) en certains temps et certains lieux.

Grishman et Kittredge (Grishman and Kittredge 1986), puis Deville (Deville 1989), définissent aussi un sous-langage comme une forme spécialisée d'une langue naturelle employée dans un domaine ou un thème particulier.

Cette définition est observationnelle et expérimentale, et prend directement en compte un contexte d'usage particulier. C'est celle qui a été utilisée dans le projet TAUM-METEO (1972-1973) et pour des manuels de maintenance d'avions dans le cadre du projet TAUM-AVIATION (1974-1981) et du PN-TAO (Projet National de TAO, 1982-87) en France.

À titre d'exemples de sous-langages, on peut citer les bulletins METEO, les manuels de maintenance d'un avion, les articles scientifiques concernant la pharmacologie, les rapports de radiologie, les annonces immobilières, etc.

Un sous-langage est alors caractérisé par un vocabulaire spécialisé, une sémantique restreinte, et dans beaucoup de cas une syntaxe spécialisée. Ainsi, les prépositions et articles normalement obligatoires peuvent être omis. Exemple : « *trappe visite réservoir avant gauche », « vent fort lac Saint-Jean », « orienté objet* ».

Cette définition a été précisée par Kittredge de la façon suivante.

Un sous-langage est un sous-ensemble d'une langue :

- qui fait référence à un *domaine particulier* ou à une famille de domaines liés,
- dont l'ensemble des phrases et des textes reflète *l'usage d'une communauté* de personnes ayant en commun des connaissances élaborées du domaine,
- qui a les propriétés fondamentales d'un *système linguistique* : consistance, complétude, économie d'expression, etc.,
- qui est *maximal* par rapport au domaine (il n'y en a pas de plus grand qui possède ces propriétés).

Type de langue / Genre de langue	**Langue générale**	**Langue standard**	**Sous-langage**
Textes	Énoncés corrects spontanés	Énoncés corrects	Énoncés spontanés
Grammaire	Usuelle + spécifique	Usuelle	Spécifique
Vocabulaire	Usuel + restreint	Usuel	Restreint

Tableau 1 : types de langues et caractéristiques associées

Le Tableau 1 résume les trois différents types de langue : langue générale, langue standard, et sous-langage.

1.1. Exemple : sous-langage de l'arabe des SMS en occasion automobile

CATS est une application de e-commerce déployée en Jordanie sur le réseau FastLink (Daoud, 2006). Elle traite des petites annonces envoyées par SMS et concernant l'occasion automobile (Cars), l'immobilier à Amman (RealEstate), l'emploi (Jobs), et autres (Misc). Elle permet de "poster" des petites annonces et de mettre en contact les personnes susceptibles d'être intéressées (Daoud, 2005). Voici quelques exemples de tels SMS, avec une traduction en français.

مطلوب سياره هونداي موديل 97 والسعر مابين 3500 الى 3750	Recherche voiture Honda, modèle 97, prix entre 3500 et 3750
مطلوب سياره سبور	Recherche voiture sport
اريد سياره مرسيدس موديل 82 لون ابيض	Je veux une voiture Mercedes modèle 82 couleur blanche

Tableau 2: Exemples de SMS arabe

Pour le domaine de l'automobile (Cars), la taille du vocabulaire utilisé est d'environ 638 entrées principales. Comme il comprend des mots étrangers translittérés, éventuellement de plusieurs façons, on y ajoute des variantes, dites entrées secondaires. Voici quelques exemples d'entrées.

Entrée arabe	**Principale /secondaire**	**UW (notation du concept dans Cars)**	**en français**
الفا روميو	P	ALFA ROMEO(country>Italy,country>europe)	Alfa Romeo
الفاروميو	S	ALFA ROMEO(country>Italy,country>europe)	Alfaromeo
روميو	S	ALFA ROMEO(country>Italy,country>europe)	Romeo
A3	P	A3(country>germany,country>europe,make>AUDI)	A3
اي3	S	A3(country>germany,country>europe,make>AUDI)	a3

Tableau 3 : Entrées du dictionnaire de CATS (arabe)

Les énoncés sont très simples et très courts (zone sombre dans la figure suivante). Si l'on parcourt un corpus de tels SMS, on observe une convergence grammaticale très rapide, mais une convergence lexicale

moins rapide, à cause des nouveaux motifs qui peuvent apparaître après un certain temps.

