Quelques contributions des statistiques à l'analyse sociolinguistique d'un corpus de SMS

Louise-Amélie Cougnon [1], Thomas François [2]

[1] CENTAL – Université catholique de Louvain – Louvain-la-Neuve (Belgique)

[2] Aspirant FNRS, CENTAL – Université catholique de Louvain – Louvain-la-Neuve (Belgique)

Résumé

Depuis son invention, le SMS a inspiré de nombreuses recherches qualitatives tant dans les milieux scientifiques qu'industriels. Des linguistes, mais aussi des sociologues, des anthropologues et même des compagnies téléphoniques ont cherché à tirer des conclusions à propos de la pratique du SMS ou des comportements que ce média entraîne. Toutefois, en l'absence de processus inférentiel, il n'est pas possible de généraliser les conclusions de ces études à l'ensemble de la population. Cherchant à pallier cette limitation, nous examinons, à partir d'un corpus francophone de 30.000 SMS, si le processus de collecte de cette ressource a permis de remplir les conditions méthodologiques d'un tel processus d'inférence. Dans une seconde partie consacrée au respect de la norme dans les SMS, nous montrons que, malgré les limitations que l'on peut observer dans un tel corpus, les méthodes quantitatives peuvent appuyer et enrichir une description qualitative des pratiques SMS.

Abstract

A wide variety of qualitative researches already focused on SMS corpora: linguists, sociologists, anthropologists or even phone companies tried to define, describe and draw conclusions about SMS practice, language or related behaviours. But the scope of their conclusions is usually limited to their sample, since no inferential process is involved in the quantitative analysis. Based on a French-speaking corpus of 30.000 SMS, our approach will first examine how the design of this resource has met or not the methodological conditions implied by such an inferential process. In a second part dedicated to the respect of the norm in SMS context, we will consider, despite the limits that can be admitted about a SMS corpus, contributions of quantitative studies for qualitative research approval or improvement.

Keywords: SMS, CMC, sociolinguistics, representativity, norm

1. Introduction

Alors que les nouvelles technologies investissent chaque jour davantage notre quotidien, fournissant de nouveaux médias de communication écrite, tels le SMS, le courriel, la messagerie instantanée ou les forums de discussion, les pratiques linguistiques en vigueur au sein de ces nouveaux canaux restent assez méconnues. Certes, il existe diverses études sur le sujet [1], mais celles-ci ne se basent généralement que sur un échantillon limité de messages – pas toujours

[1] On citera par exemple les corpus anglais *SMS Msg Corpus* (http://www.netting-it.com/) et *Linguistic Features of Mobile Phone Communication* de l'université de Hong Kong, le corpus italien de l'université de Turin (http://www.e-allora.net/SMS/ms_index.php?var=SMS) et le corpus de SMS de langue allemande (http://www.mediensprache.net/archiv/corpora/sms_os_h.pdf).

authentiques [2]. Par conséquent, leurs conclusions ne sont valables que pour cet échantillon, qui, de plus, provient souvent d'une population mal définie.

C'est pourquoi, dans cet article, nous avons étudié la possibilité d'employer des méthodes statistiques pour rendre l'analyse sociolinguistique de ce type de corpus capable de décrire des phénomènes au niveau de la population. Pour ce faire, nous avons utilisé le corpus constitué en 2004, sous l'égide de l'Université catholique de Louvain [3] (Fairon et al., 2006b). Il comprend 30.000 SMS [4], envoyés par des usagers francophones de Belgique, qui ont ensuite été anonymisés et retranscrits en graphie standard. Parmi ces participants, la majorité d'entre eux (2.436) ont accepté de répondre à un questionnaire sociolinguistique, ce qui a permis de dresser des profils pour chaque destinateur de messages.

Sur la base de ces profils, notre recherche envisagera, dans un premier temps, la question de la représentativité de ce corpus et les raisons qui rendent la modélisation de la population cible complexe. Dans une seconde partie, nous mettrons en évidence l'intérêt d'effectuer des analyses statistiques du corpus, au travers d'une thématique particulière : le rapport à la norme écrite [5]. Nous nous pencherons sur trois phénomènes courants que sont l'abréviation, les salutations et l'emprunt dans le but de préciser comment les usagers du SMS se situent vis-à-vis de la norme écrite. Nous affinerons nos résultats en détaillant comment ces pratiques linguistiques sont influencées par certaines caractéristiques des scripteurs, à savoir leur sexe, leur âge, leur niveau de formation et leur fréquence d'utilisation du SMS.

2. Constitution de corpus et inférence

2.1. Échantillon et représentativité

Constituer un corpus de SMS authentiques reste actuellement une tâche délicate. En effet, pour des raisons de respect de la vie privée, les opérateurs SMS ne sont légalement pas autorisés à transmettre des copies de leur base de données. C'est la raison pour laquelle un certain nombre de chercheurs ont constitué des corpus, qui restent toutefois de taille réduite, peu authentiques et peu fournis en données sociolinguistiques.

L'option prise par le Cental, lorsqu'il a rassemblé, en 2004, un premier corpus de SMS à l'échelle de la Belgique francophone [6], cherchait à parer les problèmes évoqués précédemment. Ainsi, la collecte ne dépendait-elle pas des opérateurs de téléphonie et mettait à contribution de très nombreux sous-groupes de la population [7]. Même si cette méthode comporte des aspects

[2] Il est par exemple arrivé que des SMS collectés n'aient pas été envoyés lors de véritables échanges conversationnels, mais qu'ils aient été inventés de toute pièce dans le but de la collecte.

