Recherche et production de corpus de messages pour la multilinguisation de sites de e-commerce en SMS, initialement en arabe

Najeh Hajlaoui

Recherche et production de corpus de messages pour la multilinguïsation de sites de e-commerce en SMS, initialement en arabe

Najeh HAJLAOUI, GETA, CLIPS, IMAG
Université Joseph Fourier, BP 53
38041 Grenoble, France
Najeh.Hajlaoui@imag.fr

Abstract

In this paper, we present our research in the framework of the CATS project (Classified Ads through SMS) [3]. CATS is a system of management of small Arabic advertisements posted in SMS of buying and selling (cars, real estate...), currently deployed in Jordan by the FastLink operator. In order to adapt this system to other languages (French, English) and in length by applying it to other sectors (employment, marriage, domestic machines, trade of mobile phone, pages yellow...), we are in the difficulty in finding or building SMS corpora functionally equivalent to a real, Arab and natural corpus. A simple translation of this starting corpus gives it the same type of corpus (real, natural)? We present, in this paper, an answer to this question and a solution for a case of multilinguisation of sites of e-commerce in SMS, initially in Arabic.

Résumé

*Dans ce papier, nous présentons notre recherche qui se déroule dans le cadre du projet CATS (*Classified Ads through SMS*) [3]. CATS est un système de gestion de petites annonces en arabe postées en SMS pour l'achat et la vente (occasion automobile, immobilier…), actuellement déployé en Jordanie par l'opérateur FastLink. Pour faire évoluer ce système en largeur en l'adaptant à d'autres langues (français, anglais) et en longueur en l'appliquant à d'autres secteurs (emploi, mariage, fourniture, appareils ménagers, commerce de téléphone portable, pages jaunes…), nous sommes confrontés à la difficulté de trouver ou de construire des corpus de SMS fonctionnellement équivalents à un corpus arabe, réel et naturel. Une simple traduction de ce corpus de départ donne-elle un corpus de même critère (réel, naturel) ? Nous présentons, dans ce papier, une réponse à cette question ainsi qu'une solution pour un cas de multilinguïsation de sites de e-commerce en SMS, initialement en arabe.*

Mots-clés *: SMS (Short Messages Services) , corpus naturel et réel, multilinguïsation, traduction brute, traduction fonctionnelle, CRLcats(Content Representation Language- Classified Ads through SMS)*

Keywords*: SMS (Short Messages Services), natural and real corpus, multilinguisation, brut translation, functional translation, CRLcats (Content Representation Language- Classified Ads through SMS)*

1 Problématique

1.1 Contexte : description du système CATS

Notre recherche se déroule autour du projet CATS (*Classified Ads through SMS*) [3]. C'est un système d'achat et de vente basé sur l'utilisation des SMS. Bien qu'il n'y ait pas de transaction directe, il aide les utilisateurs à vendre et acheter sans avoir à se déplacer. N'importe quelle personne se trouvant en Jordanie qui veut vendre une voiture peut envoyer un SMS en arabe au système indiquant sa proposition de vente avec la précision de quelques attributs comme le modèle, l'année, etc. Une autre personne intéressée par l'achat d'une voiture, peut simplement envoyer un message au système, expliquant sa requête. Dans ce cas de correspondance et grâce au système CATS, les deux personnes sont mets en contact. Par exemple un utilisateur a envoyé à partir de son téléphone mobile le SMS suivant décrivant sa requête d'achat.

6/1/2005 11:32:20 AM Message In : 962795028727,3750 الى 3500 والسعر مابين 97 مطلوب سيارة هوندای موديل

Une traduction brute de ce message donne :

6/1/2005 11:32:20 AM Message reçu : 962795028727 : recherche d'une voiture Honda, modèle 97, et le prix entre 3500 et 3750

Si c'était un Français, il écrirait plutôt (traduction fonctionnelle) :

6/1/2005 11:32:20 AM Message reçu : 962795028727 : recherche voiture Honda, modèle 97, prix entre 3500 et 3750 Dinars

Un autre utilisateur a envoyé la proposition de vente suivante.