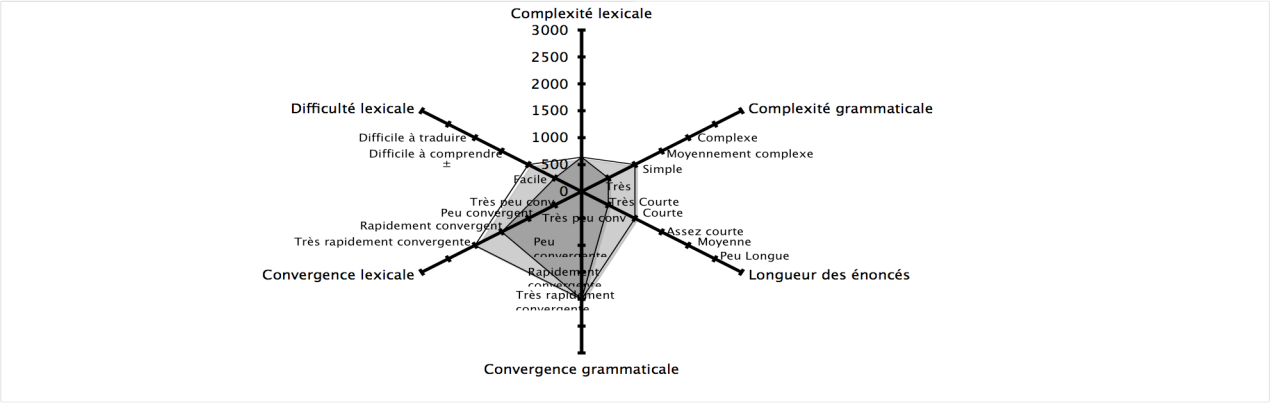

Figure 1 : analyse du sous-langage de l'automobile dans CATS

On peut aussi avoir des phrases simples et courtes (zone claire dans la figure précédente avec une convergence grammaticale et une convergence lexicale très rapides.

2. Multilinguïsation d'applications traitant des énoncés spontanés en langue naturelle

2.1. Problème et solutions possibles

Nous nous intéressons plus généralement à la multilinguïsation, 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. Tout service de ce type (soit App) est muni d'un extracteur de contenu (EC-App) produisant une forme interne spécifique (CRL-App) à partir de la langue "native" L1. Nos étapes sont les suivantes :

- Choix de l'application à porter et choix des couples des langues.
- Étude de la situation traductionnelle.
- Choix de la ou des méthodes de portage linguistique possibles, en fonction du résultat de l'étape précédente.

- Réalisation du portage linguistique.
- Évaluation du portage.

Nous avons illustré cette approche par le portage de la partie *Cars* de CATS. L'étude de la situation traductionnelle associée nous a menés à définir et expérimenter trois stratégies de portage pour ce couple de langues assez distant (arabe-français) : (1) *localisation "interne",* i.e. adaptation à L2 de l'extracteur de contenu (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* 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 intervenants pour la multilinguïsation des applications.

Les trois stratégies ont été expérimentées et ont donné de bons résultats sur le portage de l'arabe au français de la partie *Cars* de CATS (Hajlaoui, 2007).

2.2. Localisation interne

SMS en français	**CRL-CATS obtenue**
recherche voiture OPEL VECTRA	[S] wan(saloon:0A, wanted:00) mak(saloon:0A, OPEL(country>germany,country>europe):0I) mod(saloon:0A, Vectra(country>germany,country>europe,make>OPEL):0N) [/S]
recherche à l'achat NISSAN Sunny modèle 92 à 95	[S] wan(saloon:00, wanted:00) mak(saloon:00, NISSAN(country>japan):0L) mod(saloon:00, Sunny(country>japan,make>NISSAN):0S) yea(saloon:00, 92:16) yea(saloon:00, 95:1C) [/S]

Tableau 4 : exemples de résultat de la localisation interne

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'autres, une nouvelle illustration de l'analyse de R. Kittredge. Le Tableau 4 quelques résultats de SMS français.