[3] C'est en particulier le Cental, centre de recherche en traitement automatique du langage, qui s'est chargé de la collecte (www.sms4science.org).

[4] Notons que notre étude en particulier ne se base que sur 24.871 messages (ceux pour lesquels nous disposions également d'un profil sociolinguistique).

[5] L'écrit, à l'inverse de l'oral, est codifié depuis longtemps et son fonctionnement est régi par des grammaires et des dictionnaires prescriptifs. En ce sens, lorsque nous nous référons à la norme écrite, nous renvoyons aux normes morphologiques et syntaxiques prescrites dans ces ouvrages de référence.

[6] Entendue ici, et dans le reste de cet article, comme la Région Wallonne et la Région Bruxelloise. Il s'agit bien sûr d'une approximation, puisque ce découpage ne tient compte ni des 72.000 germanophones résidant en Wallonie, ni des quelques 20% de flamands habitant Bruxelles.

[7] Les usagers étaient conviés à envoyer des copies de SMS qu'ils avaient au préalable envoyé à un véritable destinataire ; pour les encourager, les responsables du projet ont eu recours à des techniques aussi diverses que

très positifs et pallie effectivement les problèmes d'authenticité ou de taille du corpus évoqués précédemment, elle reste insatisfaisante d'un point de vue purement statistique, puisqu'elle opère un échantillonnage par volontaires, lequel ne produit généralement pas un échantillon représentatif de l'ensemble des caractéristiques de la population.

2.2. Évaluation de la représentativité du corpus

Afin de vérifier la représentativité d'un jeu de données, il est possible d'effectuer une série de tests d'ajustement (test khi-carré pour une distribution non paramétrique ou test de normalité pour des données normales) pour chacune des dimensions prises en compte (âge, sexe, etc.). Toutefois, dans le cas d'un corpus de SMS, cette approche est rendue peu fiable, la population étant délicate à modéliser : si cette population est relativement simple à circonscrire – l'ensemble des utilisateurs du SMS en Belgique francophone – elle est par contre presque impossible à appréhender en termes numériques, ne fût-ce qu'en raison de la possibilité qu'un utilisateur recoure à plusieurs GSMs ou de la mixité linguistique propre au contexte belge.

Confrontés à ces difficultés méthodologiques, nous avons toutefois effectué une série de tests khi-carré d'ajustement (Agresti, 2002 : 22) en ce qui concerne les variables âge et sexe des utilisateurs. Nous avons dû émettre des hypothèses simplificatrices afin d'obtenir une population dont les paramètres sont connus. Par conséquent, comme nous l'expliquons ci-dessous, les résultats de ces tests d'ajustement peuvent être interprétés de deux manières.

En ce qui concerne le sexe des utilisateurs, on observe une plus grande proportion de femmes (57,2%) que d'hommes (42,7%). En comparant avec le nombre de résidants en Belgique francophone [8], où la proportion de femmes est de 51,6% contre 48,4% pour les hommes, il n'est pas difficile de prévoir que le résultat du test d'ajustement ($\chi^2(1) = 21{,}2$; $p < .0001$) indique que cette population ne peut correspondre à celle d'où provient l'échantillon. Ce résultat semble indiquer un biais, à moins que, pour une raison qui reste à éclaircir, les femmes aient réellement davantage tendance à communiquer via SMS que les hommes.

Au niveau de l'âge, nous avons regroupé les utilisateurs en 6 classes d'âge. Tab. 1 précise leurs frontières, leur proportion dans le corpus et dans la population [8]

Classes	*Proportion : corpus*	*Proportion : population*	*Classes*	*Proportion : corpus*	*Proportion : population*
1 : - de 15 ans	10,90%	18,20%	4 : 25 – 34 ans	15,40%	13,90%
2 : 15 – 19 ans	29,80%	6,10%	5 : 35 – 44 ans	6,20%	14,90%
3 : 20 – 24 ans	32,50%	6,30%	6 : + de 45 ans	5,10%	40,50%

Tableau 1 : Répartition des utilisateurs selon leur classe d'âge, dans le corpus et dans la population résidante

Tab. 1 montre qu'une large majorité des participants à l'étude sont des jeunes âgés de 15 à 24 ans. À nouveau, le test khi-carré d'ajustement est extrêmement significatif ($\chi^2(5) = 3941{,}5$; $p < .0001$), ce qui est révélateur soit d'un biais dans le corpus, soit d'une utilisation supérieure de la fonction SMS chez les jeunes. L'hypothèse la plus probable est ici que ces deux causes

la rédaction d'articles dans des quotidiens ou encore la mise en jeu de « prix » remis aléatoirement.

[8] Voir le site de Statbel, le principal acteur de la statistique officielle en Belgique: http://statbel.fgov.be/fr/statistiques/chiffres/population/structure/agesexe/.

jouent ensemble : le SMS constitue certainement une technologie séduisant davantage un public jeune [9], et les lots de stimulation utilisés pour la collecte du corpus ont très probablement influencé le type de population ayant participé à l'étude.