9/4/2005 12:48:12 PM Message In : +962795463274,فساتين سهرة وأطقم بحالة ممتازة للبيع بداعي السفر بأسعار مغرية

Une traduction brute de cette proposition en français donne :

9/4/2005 12:48:12 PM Message In : +962795463274, En raison de départ à vendre vetement de soirée en très bon état avec un prix excellent

Si c'était un Français, il écrirait plutôt (traduction fonctionnelle) :

9/4/2005 12:48:12 PM Message In : +962795463274, Cse départ vend vêtement de soirée en très bon état, TB prix

Ces SMS sont envoyés à un numéro spécial et sont enregistrés automatiquement dans un corpus de données précédés de la date, de l'heure et du numéro de téléphone. Leurs textes sont analysés et une réponse est envoyée automatiquement à l'expéditeur du SMS en cas de correspondance de la demande avec l'une des propositions.
Ce corpus est enrichi automatiquement par les annonces envoyées. Il est formé de phrases en arabe mélangées de dialecte jordanien, d'arabe littéraire et d'autres mots anglais importés. C'est un corpus naturel, évolutif et assez intéressant. Son point fort est la réalité du contenu, mais le rendre multilingue est une tâche assez délicate et utile en même temps pour démarrer l'adaptation du système CATS à d'autres langues.

1.2 Architecture

La figure suivante montre l'architecture du système CATS : un vendeur ou un acheteur envoie un SMS, le texte de ce SMS est analysé avec un extracteur d'information et est représenté dans un format intermédiaire CRLcats (Content Representation Language)**.** Cette représentation est traduite en requête SQL et envoyée vers une base de données. En cas de correspondance de requêtes, une réponse sera transmise à la personne intéressée.

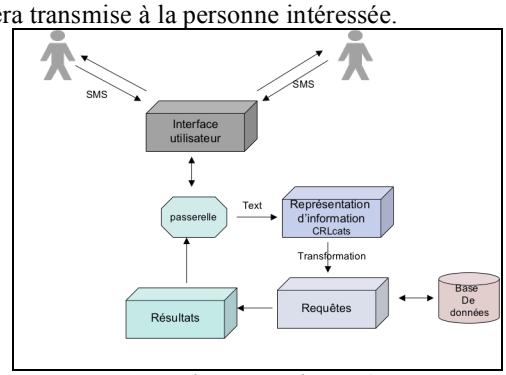

Figure 1 : architecture du système CATS

1.3 Motivation

Nous voulons porter CATS vers d'autres langues, en commençant par le français pour les deux simples raisons suivantes : la première c'est que Daoud veut porter son système au français mais il ne le parle pas et la deuxième c'est qu'on a un extracteur d'information pour le français.

1.4 Problèmes et difficultés

Nous disposons du corpus original SMS_ar_orig.txt, constitué de 12372 d'annonces en arabe, et de taille 448 Ko. Nous sommes confrontés à la difficulté de trouver des corpus fonctionnellement équivalents en français et en anglais : les corpus d'annonces et de tchat sur Internet sont rares et leurs contenus sont différents de ce que nous cherchons. Les principales différences sont culturelles, ou proviennent de l'utilisation d'un langage codé, ou du mode de saisie.

1.4.1 Différences culturelles

Les différences culturelles se concrétisent dans l'utilisation d'un dialecte pour chaque langue. Dans le cas de notre corpus arabe, quelques mots sont écrits en arabe comme ils sont prononcés en anglais, c'est une transcription en écriture arabe de la prononciation anglaise du mot. D'autre part, l'anglais est la deuxième langue en Jordanie, contrairement à d'autres pays arabes et francophones. Mais un phénomène similaire est observé dans d'autres pays, par exemple, en Tunisie, pays francophone, où on arabise des mots français et où on les emploie dans les discussions quotidiennes au lieu d'utiliser les équivalents arabes.
Exemple : dans les SMS présentés dans le paragraphe 2.1, le mot « موديل » qui veut dire « modèle » est une prononciation du mot anglais « model » mais en signifie pas "modèle" en arabe.

1.4.2 Utilisation d'un langage codé

Il existe des logiciels de rédaction des SMS "complets" pour le français et d'autres langues. Un communiqué de presse montre que les Français veulent pourtant défendre le bon langage, mais un vrai phénomène de société s'empare de la France : les SMS forment un langage codé riche en expressions familières ou abrégées du genre MDR (mort de rire), BI1 (bien) ou A2M1 (à demain). Parmi les personnes interrogées, 41 % disent utiliser toujours ou très souvent de tels termes codés lorsqu'elles envoient des messages SMS à leurs amis ou aux membres de leur famille. Cette proportion fait un bond spectaculaire chez les jeunes de 18 à 24 ans : elle passe à plus de 73 % (http://www.languefrancaise.net/news/index.php).

