Langage et sms

Langage et sms

Une étude internationale des pratiques actuelles

Louise-Amélie Cougnon

UCL PRESSES UNIVERSITAIRES DE LOUVAIN

Cet ouvrage est issu de la thèse de doctorat intitulée *L'écrit sms. Variations lexicale et syntaxique en francophonie*, présentée le 25 septembre 2012 à l'Université catholique de Louvain

CAHIERS DU CENTAL

Comité scientifique

Sergio Bolasco	Università di Roma « La Sapienza », IT
Laurence Danlos	Université Paris 7, FR
Guy Deville	Université de Namur, BE
Thierry Dutoit	Faculté Polytechnique de Mons, BE
Louisette Emirkanian	Université du Québec à Montréal, CA
Cédrick Fairon	Université catholique de Louvain, BE
Thierry Fontenelle	Microsoft Speech & Natural Language Group, USA
Sylviane Granger	Université catholique de Louvain, BE
Guy Lapalme	Université de Montréal, CA
Eric Laporte	Université de Marne-la-Vallée, FR
Denis Maurel	Université François-Rabelais de Tours, FR
Damon Mayaffre	Université de Nice – Sophia Antipolis, FR
Sébastien Paumier	Université de Marne-la-Vallée, FR
Alain Polguère	Université de Montréal, CA
Antoinette Renouf	University of Central England in Birmingham, UK
Jean Senellart	Systran, FR
Anne-Catherine Simon	Université catholique de Louvain, BE
Agnès Tutin	Université Stendhal – Grenoble 3, FR
Pierre Zweigenbaum	CHU Pitié-Salpêtrière, Université Paris 6, FR

Comité de rédaction

Anne Dister	CENTAL, Université catholique de Louvain, BE
Cédrick Fairon (directeur)	CENTAL, Université catholique de Louvain, BE
Patrick Watrin	CENTAL, Université catholique de Louvain, BE

Secrétariat de rédaction
CENTAL, Université catholique de Louvain, BE
Place Blaise Pascal, 1 – B-1348 Louvain-la-Neuve
tél. : +32 10 47 37 86 – fax : +32 10 47 26 06
email : cedrick.fairon@uclouvain.be
http://www.uclouvain.be/cental.html
Graphisme de la couverture : Olivier Vereecken http://www.aphine.com

© Presses universitaires de Louvain, 2015
Dépôt légal : D/2015/9964/9
ISBN : 978-2-87558-371-0
ISBN pour la version numérique (pdf) : 978-2-87558-372-7
Imprimé en Belgique par CIACO scrl – n° d'imprimeur 91256

Tous droits de reproduction, d'adaptation ou de traduction, par quelque procédé que ce soit, réservés pour tous pays, sauf autorisation de l'éditeur ou de ses ayants droit.

Diffusion : www.i6doc.com, l'édition universitaire en ligne
Sur commande en librairie ou à
Diffusion universitaire CIACO
Grand-Rue, 2/14, 1348 Louvain-la-Neuve, Belgique
Tél. 32 10 47 33 78 – fax 32 10 45 73 50 – duc@ciaco.com
Distributeur pour la France :
Librairie Wallonie-Bruxelles
46 rue Quincampoix - 75004 Paris
Tél. 33 1 42 71 58 03 – fax 33 1 42 71 58 09 – librairie.wb@orange.fr

Préface

La création des néologismes est souvent justifiée par l'apparition de nouvelles réalités, essentiellement matérielles et / ou techniques, qu'il faut nommer. Si cette doxa n'est pas fausse, elle est loin d'expliquer l'émergence de tous les néologismes que l'on rencontre. Bien d'autres raisons peuvent être invoquées. Parmi celles-ci il faut compter les modifications du rapport à l'écrit dans la Communication médiatisée (ou médiée) par ordinateur (CMO) que l'auteur préfère appeler communication écrite médiée par ordinateur (CEMO) qui s'est développée ces dernières décennies. Ces « écrits de proximité » autorisent en effet les scripteurs à des écarts de tous types ((ortho)graphiques, lexicaux, syntaxiques…) qu'ils ne se permettraient pas dans des productions plus formelles, tant écrites qu'orales d'ailleurs. Les SMS, comme leur nom l'indique, se singularisent par ailleurs au sein de la CEMO, entre autres caractéristiques, par la nécessité de la brièveté, ce qui n'est pas sans répercussion sur la manière d'utiliser la langue.