2.3. La présence de biais doit-elle conduire à rejeter toute approche statistique ?

Cette première partie sur la représentativité du corpus ouvre davantage de pistes de recherches qu'elle n'apporte de réponses. Nous avons montré que notre corpus se différencie nettement de la population résidant en Belgique francophone, sans pouvoir déterminer si cet écart s'explique uniquement par des biais liés à la méthode d'échantillonnage par volontaires ou s'il recouvre des caractéristiques réelles de la population d'utilisateurs des SMS. Pour intéressante que soit cette question, il conviendrait, pour la trancher, d'effectuer une nouvelle collecte de SMS qui respecte les principes d'un échantillonnage aléatoire simple. Or, cette option reste actuellement impossible, pour les raisons légales que nous avons expliquées précédemment. De plus, si dans ce genre de situation, le ré-échantillonnage des données constitue une manière de réduire les biais liés à la méthode de collecte, il faut toutefois, pour appliquer cette technique, être capable de définir la population cible. Or, dans le cas précis des corpus SMS, cette information n'est pas disponible non plus. Par conséquent, ces diverses difficultés méthodologiques limitent fortement toute tentative d'inférence statistique. Mais faut-il pour autant rejeter l'outil quantitatif ?

Dans la seconde partie de cet article, nous avons pris le parti de montrer que les méthodes quantitatives appliquées à des corpus de SMS, si elles n'autorisent pas à décrire des phénomènes sociolinguistiques au niveau de la population sans risque de biais, permettent toutefois d'explorer un corpus d'une manière systématique. Dans ce contexte, l'utilisation de tests statistiques vise à révéler certains phénomènes qui auraient échappé à une approche qualitative, mais aussi et surtout à ne pas surestimer leur importance sur la base de quelques exemples. Nous avons choisi d'en faire la démonstration en analysant la question du respect de la norme dans les SMS en fonction des caractéristiques sociodémographiques des auteurs de SMS.

3. La norme [10] écrite remise en question

L'écrit spontané qu'offre le contexte des SMS est associé, en comparaison avec d'autres types d'écrit, à un relâchement des normes, qu'elles soient typographiques, linguistiques ou communicationnelles. En effet, il semblerait que le contexte de ce type de « communication médiée par ordinateur » (Panckhurst, 1997), ou CMO, tend à désinhiber le poids normatif qui pèse traditionnellement sur les pratiques linguistiques à l'écrit : ainsi observe-t-on une tendance nette à l'abréviation, à l'usage de régionalismes, de néologismes, à l'alternance de codes ou encore au recours à l'argot.

Un tel effet désinhibiteur du médium s'observe également dans ce que l'on nomme couramment le parlé ordinaire (Gadet, 1989). En effet, le langage SMS partage avec le parlé ordinaire une certaine spontanéité, un écart par rapport aux conditions de communication écrite encadrées. Toutefois, certaines caractéristiques du SMS engendrent des pratiques qui rendent ce langage

[9] Nous en voulons comme preuve les données rapportées par le site http://www.smsmessenger.be/site/statistiques_sms.asp, estimant qu'en 2008, la proportion des 15-34 ans qui utilisent la fonction SMS de leur téléphone s'élève à 90% alors que le total pour la population est de 70%.

[10] La norme ici entendue comme « un système d'instructions définissant ce qui doit être choisi parmi les usages d'une langue donnée si l'on veut se conformer à un certain idéal esthétique ou socioculturel » (Dubois et al., 2007 : 330).

singulier, distinct du parlé ordinaire et propice à des études spécialisées (Cougnon and Ledegen, 2009). Ainsi, lorsque l'on parle de l'usage du SMS, on fait souvent référence aux contraintes de temps, d'espace et de coût qui régissent cet usage. Ces trois contraintes encouragent l'utilisation de formes abrégées autant du point de vue graphique que syntaxique, entrant ainsi en conflit avec la contrainte normative. Enfin, le manque de moyens de communication paralinguistiques, allié à l'aspect ludique propre à la CMO, encouragent les usagers du SMS à recourir à des systèmes de ponctuation émotive et à des phénomènes graphiques, tels que les smileys, pour transmettre une certaine forme d'expressivité ou encore pour indiquer le ton du message.

Le relâchement de l'écrit dans les SMS [11] a engendré, aussi bien dans les cercles spécialisés qu'auprès du grand public, une gamme de questionnements et même d'indignations (Jalabert, 2006) au sujet de l'avenir de la langue. D'autres vagues de réactions ont mis le doigt sur l'aspect positif de cette pratique qui montre la vitalité et l'adaptabilité de la langue (Fairon, Klein et Paumier, 2006a) et qui offre un regain d'intérêt pour l'écrit, notamment chez les jeunes [12].

Dans cet article, nous avons justement voulu aborder cette problématique du respect de la norme écrite sous un angle sociolinguistique en profitant des outils et du recul qu'offre l'analyse quantitative. Ainsi nous traiterons de cette question au travers de trois phénomènes particuliers : le taux d'abréviation graphique, les salutations à l'ouverture des messages et les emprunts aux langues étrangères. En outre, nous tenterons de vérifier si l'effet désinhibiteur décrit précédemment est attesté chez l'ensemble des utilisateurs, quelles que soient leurs caractéristiques sociodémographiques.