1.4.3 Absence de T9 pour certaines langues ou principe de saisie très différent

T9 Text Input, célèbre logiciel de saisie de texte utilisant l'alphabet ou les caractères est fréquemment utilisé pour certaines langues. Théoriquement l'utilisation du T9 pour certaines langues, réduit l'utilisation des abréviations dans la rédaction des SMS, mais tout dépend de la culture, des habitudes et du savoir des utilisateurs :certains ne savent utiliser le T9, d'autres n'aiment pas l'utilisation d'un dictionnaire…

1.4.4 Conclusion et hypothèse

Dans ce travail, notre problématique est de trouver des méthodes qui permettent de produire des corpus réels, spontanés, naturels et fonctionnels en

tenant compte des différences cités ci-dessus
Pour cela, nous posons l'hypothèse H suivante :

H : les traductions directes d'un corpus réel constitué de phrases naturelles ne donnent pas de résultats naturels en français, anglais etc. Nous montrons dans ce qui suit la validité de l'hypothèse H et nous proposons une solution au problème suivant : **comment construire des corpus parallèles de SMS fonctionnellement équivalents dans des situations de e-commerce ?**

2 Etude : traitement et mise en forme du corpus existant

2.1 Etude précise du corpus source

Le corpus CATS contient environ 23 % d'arabe jordanien, 5% de mots anglais importés, 70% d'arabe littéraire et 2% d'autres mots inconnus. Cela rend le traitement de ce type de corpus difficile, ajoutant d'autres difficultés liées à la structure du corpus.

2.2 Structure du corpus original

Comme le montre la figure 2, le corpus original (SMS_ar_orig.txt) contient accidentellement des lignes vides, des morceaux d'annonces incomplètes (présence de la date et/ou de l'heure de l'annonce, mais absence du cœur de l'annonce, ou annonces faites en anglais). Certes, ce type de corpus est moins propre qu'un corpus normal. Les raisons sont nombreuses : erreurs humaines naturelles (envoi d'un message incomplet en appuyant sur la touche « OK » involontairement, SMS rédigés par des personnes résidant en Jordanie mais parlant l'anglais et pas l'arabe, et n'hésitant pas à envoyer des annonces en anglais pour leurs besoins commerciaux). D'autres erreurs techniques peuvent arriver, comme la faiblesse du réseau pour tel ou tel opérateur téléphonique. Malheureusement, la dernière version du système CATS accepte encore de tels SMS "sales". Les messages en anglais doivent être traduits manuellement en arabe avant d'être passés au système.

Il est clair que ce type de corpus nécessite une étape de prétraitement (nettoyage et filtrage) pour pouvoir en extraire le contenu.

```
9/3/2005 12:52:40 PM   Message In : +962796770392,+962796770392,شقتان بخلفا للبيع ارضيات فوشان واحد حدائق كراج رمونت 190 م الواحده جديده ش القلل

9/3/2005 12:56:11 PM   90050

9/3/2005 12:56:11 PM   Message In : +962796322201,Kia  rio 2003  silver 1500cc

9/3/2005 12:57:28 PM   90050

9/3/2005 12:57:28 PM   Message In : +962795202973,شبه فيلا للبيع مع بركه سباحه 230م منطقه الروضه ناعور

9/3/2005 1:04:36 PM   90050

9/3/2005 1:04:36 PM   Message In : 962796322201,2003 للبيع كيا ريو

9/3/2005 1:05:36 PM   90050

9/3/2005 1:05:36 PM   Message In : +962795769711,
```

Figure 2: exemple du corpus SMS_ar_orig

2.3 Étape de prétraitement et extraction du contenu

Les annonces du corpus SMS_ar_orig sont constituées des champs suivants :

Champ	date	heure	Message In :	n° télé,	cœur du message
Exemples	9/3/2005	12:47:47 PM	Message In :	+962795731129,	تعم تعم شقة اريد التفاصيل
	9/9/2005	7:57:17 PM	Message In :	+962795314763,	

Tableau 1: champs du corpus original SMS_ar_orig

Les cœurs de ces annonces sont précédés par une virgule, ce qui facilite le nettoyage. Par des programmes Perl, nous anons filtré ce corpus en supprimant tous les messages vides, les lignes vides et d'autres saletés parfois incompréhensibles. Cela réduit le nombre d'annonces de 12372 à 3017. Nous obtenons ainsi un nouveau corpus propre que nous appelons SMS_ar_extr_content.

