SMS et TAL : kL 1Trè* ? (*SMS et TAL : Quel intérêt ?)

Gaëlle Chabert

▶ **To cite this version:**

Gaëlle Chabert. SMS et TAL : kL 1Trè* ? (*SMS et TAL : Quel intérêt ?). Linguistique. 2010. dumas-00561995

HAL Id: dumas-00561995

https://dumas.ccsd.cnrs.fr/dumas-00561995v1

Submitted on 2 Feb 2011

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.

UNIVERSITÉ Stendhal GRENOBLE 3

SMS et TAL : kL 1Trè ?*

*SMS et TAL : Quel intérêt ?

Nom : Chabert
Prénom : Gaëlle

UFR Sciences du Langage

Mémoire de Master 2 recherche - 30 crédits

Spécialité : Modélisation et Traitements Automatiques en Industries de la Langue

Parcours : Traitement Automatique de la Langue Ecrite et Parlée (TALEP)

Sous la direction de M. Antoniadis et Mlle Zampa

Année universitaire 2009-2010

SMS et TAL : kL 1Trè ?*

*SMS et TAL : Quel intérêt ?

Nom : Chabert
Prénom : Gaëlle

UFR Sciences du Langage

Mémoire de master 2 recherche - 30 crédits

Spécialité : Modélisation et Traitements Automatiques en Industries de la Langue

Parcours : Traitement Automatique de la Langue Ecrite et Parlée (TALEP)

Sous la direction de M. Antoniadis et Mlle Zampa

Année universitaire 2009-2010

Remerciements

Je souhaiterais remercier les deux directeurs de ce mémoire : Georges Antoniadis et Virginie Zampa. Merci à eux pour leurs idées (parfois un peu compliquées…) et leur soutien qui ont fait avancer ce mémoire.

Pour mes questions techniques qui ont trouvé leur réponse, je remercie Oliver Kraif et surtout Thomas Lebarbé.

Je remercie l'équipe d'enseignant du Master IDL pour ces deux années d'apprentissage.

Enfin, je remercie Landry, qui aura supporté mon stress pendant ces deux ans.

Sommaire

Introduction

Un « SMS » Short Message Service peut contenir de 140 à 160 caractères. Il a été créé pour permettre une communication de service de l'opérateur téléphonique à l'usager.

Il est rapidement devenu un mode de communication courant pour les usagers. Il peut en effet, être très pratique puisqu'il permet une communication « silencieuse », discrète et rapide.

Les SMS (ou « textos », mot employé au Canada et en France) sont aujourd'hui utilisés à la télévision comme moyen de vote ou encore comme support pour de la publicité.

Étant donné la longueur limitée d'un SMS, un langage SMS s'est mis peu à peu en place. Ce langage est constitué de mécanismes d'écriture permettant d'abréger un texte.

Actuellement, le langage SMS n'est plus seulement présent sur nos téléphones mobiles mais il a envahi aussi les forums de discussion sur Internet, la publicité, la littérature, etc.

Après avoir examiné les usages et usagers des SMS, nous observerons le langage SMS, soupçonné d'être un danger pour l'orthographe. Nous verrons ensuite les applications élaborées à partir des SMS puis ce que nous avons mis en place afin d'élaborer un corpus de SMS dans le cadre du projet « sms4science ».

Chapitre 1 – Usagers et Usages des SMS

Il nous a paru important de s'attarder sur les usages et usagers des SMS puisqu'ils vont avoir un impact sur le contenu et la forme des SMS.

Usagers des SMS

L'AFOM (Agence Française des Opérateurs Mobiles) et TNS Sofres ont mené entre 2006 et 2008 des études auprès de personnes équipées de mobiles, âgés de 12 ans et plus, pour en savoir plus sur leurs usages du téléphone mobile.[1]

Nous proposons ci-dessous un tableau synthétisant les informations des études de l'AFOM et TNS quant à l'utilisation des SMS « au moins de temps en temps » selon l'âge et l'année des utilisateurs :

	2006	2007	2008
Ensemble des équipés mobiles	71%	72%	79%
12-24 ans	98%	97%	97%
25-39 ans	87%	85%	93%
40 ans et +	49%	56%	59%

L'utilisation des SMS pour la jeune population (12-24 ans)

Des usages particuliers pour la tranche d'âge de 12 à 24 ans sont ressortis de l'étude menée en 2008 auprès de 1200 personnes :

72% des personnes interrogées ne trouvent pas convenable de lire ses SMS ou ses mails sur son mobile pendant un repas de famille. Pourtant, 58% des 12-24 ans l'ont déjà fait.

[1] Cf. site Internet www.afom.fr

56% des personnes interrogées ne trouvent pas convenable de faire une déclaration d'amour par SMS mais 52% des 12-17 ans pensent l'inverse.

Ecrire en langage SMS …

… à un ami : cela parait convenable pour 62% des personnes interrogés et pour 93% des 12-24 ans.

… à son conjoint : cela parait convenable pour 58% des personnes interrogées et pour 87% des 12-24 ans.

… à ses parents : cela parait convenable pour 60% des personnes interrogées mais ne parait pas convenable pour 50% des 12-24 ans.

Nous pouvons expliquer le choix des jeunes pour ce moyen de communication simplement :

a) le coût que représente un appel téléphonique par rapport au coût d'un SMS.

b) l'indépendance par rapport au téléphone fixe familial (Fairon et al, 2006).

c) la performance ludique : coder et décoder les SMS (Fairon et al, 2006). Il a été relevé, dans le corpus du CENTAL des SMS « codés » dont le but pour le destinataire est de trouver le code afin de comprendre le message.

d) les nouveaux usages qui se sont développés : C. Martin (2007) et C.-A. Rivière (2002) ont étudié les relations amoureuses dans les SMS. En effet, cette nouvelle forme de communication à distance permet à des adolescents timides de dévoiler plus facilement leurs sentiments.

Les échanges de SMS entre deux personnes amoureuses peuvent s'apparenter à des échanges épistolaires. Au sein d'une relation à distance, les SMS permettent de garder un lien, de faire part de son manque de l'autre, etc.

Cependant, l'utilisation des SMS n'est pas réservée qu'aux adolescents. Les études menées par l'AFOM et TNS Sofres entre 2006 et 2008 ont montré que l'utilisation des SMS « au moins de temps en temps » , chez les personnes de 40 ans et plus a augmenté de 10% entre 2006 et 2008.

J-P Jaffré (2006) avance même que les adultes sont tout à fait aptes à comprendre les procédés du langage SMS, comme les instituteurs qui vont « décoder » les dictées de leurs élèves, malgré les fautes d'orthographe.

Usages du SMS :

Nous présenterons dans cette partie, les usages attribués aux SMS par C.-A. Rivière (2002) et G. Gaglio (2004).

G. Gaglio précise qu'un même utilisateur peut avoir plusieurs usages. Par exemple, un adolescent utilisera le SMS comme moyen de communication pour avertir ses parents d'un éventuel retard, puis pour créer une conversation de SMS intimes avec sa petite amie et pour organiser une soirée avec des amis.

Les usages ludiques

Envoyer des SMS pour un adolescent peut s'apparenter à un passe-temps. Converser avec des SMS serait une façon de se distraire.

Les usages pratiques ou fonctionnels

Envoyer des SMS évite une conversation téléphonique à cause du coût qu'elle représente ou bien par manque de temps.

Un SMS peut se lire plus tard par le destinataire s'il est occupé, ainsi l'envoi d'un SMS peut apparaitre comme un moyen de ne pas déranger.

Un utilisateur peut préférer le SMS à une conversation téléphonique lorsque cette conversation est impossible ou difficile. En effet, ça peut être le cas dans les lieux publics, ou dès qu'il y a des personnes autour de lui dont il n'a pas envie qu'elles partagent sa conversation.

Les usages mettant en jeu l'affect

Exprimer ses sentiments peut être plus facile par SMS, ce dernier permet donc d'extérioriser et d'exprimer ses émotions.

Les usages de contact

Un SMS peut être utilisé pour demander des nouvelles à une connaissance. Un SMS est pratique aussi pour prévenir d'un retard.