Le Tableau 5 montre la répartition de l'effort pour le portage interne en terme de temps de travail et de pourcentage du code modifié ou ajouté.

Adaptation de EC-CATS	**Dictionnaire**	**Règles**
Temps de travail (H)	100	45
% du code modifié	90	5

Tableau 5: Répartition de l'effort pour le portage interne

2.3. Localisation externe

La localisation externe a été expérimentée sur une deuxième application de recherche de musique (IMRS) (Kumamoto 2007) qui traite des énoncés spontanés en japonais en adaptant le même extracteur de contenu du français construit initialement par H. Blanchon (Blanchon 2003) pour le domaine du tourisme, en restant dans la même langue, puis en changeant de langue (anglais).

Pour IMRS (Kumamoto 2007), nous avons obtenu une représentation interne (IF-Musique pour le français et IF-Music pour l'anglais) qui contient chacune un vecteur composé de dix composants. Chaque composant correspond à un axe parmi dix. La valeur d'un composant est un nombre réel entre 0 et 7 qui correspond à sept degrés de l'échelle associée à l'axe en question. Le symbole « nil » veut dire « *don't care* ». Par exemple, l'axe « *Happy – Sad* » est caractérisé par sept valeurs intermédiaires, « *very happy*, » « *happy*, » « *a little happy*, » « *medium*, » « *a little sad*, » « *sad*, » et « *very sad*, » qui correspondent respectivement aux valeurs *7.0, 6.0, 5.0, 4.0, 3.0, 2.0,* et *1.0.*

SMS en français	**(IF-CATS →CRL-CATS) obtenue**
recherche voiture OPEL VECTRA	S wan(saloon, wanted) mak(saloon, OPEL(country>germany,country>europe)) mod(saloon,

	Vectra(country>germany,country>europe,make>OPEL)) /S
recherche à l'achat NISSAN Sunny modèle 92 à 95	S wan(saloon, wanted) mak(saloon, NISSAN(country>japan)) mod(saloon, Sunny(country>japan,make>NISSAN)) yea(saloon, 95) /S
Énoncé en français	**IF-Musique obtenue**
je veux un morceau de musique calme et très solennel	{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))}
je veux un morceau de musique assez fort et clair	:{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))}
Énoncé en anglais	**IF-Music obtenue**
I want a calm and very solemn music	{c:give-information+disposition+service(service=music, music-spec=(nil 6,0 nil nil 7,0 nil nil nil nil nil))}
I want a little noisy and bright music	{c:give-information+disposition+service(service=music, music-spec=(3,0 nil nil 6,0 nil nil nil nil nil nil))}

Tableau 6 : exemples de résultats obtenus par portage externe

Adaptation de FR-IF	**Dictionnaire**	**Règles**
CATS		
Temps de travail (H)	90	140
% du code modifié/ajouté	20	15
IMRS		
Temps de travail (H) (Fr ; En)	(20 ; 30)	(10 ; 20)
% du code modifié/ajouté (Fr ; En)	(3 ; 6)	(2 ; 4)

Tableau 7 : Répartition de l'effort pour le portage externe

Nous présentons et évaluons dans la section suivante le portage de CATS de l'arabe (langue source pour le portage) vers le français par traduction statistique des énoncés du français vers l'arabe (langue cible pour la traduction) et utilisation de l'extracteur de contenu original.

3. Portage par TA statistique avec un petit corpus

Le but n'est pas de produire des traductions parfaites, mais de produire des traductions permettant à l'extracteur de contenu "natif" d'extraire l'information pertinente.

La question est de savoir si cela est possible étant donné la très petite taille du corpus d'apprentissage disponible, et si oui quelle est la taille minimale suffisante d'un tel corpus.

3.1. Corpus arabe-français

Il s'agit toujours ici du domaine Cars du système CATS. Nous avons utilisé dans cette expérience des données d'entraînement et des données de développement.