2.4 Étape de normalisation

Notre corpus de départ est un fichier texte en codage arabe (Windows). Afin de faciliter sa manipulation et son traitement sur d'autres plates-formes (UNIX, LINUX, MAC OS X), et en utilisant de simples outils tels que TexEdit ou Excel, et des programmes de conversion, nous avons transféré ce corpus dans une première étape vers un format texte avec un codage utf-8, puis dans des formats XML (codage utf-8). Le premier format, CXM (Common eXample Markup), sert à représenter des corpus monolingues et est très simple, et le deuxième format, CPXM (Common Parallel eXample Markup), est destiné à coder des fichiers multilingues.

2.5 Mise dans une base de données

Afin de faciliter la manipulation des données et pour une meilleure gestion de l'historique des versions ainsi que les différences entre les différentes versions de ce même corpus et avec d'autres types de corpus (tchat, annonces sur le Web), nous avons mis ces données dans une base de données conçue de façon à gérer la notion de versionnage.

3 Réflexions sur les méthodes possibles de multilinguisation du système CATS

3.1 Méthodes possibles

Il s'avère que CATS utilise des représentations du contenu en "CRLcats", syntaxiquement semblables à UNL, mais qui ne correspondent pas du tout à la représentation UNL standard, liée à une expression linguistique en anglais (même si elle en est une représentation profonde). CRLcats est une représentation de type *Propriété{objet, valeur}*.

Exemple : *"Je cherche une voiture française"*

Ads_type{saloon, wanted}
Country{saloon, French}

Figure 3 : exemple de représentation CRLcats

Plusieurs méthodes sont possibles pour la multilinguisation. Nous donnons dans ce qui suit une brève description de chacune des méthodes. Dans d'autres articles en cours de rédaction, nous donnons une description complètes ainsi qu'une comparaison de ces méthodes.

Localisation interne : localisation d'un CE-CATS (Content Extractor CATS) existant pour une autre langue. Concrètement, il s'agit d'adapter l'implémentation existante du CE pour l'arabe (arabe-contentcats, notée ar-cc) au français (français-contentcats, notée fr-cc).

Adaptation externe : adaptation et greffe du CE externe (CE-ext) d'un autre système pour la même langue. Concrètement, il s'agit d'adapter à CATS un CE du français existant, celui du du projet Nespole fr-IF [5].

Transformation d'un analyseur en un CE : utilisation de nouvelles voies, il s'agit d'écrire un CE en Ariane G5 (par spécialisation/adaptation d'un analyseur classique du français).

Construction d'un système de TA : construire un système de TA des SMS en français vers les SMS en arabe. Pour cela, il faudrait construire un grand corpus parallèle, par traduction manuelle d'arabe en français, ou par génération à partir des représentations"pivot" CRLcats, puis par construction automatique d'un système de TA français-arabe spécialisé aux SMS, soit par une méthode statistique (de nombreux "kits" existent), soit par une méthode analogique [7]. Le point intéressant ici est de déterminer le domaine de validité de la méthode (caractéristiques du sous-langage, taille minimale du corpus nécessaire pour obtenir le niveau de qualité souhaité).

Nous présentons ici un besoin commun pour toutes ces méthodes : un corpus réel français.

3.2 Discussion

L' idée de traduire automatiquement la langue en question L vers l'arabe et puis de l'arabe vers CRLcats est à rejeter pour l'instant car il existe seulement un seul système de TA Anglais-arabe et qui très mauvais et propriétaire, même un bon système de TA ne traduira pas les SMS. Nous testerons quand même l'hypothèse H avec des mesures de rappel.
Dans tous les cas, il faut disposer d'un corpus de SMS en L « fonctionnellement équivalent » à celui en arabe pour démarrer et tester le système. Pour cela, différentes approches sont possibles, l'idée est de réutiliser les ressources existantes et non pas de développer de nouveaux systèmes de traduction. Dans ce qui suit, nous essayons de définir un corpus réel et de présenter quelques approches qui permettent la production de tels corpus:

3.2.1 Qu'est ce qu'un corpus naturel ?

Nous définissons un corpus naturel comme suit : *Un corpus naturel est une collection de textes supposés êtres représentatifs d'une langue donnée, d'un dialecte, ou de tout autre sous-ensemble d'une langue construit d'une façon spontanée et naturelle pour être utilisé dans différents objectifs (commercial, linguistique, statistique…).*

3.2.2 Approche 1 : utilisation des déconvertisseurs

Il s'agit d'utiliser des déconvertisseurs disponibles UNL-L après avoir effectué une transformation de CRLcats en UNLstd. Si les résultats sont faibles, nous devrons étudier pourquoi. Cela est-il dû à l'utilisation d'une représentation non linguistique (CRLcats) ou est-ce le vocabulaire qui est très spécifique ?
L'opération de passage de CRLcats à UNL devrait être assez facile mais les déconvertisseurs ne produisent pas des énoncés de type « SMS ».

3.2.3 Approche 2 : approche classique

Il s'agit d'utiliser des systèmes de traduction pour obtenir des « prétraductions » en L (en anglais par exemple AR→EN) et passer à « des traductions de référence » en les révisant et les traduisant dans L pour obtenir une bonne qualité. Une fois que les traductions de référence sont obtenues, nous les utiliserons comme « accroche » pour trouver les correspondances naturelles produites directement par des utilisateurs spontanés parlant la langue L comme langue maternelle.

3.2.4 Conclusion

Dans les deux cas, les deux approches sont utiles et donnent comme résultat un corpus de SMS naturels en L fonctionnellement équivalent au corpus arabe de départ. L'approche 1 suppose qu'on dispose des représentations CRLcats ainsi qu'une transformation de CRLcats vers UNL. Dans notre cas et pour des raisons de manque de ressources logicielles, nous utiliserons dans ce que suit l'approche 2.

4 Recherche et production de corpus équivalents

Dans cette partie, nous cherchons à trouver et construire des corpus en L fonctionnellement équivalents à des corpus en arabe sous une certaine contrainte de production (dans notre cas c'est un SMS).

4.1 Traduction du corpus arabe

Nous avons traduit la totalité du corpus SMS_ar_extr_content en anglais avec Systran professionnel version 5. Puisqu'il n'existe pas de système de TA français arabe, nous avons décidé, dans première étape, de traduire manuellement une partie de ce corpus afin d'évaluer l'hypothèse H. Comme le montre la figure suivante, une traduction brute produite par un non Français est généralement très différente d'une traduction naturelle produite par un Français, c'est ce que peut dire un Français d'une façon spontanée. Nous évaluons cette différence de traduction entre brute et naturelle ou encore littérale et fonctionnelle en calculant la distance d'édition entre les deux traductions. Une mesure de distance moyenne trouvée est de 21,88 (voir Figure 4:exemple de traduction brute, littérale et fonctionnelle du corpus SMS_ar_extr_content).
La distance moyenne trouvée est le nombre minimal de suppressions, insertions ou de remplacements nécessaires pour transformer une traduction brute/littérale à une traduction naturelle/fonctionnelle.

Arabe (original)	Français (traduction brute de NH)	h_debut	h_fin	duree_incr	Français littéral (révision par CB)	h_debut_f	Français naturel (traduction fonctionnelle, CB)	Distance (brute, naturelle)
							Distance moyenne	21,88
نعم نعم شقة اريد التفاصيل	Oui je veux les détails de l'appartement	49	49	0	Oui, je veux les détails de l'appartement	46	Oui, je veux les détails de l'appartement	1
مطلوب شرا سياره تويوتا بكب 4 باي 4 كبينه واحده	Recherche d'achat de voiture Tayota 4*4 une seule cabine	49	49	0	Recherche à l'achat une voiture Toyota 4*4 à une seule cabine	46	Recherche à l'achat une Toyota 4*4	29
مطلوب كيا سبفا95فحه وباقي اقساط	Recherche Kia 95 première achat et le reste sous forme de tranches	49	49	0	Recherche Kia 95 première main	46	Recherche Kia 95 première main	38
ارض نصف دونم في ضاحيه الياسمين	Terrain dans la cité Yassamine	49	49	0	Terrain dans la ville de Yassamine	46	Terrain à Yassamine	12
للبيع رينو ميجان م 2000	A vendre Renault Megane m 2000	49	49	0	À vendre Renault Megane modèle 2000	46	À vendre Renault Mégane modèle 2000	7