3.1. La bourse ou la norme

Nous venons de voir que le contexte du SMS imposait des contraintes de temps, de coût et d'espace. En effet, un SMS est limité à 160 caractères, bien que les GSM récents permettent de dépasser cette limite en concaténant plusieurs messages à la suite, ce qui engendre toutefois un coût supplémentaire en termes de temps et d'argent. C'est probablement pourquoi, dans le corpus du Cental, 86,3% des messages font moins de 160 caractères et 9,7% en comptent exactement 160. On considère généralement que ces contraintes de production encouragent les utilisateurs de SMS à en écrire de plus restreints, en recourant à divers moyens pour abréger leur message [13], à l'instar de ce SMS : « Maman kèsk L sé kèsk L sépa ? Eseydepa c àlamésonce WE.PAPA ». Face à de tels exemples, on peut comprendre les inquiétudes de certains pour l'avenir de la langue et le respect de la norme. Toutefois, au regard du corpus entier, nous pouvons affirmer que ce genre de production est rare.

Afin d'analyser plus précisément cette problématique de l'abréviation, nous avons compté la différence de caractères entre les messages et leur traduction, obtenant une réduction moyenne

[11] Caractérisé par ce que Koch and Oesterreicher (1985) appellent la *Nähesprache* (ou « langue de proximité »), par opposition à la *Distanzsprache*.

[12] Le Conseil supérieur de la langue française au Québec, par exemple, et son président Conrad Ouellon, tentent de dédramatiser la vision plus pessimiste qu'objective des nouveaux modes de l'écrit (voir par exemple l'article http://www.cyberpresse.ca/vivre/200809/08/01-652751-le-texto-nest-pas-une-menace-pour-le-francais.php).

[13] Notons que ce phénomène d'abréviation constitue une caractéristique essentielle mais non unique de l'écrit SMS. En effet, comme explicité dans Cougnon and Ledegen (2009), d'autres propriétés du contexte SMS, telles que la volonté de transmettre une certaine forme d'expressivité, l'absence d'autorité normative, la volonté d'appartenance à un groupe socioculturel, etc., jouent sur l'écrit SMS, allant même jusqu'à inhiber les habitudes d'abréviation des usagers.

de 13,4 caractères sur une moyenne de 105 caractères par message [14]. En normalisant pour chaque message ce nombre de caractères omis en fonction du nombre de caractères de la traduction, nous observons que le taux de réduction moyen des messages atteint seulement les 9,4%. On est donc loin des 45% de réduction du message proposé plus haut en exemple, même si ce taux est significativement différent de 0, comme le révèlent les résultats d'un test T de comparaison de moyennes ($t(24828) = 151,7$; $p < 0,0001$).

Ajoutons que dans l'ensemble du corpus, 19% des messages n'ont subi aucune réduction de taille et que 2,4% des messages sont même plus longs que leur transcription (ex. : aaaaaahhhh !). Dès lors, il nous a semblé intéressant d'étudier divers facteurs explicatifs de la variabilité de ce phénomène de réduction. Pour ce faire, nous avons analysé l'effet de l'âge, du sexe, du nombre de SMS envoyés par semaine [15] (*nbsmssem*) et du niveau d'étude [16] (*nivetud*) sur la proportion de caractères omis dans les SMS (*PercentDiff*) à l'aide de tests non paramétriques. En effet, aucune des distributions conditionnelles de *PercentDiff* en fonction des quatre variables explicatives ne suit une distribution normale, ce qui s'explique notamment par la présence d'un nombre anormalement élevé de messages de 160 caractères.

Au niveau du sexe, un test de Mann-Whitney [17] indique que le sexe influe sur le degré de concision dans les SMS ($U = 6,55 * 10^7$; $p < 0,0001$). Les messages écrits par des femmes sont davantage abréviés (10% de réduction contre 8,4% pour ceux des hommes), peut être aussi parce que les femmes sont plus loquaces (109 caractères par message en moyenne contre 99 pour les hommes) [18]. Pour les trois autres variables, une série de tests de Kruskal-Wallis [19] indiquent qu'aussi bien l'âge ($X^2(5) = 2842$; $p < 0,0001$), le niveau d'étude ($X^2(4) = 1206$; $p < 0,0001$) que le nombre de SMS envoyés par semaine ($X^2(5) = 210$; $p < 0,0001$) sont significativement associés aux pratiques d'abréviations dans les SMS.

Toutefois, les mesures de corrélation révèlent que ce sont surtout l'âge ($r_s = -0,335$) et le niveau d'études ($r_s = -0,148$) qui possède une capacité explicative, alors que l'effet du sexe est plus marginal ($r_{pb} = -0,085$) et celui du nombre de sms envoyés est même non significatif ($r_s = 0,006$; $p = 0,31$).

La question qui se pose dès lors est de détailler cet effet de l'âge et du niveau d'éducation. En observant les moyennes, on peut remarquer que plus le scripteur du SMS est âgé, moins de caractères sont omis et donc, à priori, plus l'utilisateur respecte les normes orthographiques et syntaxiques du français. On passe ainsi d'une proportion d'abréviation de 15,4% pour les messages des moins de 15 ans à un ratio de 4,9 % pour ceux des 45 ans et plus. Le tableau est un peu moins clair pour le niveau d'études où les messages écrits par des diplômés du primaire et

[14] Notons que cette méthode a le mérite de livrer une idée globale du degré d'abréviation mais ne précise pas si l'abréviation porte sur les unités lexicales ou sur la ponctuation et les espaces.