Chapitre 2 – Langage SMS

En fonction des usages et des usagers des SMS, le texte du SMS va varier. S'il correspond à une réponse qui se veut la plus rapide possible ou s'il permet à l'expéditeur de donner des nouvelles de lui en un minimum de caractères, le texte du SMS va subir des procédés pour réduire le nombre de caractères.

En plus des abréviations connues des sténographes, des abréviations plus spécifiques au SMS vont faire leur apparition, comme « jtd » (je t'adore). Des procédés nouveaux vont aussi faire leur apparition. Se pose alors une question : peut-on parler d'un nouveau langage : le langage « SMS » ou bien s'agit-il uniquement d'une liste de procédés ? Nous essayerons de répondre à cette question dans cette partie.

Peut-on parler d'un « langage SMS » ?

De nombreux linguistes étudient cette question.

Le « langage SMS » peut être rapproché avec une langue étrangère que nous pouvons apprendre puisqu'il existe des cours de SMS. P. Marso (2005) commence par publier un ouvrage pour apprendre cette nouvelle langue : « CP SMS » puis il propose une classe de PMS (Phonétique Muse Service) dans un collège parisien à des adolescents de 14-15 ans qui refusent l'école ou qui sont en échec scolaire. Le langage « PMS » est un langage dérivé du langage SMS qui vise à rendre l'écriture SMS plus lisible avec notamment l'insertion de l'apostrophe dans un mot tel que « K'ré'C » (caresser) pour permettre une meilleure lisibilité des procédés d'abréviations.

Pour S. Vienney et C. Melian (2004), il existe bien un langage SMS mais les procédés utilisés ne sont pas nouveaux. J. Véronis (2004) indique que, dans un papyrus d'Egerton du IIème siècle des formes simplifiées ont été relevées. Plus tard, des cours de sténographie ont été créés pour permettre aux secrétaires d'abréger au maximum un texte pour pouvoir l'écrire aussi vite que la parole. Les auteurs citent l'exemple des lycéens qui utilisent fréquemment des abréviations dans leurs notes de cours.

J. Anis (2001) n'apporte pas de réponse définitive sur l'existence ou non d'un langage SMS. Il précise que « le langage a toujours été associé aux progrès des techniques et du savoir ». Etant données les contraintes que posent les nouvelles technologies, le langage doit s'adapter (autant sur le contenu que sur la forme). Ainsi, pour J. Anis, le langage SMS ne serait pas un nouveau langage mais une série de procédés qui viennent s'ajouter à notre langue. Il souligne comme J. Véronis que les procédés d'abréviation ne sont pas nouveaux : par exemple, le mot « Monsieur » qui s'écrit « M. ».

J. Daugmaudytë et D. Këdikaitë (2006) précisent que le langage SMS est très proche de l'oral tout en étant une forme de l'écrit. En effet, deux utilisateurs de SMS qui vont créer une conversation par ce moyen de communication vont partager le même moment d'énonciation, sans pour autant être l'un en face de l'autre. Pour palier les problèmes des mimiques et de la gestualité que l'on retrouve dans une conversation en face à face, le langage SMS propose des émoticônes (ou smileys), des formes d'écriture différentes, ainsi qu'une utilisation différente de la ponctuation. Par exemple, le smiley « ;-) » peut informer que le discours est ironique. Une ponctuation répétée informera de l'intensité de l'émotion, par exemple : « je t'adore !!!!!!! ».

La répétition de lettres dans un mot peut marquer aussi une forme d'insistance : « viens viiiiiiiite ».

Au-delà du langage SMS, certains auteurs tels que R. Panckhurst, J. Véronis ou E. Guimier De Neef parlent d'un langage lié aux nouvelles technologies.(Véronis, J. , Guimier De Neef, E, 2006 et Panckhurst,R., 2006).

A. Dejond utilise les termes de « cyberlangue » et « cyberlangage » et R. Panckhurst utilise le terme d' « eSMS » pour parler de l'écriture SMS » (2009, page 35).

Nous retrouvons des procédés similaires, que la communication se fasse au travers des SMS, des tchats, des blogs, des forums de discussions, etc.

Ces moyens de communication doivent être efficaces, c'est-à-dire que l'utilisateur doit dire un maximum d'informations en tenant compte d'un espace réduit pour les SMS, forums et blogs et la volonté d'une communication rapide pour les tchats et les SMS.

Construction du langage SMS

J. Daugmaudytë et D. Këdikaitë (2006) mentionnent dans leur article l'expression « construction du langage SMS ».

R. Panckhusrt (2009) a répertorié les éléments de construction de ce langage, c'est-à-dire les phénomènes de l'écriture SMS et plus généralement, des nouvelles technologies .

Elle les classe ces phénomènes en quatre parties : les substitutions, les réductions, les suppressions et les augmentations et ajouts.

Nous présentons ci-dessous :

I. Substitutions
Substitutions phonétisées : - entières : remplacer un son par des caractères uniques (lettres ou chiffres). L'orthographe du lexème est totalement modifiée : o (eau), 7 (cet). - partielles : remplacement de digrammes et trigrammes, qui transcrivent un phonème. L'orthographe du lexème est ainsi partiellement modifiée : ossi (aussi), allé (aller), bo (beau) ; « s » intervocalique : bizes (bises). - avec variation : bisoo (bisou)
Substitutions graphiques - élision, typographie, majuscules : remplacement de l'apostrophe d'élision ou d'un trait d'union, etc. par l'espace, « m en » (m'en), « est ce que » (est-ce que) ; mise en majuscules de l'ensemble d'un message ou, au contraire substitution majuscules/minuscules. - icônes, symboles mathématiques, caractères spéciaux, rébus : (*,+ = => @) ; à+ (à plus), de grandes @ (de grandes oreilles) - avec variation : bisoux (bisous), mwa (moi)
II. Réductions
Réductions phonétisées : - abrègements morpho-lexicaux :

- troncations : ordi (ordinateur, apocope), ‘lut, Net (salut, Internet, aphérèse). - sigles/acronymes : ASV (âge, sexe, ville), mdr (mort de rire), tvb (tout va bien), tlm (tout le monde), lol (laughing out loud). - avec variation : ui (oui), i (il).
Réductions graphiques - suppression de fins de mots muettes : échange (échanges), vou (vous), peu (peut), chian (chiant), fou (« m’en fous »), chute de « e » instables : douch (douche). - squelettes consonantiques & abréviations : dc (donc), pr (pour), ds (dans) ; consonnes doubles : ele (elle), poura (pourra) ; abréviations sémantisées (abréviations réduites à l’initiale) : t (te/tu) p (peux). - agglutinations : jattends (j’attends)
III. Suppression/ absence ou raréfication :
Suppressions graphiques - typographie & ponctuation - signes diacritiques : ca (ça), voila (voilà)
IV. Augmentations et ajouts :
Augmentations et ajouts graphiques - répétitions de caractères et/ ou de signes de ponctuation : suuuuupppeeerrr !!!!! - représentations sémiologiques (smileys/ binettes) :-) - ajout de caractères : oki (ok), les zamours (les amours) - onomatopées : mouarf, arfff, bof.

Nous avons vu que le langage SMS amène l’utilisateur à abréger les mots de sa langue française et ainsi à la modifier. De plus, nous avons vu précédemment que les jeunes utilisateurs écrivent beaucoup de SMS. C’est pourquoi l’usage de ce langage inquiète, notamment les professeurs qui ont peur de voir ce langage envahir leurs copies.

Chapitre 3 – Le langage SMS représente-t-il un danger pour l'orthographe ?

Deux opinions s'opposent : les défenseurs de la langue française pensent que les SMS représentent un danger pour l'orthographe et la grammaire, déjà malmenées par les jeunes apprenants du français, d'autres pensent que les SMS sont une façon de jouer avec la langue française et reflète l'inventivité de son créateur.

Heureusement des linguistes adoptent une approche plus scientifique et vont s'attacher à étudier ce nouveau langage plutôt que de le juger.

Les usagers différencient, pour la plupart, le langage SMS et le langage écrit plus formel utilisé par exemple en milieu scolaire.