Les données d'entraînement sont constituées d'un corpus parallèle, composé d'un corpus français obtenu par traduction manuelle (transtac_train.fr) d'un corpus arabe original (transtac_train.ar), et de ce corpus, dont une partie servira de référence pour les tests. Nous avons commencé l'expérience avec une très petite taille, 100 SMS, en calculant les scores BLEU et NIST obtenus pour la traduction des données de développement non utilisées dans l'étape d'entraînement. Nous avons progressivement augmenté la taille jusqu'à l'obtention d'une légère stabilité de la mesure.

Le Tableau 8 résume les informations concernant les données d'entraînement.

La comparaison entre la taille en mots du corpus français et arabe ne peut être définitive. Ici, la longueur moyenne d'un SMS français est plus élevée (9,93) que celle de l'arabe (8,75). Dans d'autres blocs de données, on peut observer l'inverse. Cela peut varier en fonction de la nature des données sources utilisées et de l'utilisation des variantes lexicales.

Une explication possible est que les données sources sont des données réelles rédigées par de vrais utilisateurs en Jordanie, et qu'aucune règle ne les empêche d'écrire « *VOLKSWAGEN* » en un seul mot ou en deux mots « *VOLKS WAGEN* », ou de même « *LANDROVER* » et « *LAND ROVER* » etc.

Taille du corpus d'entraînement	Taille du corpus arabe en mots	Nombre de mots par SMS arabe	Taille du corpus français en mots	Nombre de mots par SMS français	Taille du corpus arabe en octets	Taille du corpus français en octets
100	860	8,60	1010	10,10	8353	6362
200	1788	8,94	2051	10,26	17191	12828
300	2696	8,99	3052	10,17	25794	19004
400	3575	8,94	4006	10,02	33890	25015
500	4424	8,85	4954	9,91	42280	31150
600	5299	8,83	5907	9,85	50537	37222
700	6090	8,70	6809	9,73	58095	43087
800	6867	8,58	7742	9,68	65792	49040
900	7729	8,59	8811	9,79	74378	55990
1000	8685	8,69	9961	9,96	83999	63360
1100	9446	8,59	10803	9,82	91152	68716
Moyenne		8,75		9,93		

Tableau 8 : taille des données d'entraînement

3.2. Traduction statistique

La figure suivante présente l'adaptation de l'architecture générale d'un système de traduction statistique à notre cas. Étant donnée un SMS français f, nous cherchons la traduction arabe **ê** qui maximise **p(e|f)**, la probabilité qu'un SMS **e** soit la traduction de **f**, et la soumettons à CATS.

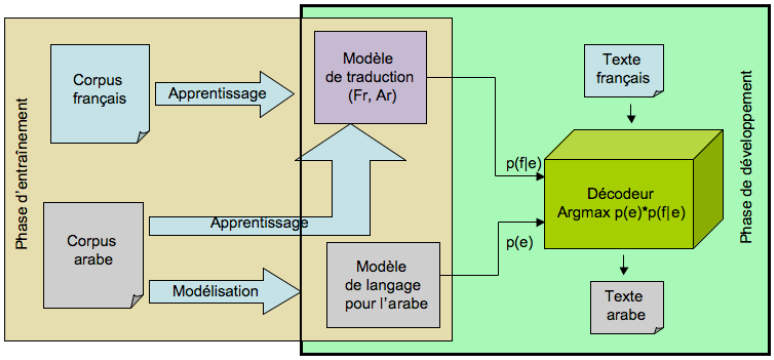

Figure 2 : Architecture générale d'un système de traduction statistique

Nous avons effectué un prétraitement des données avant de les transmettre au décodeur Pharaoh (Koehn, 2004). En particulier, nous avons aligné les données en utilisant l'outil GIZA++ (Och and Ney 2000).

Un modèle de langage pour la langue cible est nécessaire. Dans notre cas, l'arabe est la langue "native" de l'application et donc la langue cible pour la traduction. Nous devons donc construire un modèle de langage pour l'arabe. C'est une langue pour laquelle très peu de ressources sont disponibles et gratuites.