[15] Cette dimension a été paramétrée lors de la collecte sous la forme d'une variable ordinale à 6 niveaux. Ceux-ci représentent des classes de nombre de SMS envoyés par semaine et correspondent respectivement à : 1 → – de 5 ; 2 → de 5 à 10 ; 3 → de 10 à 20 ; 4 → de 20 à 50 ; 5 → de 50 à 100 ; 6 → + de 100.

[16] Cette variable compte 9 niveaux dans le corpus, mais nous l'avons réduite à 5 classes pour pallier l'insuffisance des données dans certaines classes. Celles-ci correspondent aux diplômes suivants : 1 → primaire et secondaire inférieur ; 2 → secondaire supérieur : général ; 3 → secondaire supérieur : technique et professionnel ; 4 → supérieur non universitaire ; 5 → supérieur universitaire.

[17] Décrit en détail par Howell (2008 : 675), ce test constitue l'équivalent non paramétrique d'un test T de comparaison de 2 moyennes.

[18] Cette tendance a déjà été remarquée dans la CMO en général. On retrouve par exemple dans Piette et al. (2007, 88) : « les garçons vont plus à l'essentiel, […] alors que les filles vont plus bavarder ».

[19] Ce test correspond quant à lui à une comparaison omnibus de N moyennes, c'est-à-dire une version non paramétrique de l'ANOVA. Il est décrit dans Howell (2008 : 680).

de l'inférieur montrent une tendance plus forte à l'abréviation (13,9%), suivis par les diplômés du secondaire supérieur général (11%) et technique et professionnel (9,5%), puis par des universitaires (8,5%) et enfin des diplômés du supérieur non universitaire (7%). L'éducation est donc bien en relation avec le phénomène d'abrègement lexical et syntaxique (comme le confirme la corrélation négative), mais il ne s'agit pas d'un phénomène aussi évident que pour l'âge.

3.2. Des salutations variées et orientées

Le deuxième axe de recherche s'est focalisé sur les termes de salutations occupant une fonction d'ouverture du canal de communication et constituant donc le premier mot d'un message. Comme le rappelle Traverso (1996 : 67), pour la société occidentale au moins, « les salutations relèvent de la politesse positive, et plus précisément […], des " rites de présentation ". Elles […] sont obligatoires ».

La première étape de notre analyse des salutations a consisté à paramétrer le corpus en repérant les messages comprenant des salutations [20]. Cette opération a été réalisée à l'aide d'une méthode semi-automatique d'extraction basée sur une liste initiale de salutations enrichie de manière incrémentale (voir Tab. 2 pour la liste finale, qui a été légèrement modifiée afin de rassembler des patterns similaires tels hep, heps, hép, etc.). Cette liste comprend, d'une part, les salutations normalisées qui apparaissent dans la colonne traduction du corpus SMS et, d'autre part, les différentes variantes employées dans le corpus. Nous avons ainsi repéré 5 277 occurrences de salutations (Tab. 2).

Traverso (1996 : 69) a noté que lors de conversations orales « la formulation la plus courante pour la salutation est " bonjour ". […] elle est souvent concurrencée par " salut ". Cette formule est plus familière […] Des salutations particulières sont aussi attestées mais rarissimes [tel que] " hello " ». Nous voyons que nos résultats ne suivent pas cette tendance, ce qui tend à renforcer une spécificité du langage SMS par rapport à ce que nous avons appelé précédemment le parlé ordinaire. Non seulement les formes *salut* et *bonjour* ne sont pas en concurrence puisque la première forme est prédominante, mais la forme *hello* n'est pas marginale puisque sa fréquence est même supérieure à *bonjour*. Selon Gadet (2007 : 29), « la norme, parfois dite " de référence ", a pour effet de renforcer la cohésion sociale ». D'après les observations faites à propos des salutations dans les SMS, nous aurions plutôt tendance à croire l'inverse : en effet, l'analyse qualitative autant que l'approche quantitative montrent que certains sous-groupes affectionnent plus particulièrement certaines variétés de salutation dont la fonction semble être de renforcer une cohésion identitaire dans le sous-groupe en question. En effet, pour l'ensemble des SMS du corpus, on observe une association significative ($X^2(35) = 639,6$; $p < .0001$) entre l'âge du scripteur et la forme de salutations employée [21]. De plus, l'analyse des « résidus standardisés » (cf. Sheskin, 2004 : 525) révèle que quelques formes sont significativement privilégiées par certaines classes d'âge et délaissées par d'autres.

---

[20] Les salutations apparaissent à l'ouverture de 21% des messages du corpus (5.278 SMS). Cette fréquence peu élevée n'est pas significative en soi puisqu'il est particulièrement difficile de juger des messages qui ne contiennent pas de salutations étant donné que le corpus du Cental ne présente pas des dialogues mais des messages unidirectionnels ; ainsi, il est impossible de distinguer, à l'aide d'outils quantitatifs, les messages qui initient réellement la conversation de ceux qui ne constituent qu'une réponse sans salutation ouvrante. Une étude qualitative serait à même d'affiner ces résultats.

[21] On notera toutefois qu'étant donné le grand nombre d'observations testées, il n'est pas difficile de trouver un effet entre deux variables. Dans ce cas de figure, il est recommandé d'accompagner le test khi carré d'une mesure de la taille de l'effet. Dans le cas de variables catégorielles à plusieurs niveaux, on utilise généralement le V de Cramer , qui, pour cet exemple, vaut 0,156 ($p < .0001$).