Mais certains procédés du langage SMS pourraient influencer négativement l'acquisition d'une orthographe correcte (Fairon et al, 2006). Citons par exemple, la disparition des lettres muettes en fin de mots.

De plus, l'acquisition d'un téléphone portable avant l'apprentissage de l'orthographe à l'école pourrait être problématique pour un jeune enfant puisqu'il aurait du mal à différencier le langage normé appris à l'école et le langage SMS (Fairon et al, 2006).

Les particularités de l'écriture SMS (telles que l'agglutination ou la troncation) sont considérées comme des erreurs, pourtant, l'équipe du CENTAL qui a travaillé sur les SMS (Fairon et Al. [2006]) pense qu'il est inapproprié de parler d'erreur dans ce cas puisque ce sont des mécanismes volontaire. Rappelons que le langage SMS s'est créé petit à petit afin de réduire un message pour qu'il respecte le nombre de caractère imposé par son format d'un SMS. Ce langage sera par la suite utilisé comme « jeu » (Fairon et Al, 2006) entre adolescents, nous parlons alors de SMS « codés » dont le but est de défier le destinataire à déchiffrer le SMS envoyé.

A. Dejond (2010) conforte l'idée que l'usage de procédés de l'écriture SMS ne doit pas être considéré comme une erreur. Elle nous incite à différencier ces procédés et les réelles fautes d'orthographe. Par exemple, écrire « Hier, j'ai manger chez mon père » sera

considéré comme une faute d'orthographe et rarement observé dans les SMS. Mais écrire « G manG » sera apparenté au langage SMS puisqu'il utilise un mécanisme d'écriture particulier : la phonétisation des deux mots « j'ai » et de la syllabe « gé » de manger.

En France, nous remarquons que le sujet fait débat sur des forums mais l'influence des SMS sur l'orthographe des enfants et adolescents n'a pas été observée de manière scientifique jusqu'à présent.

Au Canada, une étude sur l'influence des SMS sur l'orthographe a été menée par Connie Varnhagen (Varnhagen,C., 2009). Pour cette étude, 40 adolescents de 12 à 17 ans ont gardé l'ensemble de leurs SMS envoyés pendant une semaine. A la suite de cette semaine, C. Varnhagen et son équipe ont testé la capacité des adolescents à épeler correctement les mots. Cette étude a montré que les adolescents ayant une bonne maitrise de l'orthographe conserve cet avantage dans leurs SMS. La même chose se produit pour ceux qui ne maitrisent pas l'orthographe, c'est-à-dire que leurs SMS vont souffrir d'une mauvaise orthographe. Pour l'auteur de l'étude, écrire un SMS (en langage SMS) serait un bon moyen de faire « fonctionner les neurones » des adolescents puisqu'ils vont réfléchir à ce qu'ils veulent dire au destinataire du message et comment le dire avec le moins de caractères possibles.

Chapitre 4 – *Applications élaborées à partir des SMS*

Nous avons vu précédemment que les SMS attirent l'intérêt. Nous allons voir, dans la partie suivante, des applications élaborées pour et à partir des SMS afin de mieux les comprendre, les étudier et les utiliser.

Dictionnaire de SMS

Il existe sur Internet des dictionnaires de SMS (www.dictionnaire-sms.com, www.sos-sms.net, www.deblok.net/dicosms, pages consultées le 24 février 2010) et même une version papier créée par Phil Marso (Marso, P., 2005).

Cependant, ces dictionnaires représentent une ressource limitée puisqu'ils sont parfois obsolètes tellement le langage SMS évolue rapidement. De plus, ils ne prennent pas en compte la totalité des mécanismes de création d'un SMS, ou ne propose qu'une « version » d'écriture. En effet, il existe pour un même mot une vaste quantité de variant : exemple pour « *demain* » : 2m1, *dmain, dmin, 2main, dem1, 2min, dems, 2m, d2m1, 2mains, dem's, dms, dmai, dem'.(Fairon et al., 2006).*

Ces dictionnaires peuvent être utiles pour traduire un SMS reçus, à condition qu'ils soient mis à jour régulièrement et que le problème des formes multiples soit résolu.

TiLT : transcripteur de SMS (Traitement Linguistique de Textes)

Le logiciel TiLT a été développé par France Télécom R&D en 2007, au sein d'un projet de vocalisation de SMS sur téléphones fixes. Il a pour but de transcrire les SMS afin de rendre possible leur vocalisation.(Guimier De Neef, E., Fessard,S., 2007)

Architecture de TiLT

Le logiciel TiLT est composé de trois modules :

a) un module de segmentation qui découpe le message SMS en typant les différents éléments du message (mot ou smiley).

b) un module d'analyse lexicale qui fait correspondre un élément du SMS avec un élément du lexique.

Le lexique comporte les abréviations et sigles courant du langage SMS ainsi qu'environ 3 000 prénoms. Il a été élaboré à partir d'un corpus.

Ce module peut opérer des corrections. Par exemple, la forme étirée « *viiiiiiiiiite* » sera corrigée et rendue à sa forme standard « *vite* ».

c) un module d'analyse en chunks qui permet de choisir la meilleure correction en fonction du contexte syntaxique. Par exemple, une correction différente sera appliquée à « ***c*** » dans l'extrait de SMS suivant : « *je* ***c*** *pa ki* ***c*** » -> « *je* ***sais*** *pas qui* ***c'est*** ». Cette analyse en chunks utilise une grammaire hors contexte de 2 000 règles environ.

Le logiciel TiLT a été évalué sur deux corpus : une partie du corpus du DELIC (environ 9 700 SMS) et le corpus du CENTAL de 30 000 SMS. Cette double évaluation est intéressante puisque les corpus n'ont pas le même nombre de SMS, de plus, le premier est français alors que le second est belge.

TiLT a transcris les SMS des deux corpus et l'évaluation a consistée à comparer la transcription de TiLT avec une transcription manuelle.

Pour cette évaluation, ce sont la mesure BLEU et le coefficient de Jaccard qui ont été utilisés pour évaluer la similarité entre le SMS traduit manuellement et le SMS traduit par TiLT.

La mesure Bleu est une mesure logarithmique entre les taux de 1-gram, 2-gram, etc, en commun entre deux éléments.

Le coefficient de Jaccard représente le rapport entre le nombres de mots communs entre les deux traductions et l'ensemble des mots des deux éléments auquel on soustrait le nombre de mots communs.

Evaluation de TiLT avec le corpus du DELIC

Le système TiLT rencontre des problèmes quand le SMS ne présente pas de séparateur.

Par exemple, le SMS suivant : *Bonnefeteprofitesbiendevotredernierjourdevacance*

Sera traduit comme suit par TiLT : *Bonnefeteprof il t'est biendevotredernierjourdeva cance*

Une autre difficulté pour le système TiLT est de prendre en compte plusieurs procédés d'écriture SMS dans un même segment : *je ne pep a mpaC dtoi -> je ne peux pas me passer de toi* (procédés : agglutination et phonétique).

Le système TiLT n'arrive pas à corriger certains segments de SMS parce qu'il ne possède pas de grammaire qui établisse de liens entre la tête et ses dépendants dans une phrase. Ainsi, il ne corrigera pas la phrase : *tes vacances se passe bien*.

Lors de l'évaluation de TiLT, il a été observé un traitement erroné pour certains noms propres. Etant donné l'anonymisation dans le corpus, ce problème n'en est pas vraiment un.

Par exemple, le SMS « *gros bisous à vous tous caro* » sera traduit « *gros bisous à vous tous carreau* ».

Evaluation avec le corpus du CENTAL

Les mêmes limites que pour le corpus de DELIC ont été observées. Mais, en plus, ce sont rajoutés les problèmes dus aux belgicismes que le système corrige alors qu'il faudrait les laisser tels quels.

Correcteur automatique proposé par S. Vienney et C. Melian

Le correcteur automatique appliqué au SMS de S. Vienney et C. Melian (2004) se découpe en cinq étapes : la lecture du texte source, la segmentation du SMS, la transcription en français standard, un module d'analyses morphologiques, syntaxiques et sémantiques et enfin, la proposition d'un texte cible, correction du texte source.