Figure 4:exemple de traduction brute, littérale et fonctionnelle du corpus SMS_ar_extr_content

4.2 Evaluation de l'hypothèse H

Afin d'évaluer l'hypothèse H**,** nous utilisons les notions classiques **de rappel** et de **précision** pour un nombre limité de phrases.
Rappel = Nombre de traductions correctes / Nombre de traductions de référence
Précision = Nombre de traductions correctes / Nombre de traductions proposées
Afin de trouver les valeurs des mesures de rappel et précision, nous utilisons une méthode connue basée sur le calcul de distance d'édition entre deux chaînes de caractères (algorithme de Wagner & Fischer, 1974) [6]. La distance d'édition entre deux chaînes (caractères ou mots) est le nombre minimal de suppressions, insertions ou de remplacements nécessaires pour transformer une chaîne à un autre. La distance d'édition entre deux chaînes $\mathbf{x} = a_1 \ldots a_m$ et $\mathbf{y} = b_1 \ldots b_n$ est D(m, n), définie par :

$$D(i,j) = \min \begin{cases} D(i-1,j)+1 & \text{(suppresson du caractère i de } \mathbf{x}) \\ D(i,j-1)+1 & \text{(insertion d'un caractère dans } \mathbf{x}) \\ D(i-1,j-1)+t(i,j) & \text{(si t(i, j) } \neq 0 \text{ substituton sinon rien)} \end{cases}$$

$$t(i,j) = 0 \text{ si } \mathbf{x}(i) = \mathbf{y}(i) \text{ sinon } t(i,j) = 1$$

Exemple : comme le montre la figure 6, un calcul de distance donne la valeur 20.
Traduction brute : *Recherche de Mercedes 81 avec contrôle technique et toutes les options.*
Traduction fonctionnelle : *Recherche Mercédes 81 avec ctl tech et ttes options.*

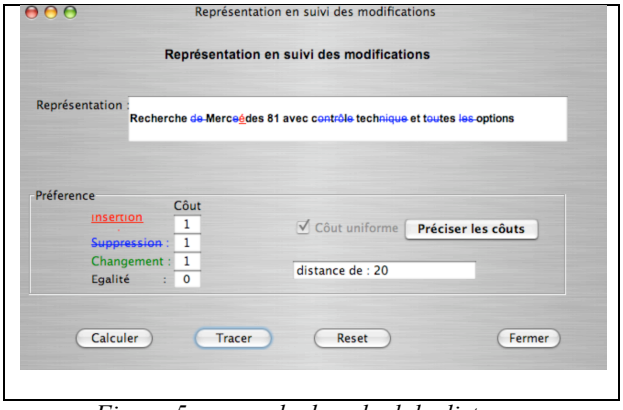

Figure 5 : exemple de calcul de distance

4.2.1 Résultats :

Avec un coût de suppression, d'insertion ou de changement fixé à 1, on obtient 2 traductions correctes sur 50 (une distance d'édition égale à 0), ce qui donne les valeurs suivantes :
Distance minimale = 0. Distance moyenne = 21,88. Distance maximale = 88. Rappel =0,04
La valeur du rappel trouvée signifie qu'il y a très peu de traductions brutes qui correspondent à des traductions fonctionnelles de référence c'est à dire que la sortie d'un système de traduction pour tels SMS (dans notre cas un non Français qui a traduit les SMS) est très différente de ce que peut dire un Français.

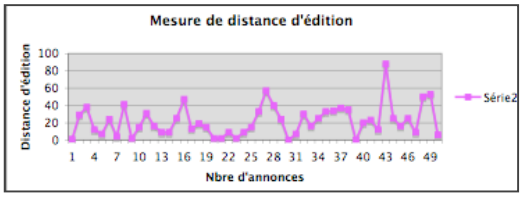

Figure 6: mesures de distance d'édition

Nous mesurons la similarité des traductions brutes et naturelles (fonctionnelles) définie comme suit :

$$s(i,j)=1/1+\alpha \times (edit(i,j)/min(length(i),length(j)))$$

Remarque : α est un nombre positif qui sert de paramètre. La valeur usuelle employée est 1.