Salutation normalisée	*Nb d'occurrences*	*Variantes (variations de casse possible)*
Salut	1657	salut, slt, salu, sl, chalu, plu, saloute, lut, saaaalu, sal, alu, saut, saaluu, lu, st, 'lu, sit, sakut
Coucou	1661	ccou, coucou, koukou, coucov, cc, couc, kkou, cou, concou, coucot, cicou, koucou, kcou, cocou, kouk, kikou, coucou, kuku
Kikou	235	kikou, kikoo, kikoooo, kik0o, kiko, likou, kkou
Bonjour	566	bjr, bonjour, bjour, bonne nuit, bijour, bichour, bonj, bj, bondjou, jour, ' jour, 'jour, boonjour, bjou, bilou, bnj, b'jour
Hello	667	elo, hello, helo, lo, heyo, hlo, l.o, llo, hilo, l=o, hell0, yello, ll, hèlo
Salam	10	salam
Hi	33	hi
Hey + Héy + Hé	169 + 2 + 42	hey, hei, Eyh, Ey, héy, ey, heyo + héy, hèy + hè, hé, he
Ciao	9	ciao
Hola	18	Hola, ola
Hallo	3	hallo
Yep	40	yep, yèp
Yo	73	yo, yop, youp, you you
Hep	87	hep, hép, ep, hèps, heps, hép

Tableau 2 : Statistiques pour les différentes formes et variantes de salutation

Ainsi, on note une nette préférence des plus de 25 ans pour la forme *bonjour*, tandis que les moins de 25 ans la dédaignent au profit de formes plus familières [22]. Parmi celles-ci, *salut* est surtout caractéristique des moins de 15 ans (z = 7,4), et des 15-19 ans (z = 2,5), alors qu'elle est sous-employée par les scripteurs plus âgés. *Kikou* et *hey* se révèlent très intéressantes, car spécifiques aux 15-19 ans (respectivement z = 5,8 et z = 6,3). *Coucou* représente une forme typique à la fois des 20-24 ans (z = 6,3) et des plus de 45 ans, et est significativement sous-employée par l'ensemble des autres classes (z inférieurs à -3). On note enfin une préférence des 25-34 ans pour les emprunts *hello* et *hi* (z = 4,9).

Ces différences de salutations s'expliquent également en fonction du sexe ($X^2(7) = 215,3$; $p < .0001$), et de manière plus importante encore ($V = 0,202$; $p < .0001$). Les différences d'emploi des formes *bonjour*, *coucou* et *salut* sont en effet significatives. Les hommes montrent un emploi excessif des formes *bonjour* (z = 2,5) et *salut* (z = 6,8) – formes que les femmes sous-emploient (respectivement z = -2 et -5,5) – et ils sous-emploient la forme *coucou* (z = 8,6), significativement privilégiée par les femmes (z = 6,9). Il apparaît donc que certaines formes sont sexuellement marquées, même s'il serait nécessaire d'affiner encore l'analyse par la prise en compte de l'interaction entre le sexe du destinateur et du destinataire. Malheureusement, cette dernière information manque.

Quant à la variable *niveau d'étude*, elle se révèle également significative ($X^2(28) = 259$; $p < 0,001$; $V = 0,11$). On observe un suremploi de *coucou (z* = 4,3) chez les diplômés du supérieur universitaire (*coucou* semble typique du monde universitaire, ce qui rejoint la prédilection des 20-24 ans envers cette forme) et de *hello*/*hi* (z = 4,4), au détriment de *bonjour* et *salut*. *Bonjour* est préférée par les porteurs d'un diplôme du supérieur non universitaire (z = 3,3)

[22] Les valeurs des résidus standardisés (z) sont respectivement de -2,9 (classe 1, cad. les – de 15 ans) ; de -3,7 (classes 2 et 3) ; 5,5 (4) ; **10,7** (5) et 8,8 (6).

ou du secondaire technique/professionnel (z = 6,3), tandis que *salut* jouit surtout de la faveur des diplômés du primaire et du secondaire inférieur (z = 4,3), ainsi que du secondaire supérieur général (z = 2,9). Les autres données concernant cette variable sont par contre beaucoup moins significatives : une même forme peut être sur-employée par des groupes très différents, voire opposés. Notons enfin que la variable *nbsmssem* n'est que peu associée à l'emploi de salutations particulières (V = 0,82 ; p < 0,001).

Ces différentes analyses nous révèlent un rapport particulier à la norme chez les utilisateurs de salutations dans notre corpus SMS. Ceux-ci recourent à deux types d'écart : l'emploi de salutations généralement considérées comme n'appartenant pas au registre courant mais plutôt familier voire argotique (par ex. kikou, hep, hello) et la sélection de variantes d'une forme courante (ex. bijour, salu ou ccou). Toutefois, le premier type d'écart ne représente que 26% des occurrences ; le second varie de 40% (pour salut) à 8% (pour coucou) de formes non standards.

3.3. Les emprunts

Cette dernière partie envisage à nouveau la question du rapport à la norme écrite dans les SMS en analysant le phénomène des emprunts aux langues étrangères. Bien que la méthode de collecte du corpus belge visait à ne rassembler que des messages de scripteurs francophones, on y rencontre des lexèmes appartenant à 16 langues étrangères [23], mélangées au français – la langue matrice [24]. Les formes que peut prendre ce « mélange » sont très variées. Nous trouvons des cas de *code-switching* [25] proprement dit, de *transfert* [26], de *code-mixing* [27] et d'*emprunt*. Toutes ces formes sont confondues dans l'analyse quantitative sous la dénomination d' « emprunt ».