Certains phénomènes de construction du langage SMS ne sont pas simples à prendre en compte pour réaliser une segmentation. Prenons par exemple, l'agglutination et

l'utilisation de sigles : comment faire pour découper « jallais » (j'allais) ou « tkt » (t'inquiètes) ?

Le module de transcription en français standard nous apparait très complexe. Il se fonde « sur un ensemble de règles de transcription traitant l'ensemble des phénomènes de néographies […] puis il calcule des hypothèses de transcription avec une analyse lexicale et combinatoire » (Vienney S. et Melian C., (2004) page 193).

Les transcriptions qui résultent de ce module vont être validées ou non par l'analyse morpho-syntaxique et sémantique.

A la suite de ces traitements, le système transmet à l'utilisateur un texte en français standard. Cependant, quelques ambigüités demeurent comme, par exemple, avec des systèmes de traitement automatique de l'oral : les pois sont verts/ les poissons verts.

Les smileys amènent une difficulté supplémentaire. Les auteurs proposent de les remplacer par « une structure sémantique équivalente » ou de les supprimer « s'ils sont utilisés pour « signer » le texte : *je viens 2m1 ;-)* ». Nous trouvons la dernière solution non adéquate puisque même en tant que signature, le smiley transmet une information.

Architecture pour le traitement automatique des SMS proposé par F. Yvon. (Yvon, 2008 et Kobus et Al., 2008)

Après avoir observé une proximité entre les formes d'écriture utilisées dans les SMS et la langue orale, F. Yvon (2008) propose un système de normalisation des SMS inspiré des systèmes de reconnaissance de la parole.

Ce système est optimisé pour un système de vocalisation de SMS puisqu'il passe par une étape de phonétisation.

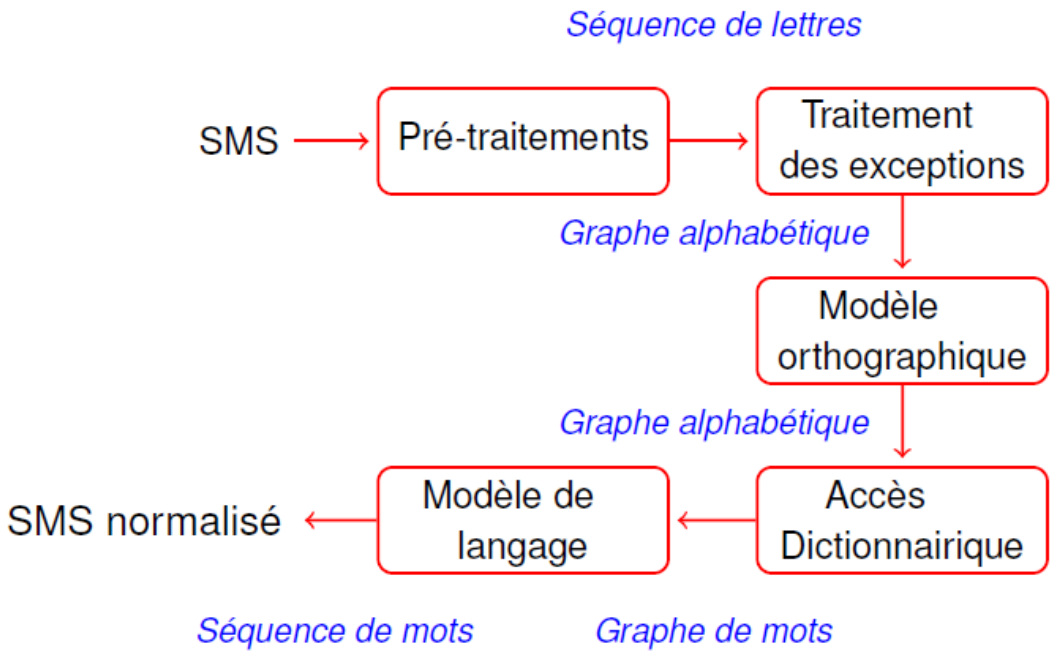


Schéma du système (Kobus et al., 2008, page 130)

Présentation des modules du système de F. Yvon

Nous reprenons chaque module du système pour les détailler :

Dans un premier temps, le SMS passe dans le module de traitement. Ce dernier converti le message en entrée en un automate fini. Puis il réalise des opérations de normalisation telles que le traitement des chiffres, l'insertion de marques de débuts et fin de phrase et fin de mots, la suppression de la ponctuation, des majuscules et des smileys.

Le second module va traiter les exceptions en utilisant un dictionnaire des formes abrégées avec leur correspondance en français standard. Exemple : pr => pour. Ce dictionnaire a été élaboré manuellement à partir d'une étude de corpus.

Le troisième module de modèle orthographique vise à modéliser les récritures graphiques en traitant les phénomènes tels que l'absence d'accents, les fautes de frappes, l'écriture phonétique et les rébus. Il comporte deux transducteurs, l'un pour établir les

modèles d'erreur et l'autre pour les modèles phonétiques. Chaque élément du message va être phonétisé en fonction de son contexte. Ce module crée une liste de phonétisations possibles. Les exceptions détectées au module précédent vont être phonétisées à partir d'un dictionnaire de prononciation.

Le quatrième module correspond à l'accès dictionnairique. Une fois la liste de phonétisations possibles établie, le module utilise un dictionnaire de prononciation qui associe une séquence de phonèmes en séquences de mots. Si les mots du message sont connus, ils seront associés aux mots du dictionnaire, par contre, s'ils sont inconnus, ils seront associés à « <unk> ».

Le cinquième module applique un modèle de langage statistique de type n-gram pour faire ressortir la séquence de mots la plus probable.

Avantages et Inconvénients du système

Le système corrige les agglutinations mais ne corrige pas les erreurs de segmentation. Par exemple , dans la phrase « *je ne pep a mpaC dtoi* », le segment « pep a » sera traduit « *pe p a »* au lieu de « *pe pa »*.

L' implémentation avec des transducteurs permet de gérer les agglutinations telles que « lbac blanc ».

Le système crée beaucoup d'erreurs de type « *c* » traduit en « *c'est* » au lieu de « *sait* ». Mais ce sont des erreurs qui ne sont pas significatives si le système est utilisé pour la vocalisation de SMS.

F. Yvon (2008) concède lui-même que son système pourrait être amélioré. Il faudrait traiter les accords, les dépendances longue distance, le mélange des langues (utilisation de mots anglais notamment), les ponctuations, les smileys, les structures

consonantiques, etc. De plus, les ressources pourraient être améliorées (lexiques, règles, modèles d’erreurs et de langage).

Chapitre 5 – Constitution de corpus SMS

Nous avons vu précédemment que les SMS connaissent un essor considérable. De plus, l'utilisation du langage créé par ce moyen de communication s'étend en dehors du téléphone mobile. Il parait intéressant de s'attarder sur ce langage et ses spécificités. Pour cela, il est nécessaire de pouvoir observer des corpus de SMS. Nous verrons dans cette partie, quelques corpus de SMS et comment ces derniers ont été récoltés.

Corpus du DELIC (utilisé pour tester le système TiLT)

Le corpus compte 13 400 SMS pour environ 156 620 mots.

Une partie des SMS ont été récoltés entre 2000 et 2004 par des étudiants de l'université de Provence, dans le cadre de travaux pratiques ou mémoires.

Se sont rajoutés des SMS provenant d'utilisateurs ayant donné leur accord à Orange pour que leurs SMS soient utilisés à des fin de recherche.

Seuls 9 700 SMS ont été utilisés pour l'évaluation du système TiLT, parce qu'ils ont été traduits manuellement (Hocq, 2006).

Corpus de l'Université de Singapour : « NUS SMS Corpus »[2]

Yijue How et Mingfeng Lee ont constitué un corpus de 10 117 SMS anglais. Un appel par mail a été lancé aux étudiants de l'Université de Singapour pour récolter ces SMS, par le Département de recherche en Informatique (the Department of Computer Science at the National University of Singapore).