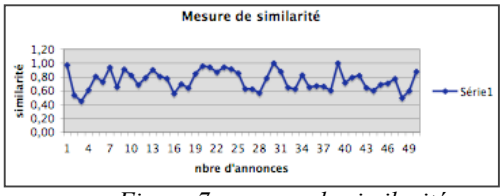

Figure 7:mesures de similarité

Pour un seuil de similarité fixé à 0,90, nous avons obtenu une similarité moyenne de 0,75 et 20% de traductions correctes.

4.2.2 Conclusion

La valeur de similarité est très faible, cela signifie que notre hypothèse H posée au départ est vraie : **H les traductions directes d'un corpus réel constitué de phrases naturelles ne donnent pas de résultats naturels en français.**

4.3 Production du corpus équivalent

Nous avons utilisé l'approche classique pour produire un petit corpus français fonctionnellement équivalent au corpus arabe initial. Afin de développer ce corpus, nous avons adopté la technique suivante : à partir du modèle d'annonces composé de 50 types de SMS révisées et jugées fonctionnelles, nous avons multiplié le nombre de ces annonces en formant des combinaisons différentes des arguments utilisés (type, modèle, année, couleur, prix…). Il s'agit de remplacer par exemple une année par une autre « je cherche une voiture modèle 98 » par « je cherche une voiture modèle 99 »…) Une marque par une autre « A vendre BMW rouge » par « A vendre PEUGEOT noire »…
Il reste à améliorer encore le corpus obtenu pour pouvoir tester et démarrer le système CATS porté vers le français en supprimant des tas de mots inutiles (articles, prépositions longues, etc), en introduisant des abréviations usuelles dans les petites annonces et en ajoutant des abréviations de type phonétique ("ke" pour "que", "g" pour "j'ai", etc.)
Cette méthode nous a permis d'obtenir un corpus fonctionnellement équivalent en français constitué de quelques centaines d'annonces, il est possible de le faire évoluer à des milliers d'annonces.

5 Conclusion et perspectives

Un nouvel axe de recherche a été décrit dans cet article sur la définition, la recherche, la production d'un type très particulier de corpus : c'est les corpus naturels, réels, spontanés rédigés par des êtres humains dans leurs langages maternels. Leur point fort est la réalité du contenu vu qu'ils sont construits d'une façon très naturelle, ce qui garantit l'évidence de l'hypothèse H (testée pour le français) que nous avons prouvé dans ce travail: **les traductions directes d'un corpus réel constitué de phrases naturelles ne donnent pas de résultats naturels en français.**
Nous avons proposé quelques méthodes de production de corpus réels. Nous avons obtenu un corpus réel en français avec lequel nous avons démarré et testé le système CATS.
Une perspective importante de ce travail serait de produire de tel corpus en utilisant le passage pivot CRLcats comme c'est décrit dans la section Réflexions sur les méthodes possibles de multilinguisation du système CATS. Ce pivot nous permettra d'obtenir des corpus réels multilingues, par la suite nous avons l'idée de construire un système de traduction pour les SMS basé sur l'ensemble des corpus réels multilingues obtenus.

6 Références

[1] Achille, FALAISE " Constitution d'un corpus de français tchaté," TALN, 6-10 Juin 2005

[2] Daoud Maher DAOUD " Arabic Deconversion: Problems and Prospects " UNL 99 European workshop, Perugia, Italy, July 1999.

[3] Daoud Maher DAOUD " Building SMS-based System using Information Extraction Technology " ACIDCA-ICMI'2005 Tozeur, Tunisia., 5th to 7th November 2005

[4] Jean-Gabriel GANASCIA Extraction automatique de motifs syntaxiques, TALN 2001 Tours, 2-5 juillet 2001.

[5] NESPOLE « *NEgotiating through SPOken Language in E-commerce* ». http://nespole.itc.it/

[6] R.WAGNER & Michael.FISCHER *The String-to-String Correction Problem* ACM Journal of the Association for Computing Machinery, Vol. 21, No 1 Janvier 1974.

[7] Yves LEPAGE (*traduction par analogie)*, MTS-2006, Pukhet.

[8] UNDL The Universal Networking Language specification, , Http://www.undl.org