Nous avons détecté dans le corpus du Cental 3.230 occurrences de mots étrangers, représentant 536 formes distinctes. 9,5% des messages du corpus contiennent des emprunts [28]. Les mots empruntés de plus de 100 occurrences proviennent tous de la langue anglaise ; ils servent soit à l'ouverture et à la clôture des messages (*hi*, *kiss*), soit d'équivalents brefs à une forme française trop contraignante à l'encodage (*today*, *now*). La deuxième langue d'emprunt est l'italien (147 tokens) et concerne presque exclusivement le champ lexical des sentiments (*amore*, *bello*, *ti amo*, *baci*, etc.).

À partir des données sociolinguistiques (et en particulier le détail des langues maîtrisées par chaque usager), nous avons également pu distinguer, pour chaque message d'un usager, les mots empruntés à une langue étrangère que l'usager connaît (et qui pourrait expliquer le recours à cette langue) de ceux qui appartiennent à une langue inconnue de l'usager. Or, justement, les

[23] Il s'agit de l'allemand, l'anglais, l'arabe, l'espagnol, le grec, l'italien, le japonais, le latin, le libanais, le macédonien, le marocain, le néerlandais, le portugais, le russe, le tunisien et le turc.

[24] Nous empruntons ce concept à Myers-Scotton (1993). La norme écrite dans ce contexte précis est entendue comme l'emploi unique de la langue matrice.

[25] Il s'agit de séquences de mots de longueur variable provenant de deux langues distinctes, juxtaposes au sein du même échange verbal, qui montrent tous les signes d'un échange monolingue et qui respectent les règles grammaticales des deux langues (Cougnon, 2007). On trouve par exemple : *Attente insupportable mais* ***restons open! Be aware! But he makes me crazy! Yes he does!*** [smsBF].

[26] Il s'agit d'un échange verbal au sein duquel le passage à la deuxième langue survient mais est limité à une construction spécifique suivi d'un retour à la langue matrice.

[27] Il s'agit de l'usage d'une deuxième langue au sein d'un contexte en langue matrice qui ne respecte pas les règles grammaticales de la langue étrangère.

[28] Si 90,5 % des messages ne contiennent aucun emprunt, nous pouvons clairement conclure à une tendance unilingue dans les SMS.

emprunts à des langues connues de l'usager constituent 30% des cas [29]. Comment justifier alors les 70% restant ? On peut évidemment citer les fonctions que revêtent ces emprunts dans les SMS et dont nous avons déjà parlé précédemment (transmettre des émotions, utiliser un mot plus court, etc.). Notons qu'il est également possible de justifier ces emplois par une influence indirecte des facteurs sociologiques, géographiques ou politiques [30].

L'analyse de nos variables a aussi révélé des points intéressants. Les deux variables ayant un effet sensible sur l'usage des emprunts dans les SMS sont l'âge ($X^2(5) = 127,5$; $p < 0,0001$; $V = 0,07$) et le niveau d'étude ($X^2(4) = 127$; $p < 0,0001$; $V = 0,07$), alors que le sexe et le nombre de sms envoyés par semaine ont un effet minime ($\varphi = -0,025$ et $V = 0,03$). L'analyse des résidus standardisés montre, sans surprise, que les emprunts sont plus volontiers utilisés par les plus jeunes [31], avec une charnière autour des 25 ans, âge au-delà duquel on aurait moins tendance à recourir à des emprunts. Quant au niveau d'éducation, ce sont les diplômés du primaire et du secondaire inférieure ($z = 3,8$), ainsi que du secondaire supérieur technique et professionnel ($z = 4,9$) qui préfèrent emprunter à une autre langue, alors que ceux sortis de l'enseignement supérieur non universitaire y semblent peu disposés ($z = -8,6$).

4. Conclusions

Partant d'une volonté de décrire les pratiques du SMS à l'échelle de la Belgique francophone, nous avons étudié dans la première partie la représentativité du corpus SMS du Cental. Cette question n'a pu être résolue puisque la population des usagers du SMS n'est pas connue. Dès lors, il n'est pas possible de déterminer si les différences de constitution observées au niveau du corpus sont le reflet de biais dans sa collecte ou d'une différence entre la population d'utilisateurs de SMS et de la population résidant en Belgique francophone.

Dans la seconde partie, au travers du rapport à la norme écrite envisagé via trois pratiques linguistiques, nous avons montré comment l'outil statistique peut constituer une précieuse méthode exploratoire pour une description sociolinguistique des pratiques du SMS. Le choix de cette thématique n'était pas inconsidéré, puisque cette question se situe au centre des préoccupations actuelles à l'égard des SMS et nombreux sont ceux qui voient dans des phénomènes, tels que l'abrègement syntaxique et morphologique, une caractéristique essentielle des SMS, propre à inquiéter les gardiens de la langue française.