Les participants, de 18 à 22 ans, tapaient leur SMS (en anglais, qu'il soit reçu ou envoyé) et leur numéro de téléphone dans un formulaire présent sur le site Internet dédié à la récolte. Ainsi, 6 167 SMS ont été collectés.

Puis 602 SMS ont été récoltés sur un chat SMS sur le site « Yahoo ».

[2] Site du corpus avec lien pour le télécharger :
http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus/

Corpus SMS de Caroline Tagg

Caroline Tagg, doctorante à l'université de Birmingham en Angleterre, a analysé l'orthographe, la grammaire et les abréviations dans les SMS (Tagg, C., 2009). Pour son étude, elle a demandé à ses amis et aux membres de sa famille de lui renvoyer tous les messages qu'ils recevaient ou envoyaient. Elle a ainsi récolté environ 11 000 SMS, soit près de 190 000 mots.

Projet « SMS4science »

Le projet en Belgique

Le projet, qui se veut international, a débuté en Belgique en 2004. Il a permis de récolter plus de 75 000 SMS grâce à 3600 participants. Parmi ces participants, 2000 ont acceptés de répondre à un questionnaire « portant sur leur profil et leurs pratiques ».

Le corpus constitué est à la fois homogène (il ne s'agit que de SMS, pas de chat, etc.) et diversifié puisque les participants sont différents donc leur utilisation des SMS l'est aussi. En effet, les participants ont entre 12 et 73 ans et appartiennent à des catégories socioprofessionnelles différentes.

La récolte des SMS s'est faite par un numéro spécial. Les participants envoyaient leurs SMS directement à ce numéro gratuit. Ce procédé a évité des erreurs de saisie ou d'interprétation en recopiant manuellement les SMS. C'est un problème qui a été observé dans des projets d'analyse de SMS qui se sont réalisés dans différentes universités.

Depuis 2007, huit partenaires se sont joints au projet SMS 4 science :

- La France,
- La Réunion,
- Le Royaume-Uni,
- L'Espagne,
- La Grèce,
- La Suisse,

- Le Pays Basque,

- et le Canada.

Le projet en France : « smsAlpins »[3]

Le projet sera réalisé en collaboration avec le Conseil Régional des Hautes-Alpes et l'Université Stendhal à Grenoble.

La récolte est prévue d'octobre 2010 à janvier 2011.

Le projet à la Réunion : « LaRéunion4science »[4]

La récolte de SMS de plus de 20 000 SMS s'est déroulée du 15 avril 2008 au 30 juin 2008.

L'étude de SMS récoltés à la Réunion reprend les enjeux de l'analyse des SMS en ajoutant une dimension multilingue. En effet, les utilisateurs de SMS sont bilingues français-créole et alternent souvent plusieurs langues (français, créole, anglais, espagnol et même la langue des signes pour les jeunes sourds).

Le projet en Suisse : « sms4science »[5]

Pour ce projet, en Suisse, 26 888 SMS ont été récoltés pendant deux mois, de novembre à décembre 2009. Parmi les 2 674 « donateurs » de SMS, 1 387 ont remplis un questionnaire afin de récolter des informations sur leur âge, leur catégorie socioprofessionnelle, etc.

Ce sont les Universités de Neuchâtel et Zurich ont menés ce projet.

[3] Cf. site Internet du projet : www.sms4science.org
[4] Cf. site Internet du projet : www.lareunion4science.org
[5] Cf. site Internet du projet : www.sms4science.ch

Le projet au Canada : « Texto4science »[6]

Le projet « sms4science » est mené au Canada par Patrick Drouin et Philippe Langlais de l'Université de Montréal, en collaboration avec les Universités d'Ottawa et Simon Fraser. Ils espèrent récolter autant de SMS en français que de SMS en anglais.

La récolte est en cours. Nous savons déjà que 295 participants ont rempli le questionnaire sur le site Internent de l'opération.

Le projet dans les autres pays partenaires

- Pour le Royaume-Uni, le projet va être mené par l'université de Birmingham ;

- Pour l'Espagne, ce sera l'Université Autonome de Barcelone ;

- Pour la Grèce, ce sera l'Université de Thessalonique ;

- Pour le Pays Basque, ce sera l'Institut Elhuyar Fundazioa.

[6] Cf. site Internet du projet : www.texto4science.ca

Chapitre 6 – Présentation des outils réalisés en vue du projet « SMS4science » en France

Dans le cadre du projet « SMS4science », nous espérons recevoir entre 20 000 et 35 000 SMS. Ces SMS seront envoyés par des participants volontaires vers un numéro court. Ces participants seront invités à remplir un questionnaire en ligne pour en savoir plus sur leur pratique du SMS.

Mon objectif pour ce mémoire a été de penser et concevoir des scripts pour exploiter et utiliser les SMS reçus.

Dans un premier temps, nous avons pensé à l'exploitation des SMS de leur réception jusqu'à leur transcription.

Puis dans un second temps, nous présenterons les applications conçues pour utiliser les SMS et des applications à venir.

Exploitation des SMS

Nous présenterons, dans cette partie, ce qui a été mis en place afin d'exploiter les SMS qui seront reçus lors du projet « SMS4science » en France.

Anonymisation des SMS bruts

L'anonymisation des données personnelles au sein des SMS est l'élément primordiale à mettre en place parce que nous sommes tenus d'assurer la confidentialité des participants au projet.

Nous avons choisi de reprendre le protocole utilisé par l'équipe du CENTAL sur le projet « SMS4science » en Belgique afin de rendre les corpus similaires.

Quelles données à anonymiser ?

Quant à l'anonymisation, nous avons choisi de mettre en place des scripts en PHP pour détecter automatiquement certaines données personnelles pouvant apparaitre dans les

SMS. Ces scripts recherchent les numéros de téléphone, les adresses mail, les adresses de site Internet et les coordonnées bancaires dont le format est fixe.

Fairon et al (2008) proposent d'autres éléments à anonymiser dont la détection produirait trop de bruit, comme les numéros de rue, les adresses de blog et numéros divers. D'autres éléments ne sont pas aisés à détecter comme les noms propres à cause de l'emploi multiples des majuscules dans le langage SMS.

Interface d'anonymisation et outils

Nous avons pensé qu'un contrôle était nécessaire pour vérifier si les éléments détectés devaient vraiment être masqués ou si des éléments n'avaient pas été oubliés par les scripts.

Nous avons donc créé une interface pour l'utilisateur qui va anonymiser les SMS. Cette interface va lui afficher les SMS avec les informations détectées comme personnelles, qu'il va pouvoir masquer ou garder, ainsi qu'un champ de texte libre, pour modifier des éléments non détectés.

S'il trouve d'autres éléments à anonymiser au sein du SMS, il va pouvoir le spécifier dans le champ de texte libre. Pour cela, il va encadrer la donnée par le symbole « # » puis il va choisir dans la liste déroulante, le type de donnée détectée. Si l'utilisateur ne trouve pas un type de donnée approprié à son élément, il pourra définir un nouveau type en choisissant « Autre type de donnée » dans la liste déroulante.

Interface d'anonymisation

Transcription des SMS bruts

Choix du terme « transcription » et non « traduction »

Nous parlerons de transcription au lieu de traduction. En effet, la transcription vise à rendre le SMS plus compréhensible pour des utilisateurs connaissant peu ou mal le langage SMS. Elle ne vise pas à traduire le SMS en langage « standard ».

Méthodologie de transcription

Nous avons choisi de faire transcrire les SMS par des individus et non pas par les traducteurs automatiques de SMS qui ont pu être établis. Ce système de traduction va permettre de créer un lexique au fur et à mesure des transcriptions. En effet, nous proposons à l'utilisateur de transcrire le texte du SMS mot par mot. C'est en récupérant l'association de chaque mot à traduire par chaque mot traduit que nous créons ce lexique.