Nous avons trouvé un modèle de langage pour l'arabe, mais construit à partir du Coran, ce qui n'est pas du tout adapté au cas du sous-langage traité. Nous avons donc dû construire un modèle de langage pour le sous-langage *Cars* en nous basant sur ce que nous avions comme données. Nous avons utilisé pour cela le générateur de modèles de langage de Stolcke, disponible gratuitement sur le Web (Stolcke 2002) (*http://www.speech.sri.com/projects/srilm/*).

Nous avons construit notre modèle de langage en utilisant le même corpus d'entraînement que celui utilisé pour l'entraînement du décodeur de traduction. De la même façon, nous sommes partis d'une taille de corpus minimale et nous avons augmenté la taille au fur et à mesure jusqu'à obtenir des résultats satisfaisants.

3.3. Évaluation des résultats

3.3.1. Exemples

La Figure 3 montre quelques résultats obtenus pour une taille de corpus d'entraînement limitée à 400 SMS.

Portage de sous-langage restreint par TA statistique avec petit corpus

Langue référence (arabe originale)	Langue source (français)	Langue cible (TA statistique Fr → Ar)
مطلوب نيسان صني غير اوتوماتيك موديل ٩٣ الى ٩٧.	recherche NISSAN Sunny manuelle modèle 93 à 97	مطلوب نيسان صني عادي موديل ٩٣ ل٩٠٠٥٠ ٩٧
مطلوب رينو كليو غير اوتوماتيك موديل ٩٥ الى ٢٠٠٠.	recherche RENAULT Clio manuelle modèle 95 à 2000	مطلوب رينو Clio عادي موديل ٩٥ ل٩٠٠٥٠ ٢٠٠٠
مطلوب سيارة هونداي.	recherche voiture HYUNDAI	مطلوب سياره HYUNDAI
مطلوب سيارة ميتسوبيشي.	recherche voiture MITSUBISHI	مطلوب سياره ميتسوبيشي
أبحث عن سياره بي إم دبليو موديل ٩٢.	je cherche une voiture BMW modèle 92	اريد cherche a سياره بي ام دبليو موديل ٩٢
مطلوب دابو لمنز.	recherche DAEWOO	مطلوب دابو
مطلوب نيسان صني ٩٣ الى ٩٥ غير اوتوماتيك.	recherche NISSAN Sunny modèle 93 à 95 manuelle	مطلوب نيسان صني موديل ٩٣ ل٩٠٠٥٠ ٩٥ عادي
مطلوب هوندا سيفك موديل ٩٤ فل اويشن غير اوتوماتيك.	recherche HONDA Civic modèle 94 toutes options ma	مطلوب هوندا سيفك موديل ٩٤ فل اويشن عادي
مطلوب كيا سيفيا .	recherche KIA Sephia	مطلوب كيا سيفيا
مطلوب دابو لانوس ٩٥ الى ٩٧ غير اوتوماتيك .	recherche DAEWOO Lanos 95 à 97 manuelle	مطلوب دابو لانوس ٩٥ ل٩٠٠٥٠ ٩٧ عادي

Figure 3 : exemple de résultats obtenus par traduction statistique (Pharaoh)

En tenant compte de la complexité et de la richesse de la langue arabe, le résultat obtenu pour cette taille est encourageant. En effet, très peu de mots sont inconnus, comme HYUNDAI.

Nous avons utilisé le même corpus d'évaluation que celui utilisé dans l'évaluation de la version originale de CATS (200 SMS, 100 d'achat et 100 de vente).

Dans ce qui suit, nous évaluons d'abord la traduction par les deux méthodes NIST et BLEU, classiquement utilisées en TA statistique. Pour mesurer l'adéquation à la tâche, nous évaluons ensuite l'extraction d'information par une mesure de rappel et de précision.

3.4. Évaluation de la traduction par NIST et BLEU

L'évaluation par NIST et BLEU suppose d'avoir au moins une traduction de référence et une traduction candidate pour chaque énoncé source (Papineni, Roukos et al. 2002).

Dans notre cas, la traduction de référence est le SMS arabe original, la traduction candidate est le résultat produit par le système Pharaoh, et l'énoncé source est le SMS du corpus d'évaluation en français.