Or, nos résultats ont montré que l'importance de cet écart par rapport à la norme est à relativiser. Ainsi, 20% des messages du corpus ne sont nullement abrégés. De même, la majorité des formes de salutations employées sont les formes standards de *bonjour*, *salut*, *coucou* et *hello* et plus de 90% des messages ne renferment aucun emprunts. Il existe certes certains types de locuteurs qui utilisent en effet le SMS de manière relâchée et qui profitent de ce nouvel espace de communication pour s'éloigner de la norme écrite. Nos analyses révèlent que le facteur le plus explicatif en ce sens est l'âge. On note ainsi, dans les messages des plus jeunes, une

[29] On recense 1.090 tokens provenant de langues connues par l'usager.

[30] La Belgique est en effet un cas exceptionnel en matière de mixité linguistique et culturelle : elle montre trois langues officielles – qui sont en usage dans les pays adjacents – mais aussi un nombre croissant d'immigrants d'Europe et d'Afrique du Nord venus habiter ou travailler en Belgique, attirés par les Institutions européennes et autres compagnies internationales.

[31] Pour le détail des résidus : - de 15 ans : $z = 2,9$; 15 - 19 : $z = 3,9$ et 20 - 24 : $z = 3,2$, tandis que les 25 – 34 : $z = -5,1$; 35 – 44 : $z = -6,9$ et les 45 ans et + : $z = -2,9$.

tendance marquée à l'abrègement [32], à la présence de salutations moins standards telles que *hey* ou *kikou* ou d'emprunts. Par ailleurs, les autres variables n'offrent pas une vision unifiée quant à cette distance vis-à-vis de la norme : elle n'est pas toujours le fait de populations moins diplômées ou plus portées sur les SMS. Au travers de cette analyse, se dessine plutôt l'existence de sous-groupes présentant chacun leurs particularismes.

On peut dès lors penser que diverses attitudes vis-à-vis de la norme écrite coexistent au sein de la pratique du SMS : pour certains, le SMS constitue un lieu de relâchement linguistique ponctuel, pour d'autres, ses particularités n'engendrent pas un écart par rapport à la norme ; enfin, il joue également un rôle au sein de divers groupes sociaux où certains écarts observés se constituent en une nouvelle norme.

Références

Agresti A. (2002). *Categorical Data Analysis. 2nd edition*. New York : Wiley-Interscience.

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

Cougnon L.-A. (2007). 'Trop de langues dans nos assiettes !'. Analyse sociolinguistique de l'alternance français-anglais dans une famille au Grand-Duché de Luxembourg. Master's thesis, Louvain-la-Neuve, Université catholique de Louvain, September 2007.

Cougnon L.-A. and Beaufort R. (in press) SSLD: a French SMS to Standard Language Dictionary. *Cahiers du Cental*, 8.

Cougnon L.-A. and Ledegen G. (2009). *C'est écrire comme je parle*. Une étude comparatiste de variétés du français dans l''écrit sms'. In *Les voix des Français : usages et représentations*. Colloque AFLS à Oxford en avril. Peter Lang.

Dubois J. (editor) (2007). *Dictionnaire de linguistique et des sciences du langage*. Paris : Larousse.

Fairon C., Klein J. and Paumier S. (2006a). *Le langage SMS*. Louvain-la-Neuve : P.U.Louvain, Cahiers du Cental, 3.1.

Fairon C., Klein J. and Paumier S. (2006b). *Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation*. CD-Rom, Louvain-la-Neuve : P.U.Louvain, Cahiers du Cental, 3.2.

Gadet F. (1989). *Le français ordinaire*. Paris : Colin.

Gadet F. (2007). *La variation sociale en français*, 2e édition. Paris : Ophrys.

Howell D.C. (2008). *Méthodes statistiques en sciences humaines*, 6e édition. Bruxelles : de Boeck.

Jalabert R. (2006). MoliR, revi1 vit... il son 2vnu foo! (Molière, reviens vite ... ils sont devenus fous !). *Cahiers pédagogiques*, 440. [En ligne] http://www.cahierspedagogiques.com/article.php3?id_article=2165 (4 mars 2009).

Kerbrat-Orecchinoni C. (1992). *Les Interactions verbales*. Paris : Colin, 3 voll.

Koch P. and Oesterreicher W. (1985). Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36 : 15-43.

Myers-Scotton C. (1993). *Social motivations for codeswitching*. Oxford : Clarendon Press.

[32] Toutefois, ce taux d'abrègement ne dépasse pas 15,4%. Il n'est pas certain que ce taux soit suffisant pour conclure à une détérioration de la langue chez les jeunes lorsqu'ils utilisent le SMS.

Panckhurst R. (1997). La communication médiatisée par ordinateur ou la communication médiée par ordinateur ?. *TERMINOLOGIES NOUVELLES*, 17 : 56-58.

Panckhurst R. (2008). Short Message Service (SMS) : typologie et problématiques futures. In Arnavielle, T., editor, *Polyphonies*, Montpellier : Éditions LU.

Piette J., Pons Ch.-M. and Giroux L. (2007). Les jeunes et Internet (Appropriation des nouvelles technologies). Rapport final de l'enquête menée au Québec pour le Ministère de la Culture et des Communications, Gouvernement du Québec [En ligne] http://www.infobourg.com/data/fichiers/156/Les%20Jeunes%20et%20Internet%202006.pdf (Consulté le 17 octobre 2009)].

Sheskin D. (2004). *Handbook of parametric and nonparametric statistical procedures*. 3nd edition. Boca Raton (Floride) : Chapman and Hall/CRC.

Traverso V. (1996). *La conversation familière. Analyse pragmatique des interactions*. Lyon : Presses universitaires de Lyon.