Deux méthodes de découpage du texte du SMS

Nous proposons à l'utilisateur un découpage automatique, fondé sur des séparateurs définis tels que les ponctuations ou espaces. Si ce découpage ne convient pas à l'utilisateur, nous lui proposons de le modifier. Il va ainsi pouvoir découper de nouveau des mots qui n'auraient pas été séparés automatiquement, en insérant le symbole « # » entre chaque mot. Cette possibilité sera intéressante lorsqu'on rencontrera des SMS écrit sans espaces et sans ponctuations, comme a pu le voir l'équipe du Cental.

Outils pour la transcription

Une fois le découpage effectué, chaque mot du texte sera affiché avec une zone de texte chargée de recueillir la traduction de l'utilisateur. Nous proposons à l'utilisateur quatre autres outils pour la transcription :

a) une case à cocher pour garder l'intégralité du SMS s'il n'a pas besoin d'être modifié.

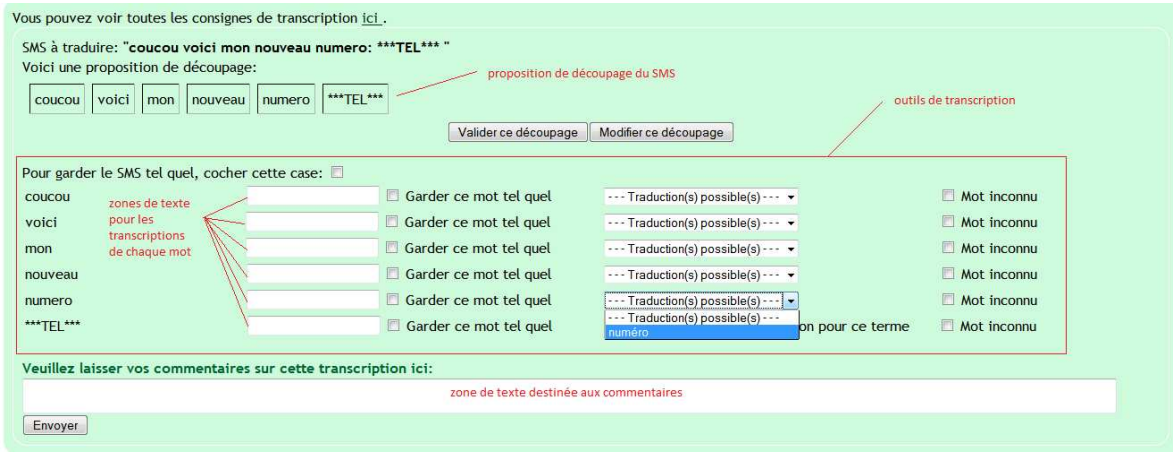
b) une case à cocher pour garder le mot à transcrire s'il n'a pas besoin d'être modifié.

Ces deux cases à cocher ont été créées après observation d'un extrait du corpus des SMS récoltés en Belgique pour le projet « SMS4science ».

c) une liste déroulante proposant les traductions possibles pour ce mot (elle s'appuie sur le lexique constitué au fil des transcriptions).

d) une case à cocher « Mot inconnu » pour permettre à l'utilisateur de spécifier qu'il n'arrive pas à traduire ce mot. Il parait primordial de donner la possibilité à l'utilisateur de ne pas savoir transcrire certains termes. Lors d'une relecture, un autre utilisateur pourra combler les lacunes.

Nous proposons pour chaque transcription, une zone de texte libre, afin que l'utilisateur ajoute des commentaires sur la transcription proposée.



Interface de transcription

Applications des SMS mises en place

Lexique SMS-Français et Lexique Français-SMS

Il parait primordial de concevoir un lexique SMS-Français pour permettre à des utilisateurs de « décoder » leurs SMS.

Nous avons choisi de créer aussi un lexique Français-SMS pour proposer aux utilisateurs un outil afin de simplifier et réduire leurs SMS.

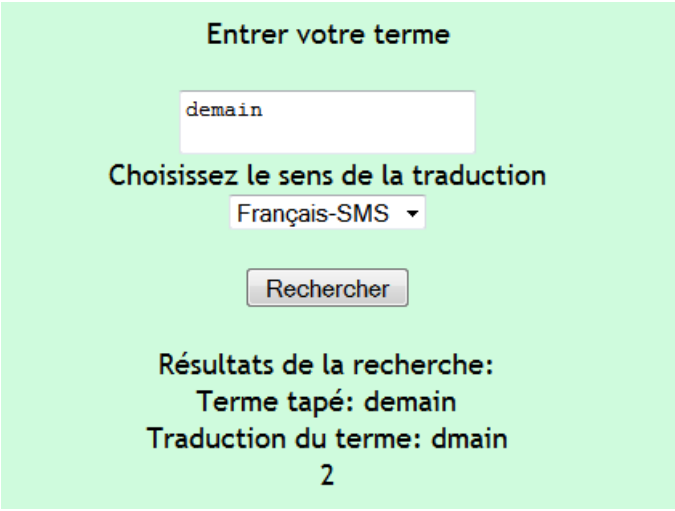
Dans chacune des interfaces, nous proposons un bouton pour chaque lettre de l'alphabet qui va afficher les termes du lexique commençant par la lettre associée au bouton. Le bouton « All words » va afficher tous les mots du lexique.

Interface du lexique

Transcripteur de terme

Nous proposons un moteur de recherche, où l'utilisateur va saisir le mot qu'il cherche et choisir le sens de la traduction (SMS-français ou français-SMS). Si le terme saisi est bien présent dans le lexique, nous le lui afficherons ainsi que sa ou ses traductions.

Nous lui affichons aussi la fréquence d'apparition de la traduction dans le corpus.



Transcripteur

Applications des SMS souhaitées

Interface de consultation des SMS

Il apparait nécessaire de créer une interface de consultation des SMS. Nous pourrions envisager le choix des SMS à consulter en fonction des informations des expéditeurs grâce au questionnaire remplis sur Internet : sexe, âge, nombre de SMS envoyés par semaine, etc.

Nous pourrions penser à une consultation par thématique des SMS. Il serait intéressant de consulter les SMS selon qu'ils traitent des relations amoureuses, de l'organisation d'une soirée entre amis, etc.

Etiquetage morphosyntaxique

Un étiquetage morphosyntaxique appliqué aux SMS pourrait mettre en avant quelles catégories syntaxiques sont les plus touchées par les procédés d'abréviations du langage SMS. Il pourrait aussi être utile afin de faire une recherche d'informations dans les SMS.

Traduction d'un texte entier (Français-SMS et SMS-Français)

La traduction d'un texte entier s'appuierait sur les contextes gauche et droit des mots de la phrase grâce à l'alignement des SMS originaux et leur transcription, ainsi que sur l'étiquetage morphosyntaxique pour réduire au maximum les ambigüités.

Chapitre 7 – Description technique des outils réalisés en vue du projet « SMS4science » en France

Nous présenterons dans cette partie les points techniques de ce que nous avons développé pour le projet « SMS4science ».

Présentation de la Base De Données *(voir schéma page 35)*

Notre base de données a été implémentée sous PHP MySQL. Elle comporte quatre tables :

a) La table des **données personnelles** répertorie les types de données à masquer dans les SMS au moment de l'anonymisation. Nous avons établi une première liste de données grâce aux éléments renseignés par l'équipe du CENTAL dans leur ouvrage sur le projet « sms4science » mené en Belgique (Fairon et al., 2006). Cette table pourra être complétée par la suite si un utilisateur détecte une donnée dont le type n'est pas répertorié.

b) Le **lexique** contient chaque mot des SMS d'origine associé à leur traduction. Cette table sera remplie au moment de la transcription des SMS. Si un mot n'a pas pu être traduit, nous noterons « Non traduit » comme traduction.

c) La table des **SMS** contient pour chaque SMS, son identifiant et l'identifiant de l'expéditeur, le texte original, le texte anonymisé, le texte transcris et le texte aligné ainsi que les commentaires de l'utilisateur sur la transcription.

L'identifiant du SMS est incrémenté à chaque enregistrement d'un nouvel SMS.

L'identifiant de l'expéditeur correspond à un codage du numéro de téléphone de celui-ci.

d) La quatrième table contient les données de chaque **expéditeur** de SMS qui aura rempli un questionnaire sur le site Internet de l'opération.