Le Tableau 9 présente les différents scores NIST et BLEU obtenus pour le corpus d'évaluation en fonction de la taille du corpus d'entraînement.

On observe que ces scores n'augmentent presque plus à partir de 500 SMS, et qu'ils sont très faibles par rapport aux scores obtenables

avec de très gros corpus. Mais il est possible que les résultats soient malgré cela utilisables pour en extraire une information correcte.

Taille du corpus d'entraînement	NIST	BLEU
100	3,52	0,14
200	4,23	0,20
300	4,42	0,21
400	4,64	0,23
500	4.95	0.25
600	5,00	0.25
700	5.04	0.25
800	5.05	0.26
900	5.01	0.25
1000	5.07	0.26
1100	5.05	0.26

Tableau 9 : scores NIST et BLEU en fonction de la taille du corpus d'entraînement

La courbe de la Figure 4 montre une croissance très faible du score BLEU à partir de la valeur 0,26 qui correspond à une taille du corpus d'entraînement égale à 800 SMS. À partir de cette même taille de corpus, la courbe de la Figure 5, représentant la mesure NIST, croit aussi très faiblement à partir de la valeur 5,05. Cela veut dire qu'une augmentation de la taille du corpus d'entraînement ne modifie presque pas la valeur du score BLEU.

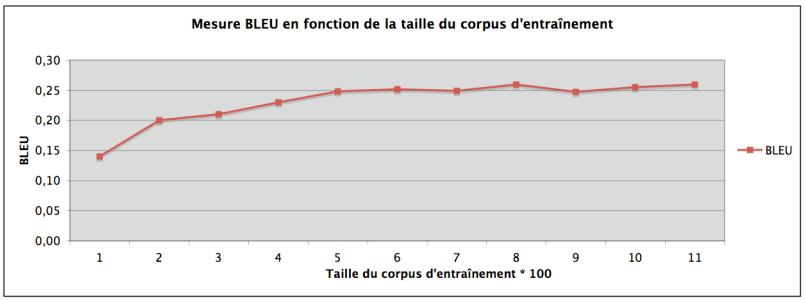

Figure 4 : BLEU en fonction de la taille du corpus d'entraînement

Nous ne garantissons pas que les mesures ne peuvent s'améliorer après une certaine augmentation de la taille du corpus d'entraînement ou l'ajout d'autres ressources et outils comme un analyseur morphologique pour le français. Mais, rappelons-le, notre objectif était de proposer des solutions de multilinguïsation simples, et applicables sur le terrain avec le coût le plus faible possible. Or, un examen rapide des résultats nous indique qu'il semble n'y avoir que peu ou pas de perte d'information. Nous vérifierons ce point plus loin, en appliquant l'extracteur de contenu à ces résultats, et en comparant les CRL-CATS obtenus avec ceux obtenus à partir des SMS originaux.

Notons qu'on arrive approximativement au même score BLEU que d'autres expériences sur le couple de langues anglais—arabe. Ainsi, d'autres chercheurs de notre équipe sont arrivés à la valeur 0,25 pour BLEU, mais cela leur a demandé une taille de corpus d'environ 42000 phrases car il s'agissait d'une variante de l'arabe plus complexe et plus générale que la nôtre (dialecte irakien, dialogues informationnels) (Besacier, 2007).