Relier les SMS et ces données nous permettrons d'étudier les variations du langage SMS selon des critères d'âge, de profession, de sexe, etc.

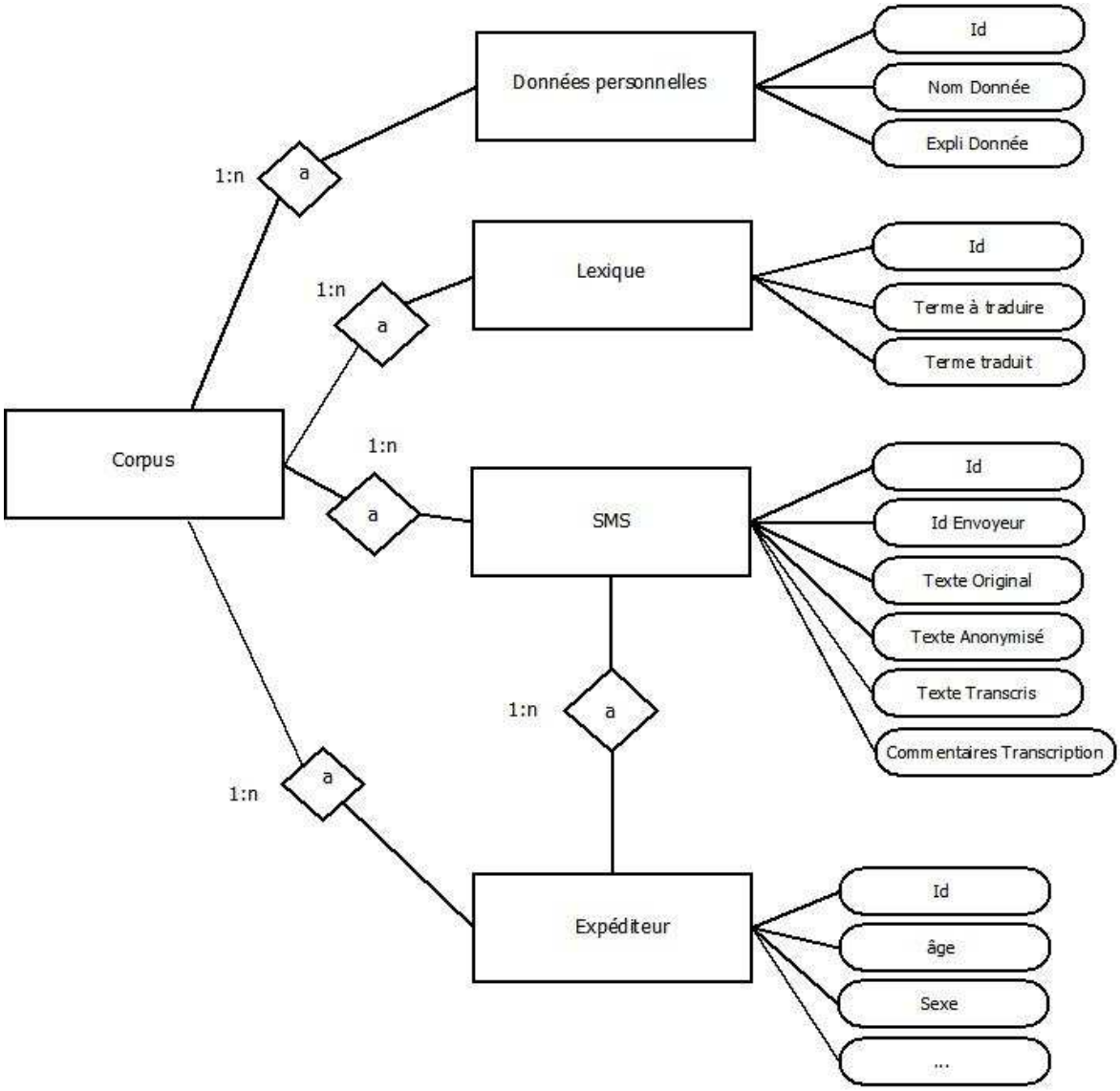


Schéma de la base de données

Extraction des SMS bruts et enregistrement dans la base de données

Les SMS de l'opération « sms4science » vont être transmis vers une boite mail.

L'entête du mail est de la forme « Vous avez reçu un SMS du 000000000000 sur le 31014 ». Les 12 chiffres correspondent au numéro de téléphone codé du participant. Nous l'appellerons « alias ».

Puis dans le corps du mail, nous retrouvons le texte du SMS.

Nous allons nous servir de l'alias comme identifiant du participant dans notre base de données. Il va être utile pour relier les SMS envoyés par un même participant.

Nous allons travailler sur le fichier créé par le client email, qui contient l'ensemble des emails reçus.

Grâce à des expressions régulières, nous récupérons l'entête et le corps de chaque mail. Nous enregistrons ensuite l'alias extrait de l'entête et le texte brut du SMS dans la base de données.

Anonymisation

L'anonymisation est une étape importante à effectuer sur les SMS. En effet, ce projet est enregistré à la CNIL et doit protéger les informations personnelles des participants.

Afin de simplifier le travail d'anonymisation, nous avons pensé qu'il serait intéressant de faire apparaitre certaines données automatiquement. Dans un premier temps, nous avions pensé afficher en rouge les éléments détectés au milieu d'une zone de texte. Mais l'utilisateur aurait dû alors modifier cet élément, avec le risque de modifier par erreur un autre terme du SMS.

Après avoir rappelé le SMS à anonymiser, nous inscrivons à l'écran chaque élément détecté, suivi de deux cases à cocher. Ces cases vont permettre de garder ou masquer l'élément. Si l'utilisateur choisi de le masquer, nous remplaçons l'élément par un code, à savoir « ***(abréviation du type de la donnée)*** ». Nous noterons que cette méthode ne fera pas la différence entre deux noms de famille au sein d'un même SMS. C'est-à-dire qu'ils seront remplacés par la même abréviation, pouvant créer une confusion pour la compréhension du SMS. Nous penserons aux améliorations possibles par la suite.

Dans le cas où l'utilisateur indique une donnée personnelle non détectée, nous lui proposons de choisir son type dans une liste déroulante. Nous ne voulions pas que la page se recharge, c'est pour cela que nous avons utilisé l'AJAX. Cette méthode permet de faire une requête au serveur et d'en retirer un résultat sans que l'utilisateur sans aperçoive. Dans notre cas, nous récoltons les types de données personnelles présentes dans notre base de données pour créer une liste déroulante. Cela permet de mettre à jour la liste des types de données pour prendre en compte les ajouts des utilisateurs.

La liste déroulante ne contient que des abréviations des types de données. C'est lorsque l'utilisateur choisit un type de donnée dans la liste déroulante que nous lui affichons l'explication de l'abréviation.

Si l'utilisateur ne trouve pas le type de donnée qui lui convient, il pourra en renseigner un nouveau. Pour cela, il choisira « Autre type de donnée ». Nous lui afficherons alors deux zones de texte afin qu'il précise le type de donnée en forme abrégée et son explication.

Transcription

Concernant la transcription, le principal problème a porté sur le découpage. Au départ, nous avions pensé proposer à l'utilisateur deux méthodes de découpage : l'un automatique et l'autre manuel. Le premier se fonde sur les ponctuations comme séparateurs tandis que pour l'autre, c'est à l'utilisateur de placer le séparateur utile au découpage. Cette dernière méthode parait importante dans le cas où le SMS ne possède pas de ponctuations, donc que le découpage automatique n'est pas possible. Mais parfois, seule une portion du SMS est mal ponctuée et posera ainsi problème.

Nous avons donc décidé de présenter le découpage automatique à l'utilisateur, puis nous lui proposons de le valider ou bien de le modifier. Dans le cas où il y a des modifications à faire, nous affichons chaque élément découpé suivis de deux cases à cocher, « ok » ou « pas ok ». Si le choix de l'utilisateur est le second, alors nous présentons l'élément dans une zone de texte pour faire les modifications.