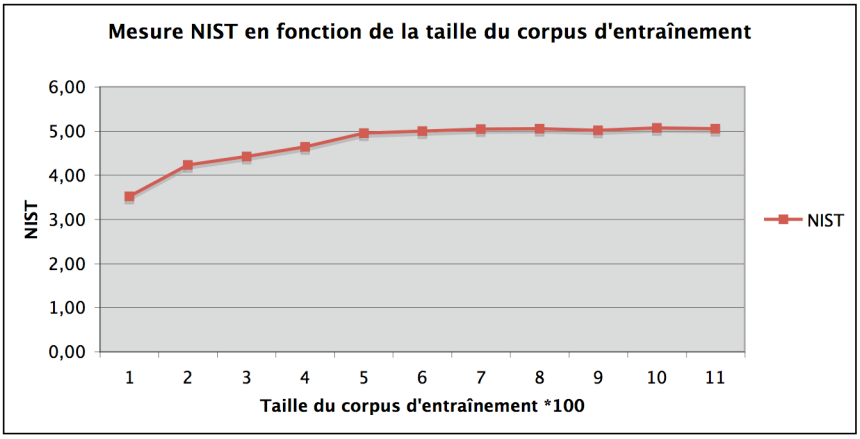

Figure 5 : score NIST en fonction de la taille du corpus d'entraînement

3.5. Évaluation de l'extraction d'information

Les résultats d'extraction de contenu de la version arabe obtenue par TA statistique des SMS français et ceux obtenus à partir de la version originale (arabe) sont regroupés dans le tableau suivant, pour les propriétés les plus importantes. Les pourcentages de portage par rapport

à la version originale varient entre 85% et 98%, avec une moyenne de 93 %.

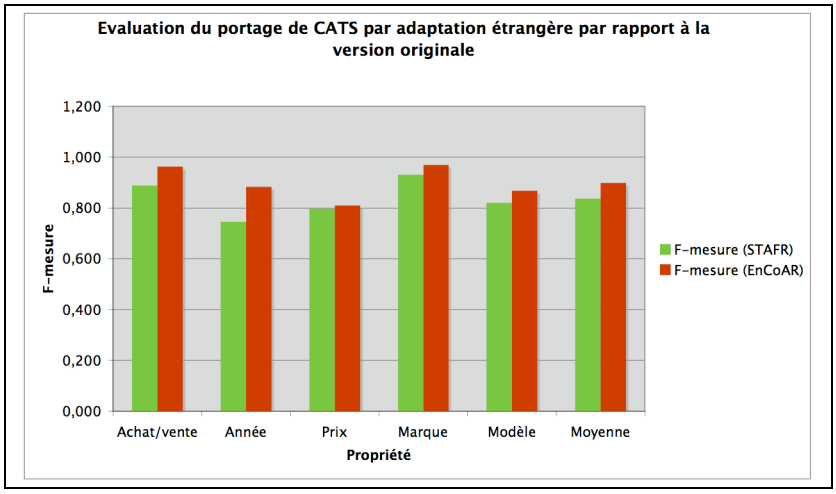

Figure 6 : comparaison entre F-mesure (par rapport à la version originale)

L'avantage de cette méthode est qu'elle ne nécessite aucun accès aux ressources de l'application-mère. La Figure 6 permet de mieux visualiser la comparaison entre les valeurs de F-mesure trouvées pour chacune des versions du système.

	SMT-FR			EnCo-AR			
Propriété	Précision	Rappel	F-mesure (SMT-FR)	Précision	Rappel	F-mesure (EnCo-AR)	% portage
Achat/vente	1,000	0,800	0,889	0,956	0,970	0,963	92
Année	0,753	0,740	0,746	0,817	0,960	0,883	85
Prix	0,883	0,726	0,797	0,800	0,822	0,811	98
Marque	0,964	0,901	0,931	0,978	0,963	0,970	96
Modèle	0,957	0,718	0,820	0,901	0,837	0,868	95
Moyenne	0,912	0,777	0,837	0,890	0,910	0,899	93

Tableau 10 : Comparaison entre les résultats d'extraction de contenu pour 200 SMS en arabe obtenus par traduction statistique, par rapport aux SMS de référence

Conclusion

Nous avons présenté une application de la traduction automatique statistique (SMT) au "portage linguistique" de l'arabe au français de CATS, un système traitant le contenu de brefs messages spontanés en langue naturelle (SMS). Il s'agit d'énoncés réels, car CATS est une application déployée sur le réseau FastLink en Jordanie. Nous avons travaillé sur la partie "occasion automobile" (Cars), où il s'agit d'un "sous-langage" très restreint.

Nous avions préalablement expérimenté deux autres méthodes, l'une demandant un accès au code de l'extracteur de contenu "natif", et l'autre consistant à adapter un extracteur de contenu du français existant. Il nous avait suffi pour cela de construire un très petit corpus parallèle, augmenté d'un dictionnaire bilingue assez complet lié à l'application choisie (petites annonces en occasion automobile), et nous nous sommes limités à ces ressources pour construire avec Pharaoh un système de TA statistique français-arabe pour des SMS en français et évaluer la faisabilité d'un portage de CATS en français par cette méthode.