Nous avons rencontré des problèmes pour reconstituer le SMS après découpage. En utilisant les ponctuations comme séparateurs, ils étaient effacés. Il nous as fallu encadrer les ponctuations par un symbole et se servir de ce symbole comme séparateur. Ainsi les ponctuations étaient conservées.

Conclusion

D’un point de vue linguistique, le langage SMS est intéressant à étudier puisqu’il « joue » avec la langue afin de répondre au besoin de l’utilisateur de réduire le nombre de caractère de son texte.

D’un point de vue sociologique, l’usage des SMS est complexe. Les SMS sont utilisés par une population très large mais pas de façon différente.

Il parait intéressant de réaliser un projet tel que « sms4science ». En effet, par plusieurs points les corpus recueillis dans le cadre du projet « sms4science » diffèrent des corpus observés précédemment. Ils sont diversifiés, c'est-à-dire que les SMS proviennent d’horizons différentes, contrairement au corpus de Caroline Tagg par exemple, où les SMS ont été récoltés dans son contexte familial et amical.

Les SMS ne subiront pas d’erreurs de saisie puisque les participants au projet vont nous transmettre leurs SMS directement de leur boite d’envoi de leur mobile vers notre numéro court. Ils n’auront pas à retranscrire leurs SMS sur un ordinateur par exemple, comme on été amené à le faire les participants à la récolte de SMS menée à Singapour.

Les corpus constitués dans le cadre du projet « sms4science » seront multilingues puisque le projet se veut international. Il a déjà été mené en Belgique, en Suisse, au Canada et sur l’île de La Réunion. Il sera bientôt réalisé en France (début de la collecte en octobre 2010). Puis d’autres pays viendront tels que l’Espagne, la Grèce, Le Royaume-Uni et le Pays Basque.

Lors de la récolte des SMS, les participants sont invités à remplir un questionnaire sur le site Internet du projet. Ainsi, nous pourrons lier les informations des participants à leurs SMS afin de faire de nouvelles observations concernant les SMS et l’âge, le sexe, la langue maternelle, etc.

Nous souhaitons reprendre, pour conclure, la question posée par le titre de ce mémoire. Quel intérêt représente le TAL pour les SMS et les SMS pour le TAL ?

Nous pensons que le TAL a pu apporter des outils existants appliqués aux SMS. Par exemple, les correcteurs orthographiques de SMS se fondent sur les mêmes mécanismes pour corriger les abréviations que des fautes d'orthographe.

Un corpus de SMS comme ceux des projets « sms4science » vont apporter aux applications TAL un lexique, des nouvelles formes d'abréviations qui vont pouvoir être utiles, pour la traduction par exemple. Nous pouvons penser qu'ajouter un lexique SMS/français standard aux moteurs de recherche améliorerait certaines recherches, si la recherche par mots-clés tient compte de la « version » abrégée de ces mots-clés.

Bibliographie

ANIS, J. (2001). (Dir.), (2001), « *Parlez-vous texto ?*», Paris : Le Cherche Midi Éditeur.

ANIS, J. (2002). « *Communication électronique scripturale et formes langagières : chats et sms* ». In Actes des Quatrièmes Rencontres Réseaux Humains / Réseaux technologiques, Université de Poitiers, Poitiers.

BOVE, R. (2005). « *Etude de quelques problèmes de phonétisation dans un système de synthèse de la parole à partir de sms* ». In *Actes de RECITAL 2005,* Dourdan.

DEJOND A. (2002). « *La cyberl@ngue française »*, Tournai, La renaissance du livre.

DEJOND A. (2010). « *Ki a peur du cyberl@ngage?* ». In *Correspondance*, mars 2010.

DAUGMAUDYTË, J., KËDIKAITË, D. (2006). « *Le langage SMS dans le français* ». In *Kalbotyra*, 56(3), pp. 39-47.

FAIRON, C., KLEIN, J.R., PAUMIER, S. (2006). *« Faites don de vos SMS à la science ». Un corpus pour l'étude du langage SMS*. Coll. Cahiers du CENTAL, Presses universitaires de Louvain.

FAIRON, C., KLEIN, J.R., PAUMIER, S. (2006). « *Le langage sms : révélateur d'1compétence* ». In *Le français m'a tuer*, 1, pp. 33–42.

GAGLIO, G. (2004). « *La pratique du sms : analyse d'un comportement de communication en tant que phénomène social* ». In *Consommations et sociétés*, 4.

GUIMIER DE NEEF, E. and VERONIS, J. (2004). « *1 pw1 sr la kestion ;-)*». Paper presented at the Journée d'Etude de l'ATALA, *Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.)*, Paris.

GUIMIER DE NEEF, E., DEBEURME, A., Park J. (2007). « *TiLT correcteur de SMS : évaluation et bilan quantitatif* », TALN 2007, Toulouse, p. 123-132, 2007.

GUIMIER DE NEEF, E., FESSARD, S. (2007). « *Evaluation d'un système de transcription de SMS* ». Acte du 26e Colloque international Lexique Grammaire, Bonifacio, France, 2-6 octobre 2007.

HOCQ, S., (2006). *« Etude des SMS en francais : constitution et exploitation d'un corpus aligné SMS – langue standard ». Rapport de Master II "Industries des Langues"*, Aix-en-Provence.

JAFFRE, J.-P. (2003). *« L'écriture et les nouvelles technologies : ce que les unes nous apprennent de l'autre »*. Actes des Quatrièmes Rencontres Réseaux Humains / Réseaux Technologiques. Poitiers, 31 mai et 1er juin 2002.

KOBUS, C., YVON, F. and DAMNATI, G. (2008). *« Transcrire les SMS comme on reconnaît la parole* ». In *Actes de la Conférence sur le traitement Automatique des Langues* (TALN'08), Pages 128-138, Avignon, France.

MARTIN, C. (2007). « *Téléphone portable et relation amoureuse : les SMS, des messages vraiment désincarnés ?* », Corps, 3, pp. 105-110.

MARSO, P. (2005). « *CP SMS* », Editions Megacom-ik.

PANCKHURST, R. (2009), « *Short Message Service (SMS) : typologie et problématiques futures* », in Arnavielle T.(coord.),*Polyphonies, pour Michelle Lanvin*, Université Paul-Valéry Montpellier3, p.33-52.

RIVIERE, C.-A (2002). *« La pratique du mini-message. Une double stratégie d'extériorisation et de retrait de l'intimité dans les interactions quotidiennes ».* In *Réseaux*, 112-113 : 139–168.

TAGG, C. (2009). *« A corpus linguistics study of SMS text messaging». Thèse soutenue en mars 2009.*

VARNHAGEN, C. (2009), *« lol : new language and spelling in instant messaging ».* In *Reading and writing.*

VERONIS, J. et GUIMIER DE NEEF, E. (2006), *« Le traitement des nouvelles formes de communication écrite »*, In Gérard Sabah (Éd.), Compréhension des langues et interaction, Lavoisier, Paris : 227-247.

VIENNEY, S., MELIAN, C., (2004), « *La correction automatique du langage des nouvelles formes de communication écrite* » In BULAG N°29, 2004

YVON, F., (2008), « *Une architecture pour le traitement automatique des SMS* », présentation de la conférence du 21 novembre 2008 à l'Université Stendhal.

Table des matières

MOTS-CLÉS : SMS, langage, corpus, traitement automatique, transcription, anonymisation.

RÉSUMÉ

Ce mémoire présente le travail réalisé en préparation d'une collecte de SMS en France métropolitaine : alpes4science. Ce projet vise à constituer un corpus afin de proposer des données nombreuses et diverses comme outils d'études aux chercheurs travaillant sur le SMS, les pratiques qui lui sont associées et le langage SMS qui peut découler de ce mode de communication.
Le SMS peut apparaitre comme un mode de communication écrit mais aussi oral, par son immédiateté et le langage utilisé.
Dans un premier temps, nous proposons un état de l'art concernant le SMS et ses pratiques. Puis nous présentons les applications élaborées à partir de et pour les SMS, ainsi que les corpus réalisés jusqu'alors. Dans une dernière partie, nous expliquons ce qui a été mis en place pour constituer et exploiter le corpus alpes4science.