Bien que la TA statistique soit réputée ne fonctionner assez bien que si l'on dispose de très grands corpus parallèles, ce système a produit des résultats satisfaisants, au sens où les descripteurs de contenu produits par l'extracteur de contenu de CATS sont très proches de ceux produits à partir des SMS de référence correspondants en arabe, en termes de rappel et de précision, alors même que les scores BLEU et NIST sont assez mauvais.

Il semble donc qu'on puisse se limiter à de très petits corpus pour utiliser efficacement la TA statistique sur des "sous-langages" très restreints, du moment qu'on a un dictionnaire bilingue assez complet : même si les traductions ne sont pas très "fluides", elles peuvent être "adéquates", même si les deux "langues-mères" des deux sous-langages considérés sont assez distantes.

On a ici une illustration de la validité de l'affirmation de Kittredge selon laquelle deux sous-langages qui se correspondent dans deux langues différentes sont très proches entre eux, et souvent plus proches entre eux qu'ils ne le sont chacun de leur langue-mère respective, ce qui

permet de les considérer et de les traiter comme des variantes l'un de l'autre.

Bibliographie

Besacier, L. (2007). *Transcription enrichie de documents dans un monde multilingue et multimodal*. Grenoble, Université Joseph Fourier. HDR, 300 p.

Biber, D. (1993). *Using register-diversified corpora for general language studies* (Special issue on using large corpora): 219-241. MIT Press Cambridge, MA, USA.

Blanchon, h. (2004). *Comment définir, mesurer et améliorer la qualitél'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*. Grenoble, Université Joseph Fourier. HDR, 380 p.

Bross, I. D. J., P. A. Shapiro, et al. (1972). *How information is carried in scientific sub-languages*. Science, pp. 1303-1307.

Chandioux, J. (1988). *10 ans de METEO*. Traduction Assistée par ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires, OFIL, A. Abbou, ed. Paris.

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*. GETA - CLIPS. Grenoble, Université Joseph Fourier, Thèse, 296 p.

Daoud, D. M. (2005). *Building SMS-based System using Information Extraction Technology*. ACIDCA-ICMI, Tozeur, Tunisia. 7 p.

Deville, G. (1989). *Modelization of Task-oriented Utterances in a Man-Machine Dialogue System*, University of Antwerpen, Belgique. PhD, 200 p.

Grishman, R. and R. Kittredge (1986). *Analyzing language in restricted domains*, Hillsdale NJ.

Hajlaoui, N. and C. Boitet (2007). *Portage linguistique d'applications de gestion de contenu*. TOTh Conférence sur la Terminologie & Ontologie: Théories et Applications, Annecy France, 13 p.

Harris, Z. (1968). *Mathematical structures of language*. New York, Wiley-Interscience.

Kittredge, R. (1982b). *Variation and Homogeneity of Sublanguages. in Sublanguage* - Studies of Language in Restricted Semantic Domains. Walter de Gruyter. Berlin / New York.

Koehn, P. (2004). Pharaoh: a Beam Search Decoder for Phrase-Based SMT. 6th AMTA, Washington.

Kumamoto, T. (2007). A Natural Language Dialogue System for Impression-based Music-Retrieval. CICLING 07 (Computational Linguistics and Intelligent Text Processing), Mexique.

Och, F. J. and H. Ney (2000). Improved statistical alignment models. The 38th Annual Meeting of the Association for Computational Linguistics. pp. 440-447.

Papineni, K., S. Roukos, et al. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia.

Slocum, J. (1986). *How one might automatically identify and adapt to a sublanguage*. Book section « Analyzing language in restricted domains », pp. 195-210.

Stolcke, A. (2002). *SRILM - an Extensible Language Modeling Toolkit*. ICSLP, Denver, USA.

Sekine, S. (1994). *A new direction for sublanguage NLP*. International Conference on New Methods in Language Processing, 8 p.