Dans son livre qui associe les outils de la linguistique de corpus, de la lexicologie, de la syntaxe et de la sociolinguistique essentiellement, Louise-Amélie Cougnon s'attelle précisément à étudier les spécificités de l'écrit SMS tant dans le vocabulaire employé (divers types de néologismes entre autres, dont des emprunts, des conversions, des accourcissements…) que dans des faits de syntaxe récurrents (code-mixing ou infractions par rapport aux normes dans l'expression de la négation, l'interrogation, dans l'emploi des temps…). Son étude est novatrice dans la mesure où peu de travaux ont été conduits sur ce phénomène récent (et il est fait état de ceux-ci au début du livre), en particulier sur la langue française, et ce par comparaison de plusieurs aires francophones.

La recherche a été conduite à l'Université de Louvain-la-Neuve -où ont été menés les travaux précurseurs dans le domaine de Cédric Fairon et Jean Klein- et elle s'appuie sur un travail de collecte de grande ampleur : 51 280 messages produits en Belgique, Canada, France et Suisse. Les disparités dans le nombre des SMS recueillis dans chaque pays ne nuisent pas aux conclusions qui peuvent être tirées dans la mesure où les données sont toujours suffisamment nombreuses pour être significatives.

La collecte des données, leur examen posent de nombreux problèmes méthodologiques (liés à des contraintes matérielles, éthiques, etc. diverses selon les pays d'origine des SMS du corpus) qui sont exposés clairement et franchement, avec des solutions pour les résoudre ou les contourner. Autant dire que la lecture de ce livre s'impose tant à ceux qui s'intéressent aux évolutions technologiques et à leurs retombées sociétales, qu'à ceux qui s'intéressent à la langue française et à ses évolutions.

Jean-François Sablayrolles, Université Paris 13

Table des matières

I Introduction **1**

1 Problématique **3**

2 Objectifs généraux de la thèse **7**
2.1 Objectifs méthodologiques 7
2.2 Objectifs théoriques 8
2.3 Objectifs empiriques 9
2.4 Questions de recherche 10

3 Plan de la thèse **11**

4 Conventions **13**
4.1 Traductions 13
4.2 Transcriptions 13
4.3 Etiquettes 13

II De la théorie au recueil de données **15**

5 sms et communication écrite médiée par ordinateur **17**
5.1 Communication, médiation et informatisation 18
5.1.1 Un état de l'art 20
5.1.1.1 Communication écrite médiée par ordinateur 20
5.1.1.2 Écrit sms 24
5.1.2 Traits de la communication médiée 27
5.1.3 Principaux types de la Cémo 28
5.1.3.1 Messagerie instantanée et chat 29
5.1.3.2 Courrier électronique 30
5.1.3.3 Forums de discussion 32
5.1.3.4 Sites de réseautage social 33
5.1.3.5 Muds ou communication dans les univers virtuels . . 36
5.2 L'écrit sms et ses phénomènes graphiques 41
5.2.1 Un nouveau code graphique 42
5.2.2 Phénomènes graphiques et lexicaux 46
5.2.2.1 Abréviations graphiques 48
5.2.2.2 Emoticônes 52
5.2.2.3 Ponctuation et casse expressives, caractères échos . 55
5.2.3 Un écrit oralisé, spontané ou simplement un langage ordinaire ? 56
5.2.4 Conclusions 60
5.3 Conclusions 61

6 Recueil et traitement de données 63
6.1 Constitution de corpus de sms . . . 64
6.1.1 Linguistique de corpus et sms . . . 66
6.1.2 Des sms comme corpus de données textuelles . . . 70
6.1.3 Collecte des données . . . 73
6.1.3.1 Méthodologie de collecte . . . 73
6.1.3.2 Collectes régionales . . . 77
6.1.3.3 L'échantillon de données sms . . . 80
6.1.4 Comparabilité des données . . . 83
6.1.4.1 Format, annotation et nettoyage . . . 83
6.1.4.2 Alignement : transcription en graphie standard . . . 92
6.2 Recueil de données sociodémographiques . . . 101
6.2.1 Constitution et traitement de formulaires . . . 102
6.2.1.1 Tronc commun . . . 103
6.2.1.2 Limites des questionnaires élaborés . . . 104
6.2.2 Distribution des variables . . . 110
6.2.3 L'échantillon de locuteurs . . . 118
6.3 Conclusions . . . 124

III De la variation des données 125

7 Préliminaires 127
7.1 La notion de variation pour un corpus de sms . . . 127
7.2 Approches guidées par corpus et fondées sur corpus . . . 129
7.3 Outils d'analyse . . . 131

8 Variation lexicale 135
8.1 Questions générales de lexicologie . . . 135
8.1.1 Lexicométrie . . . 136
8.1.2 Cooccurrences . . . 141
8.1.3 Néologismes . . . 143
8.1.3.1 Introduction et méthodologie . . . 143
8.1.3.2 Approche qualitative . . . 146
8.1.3.3 Variation diatopique . . . 148
8.1.3.4 Variation diastratique . . . 149
8.1.4 Conclusions . . . 151
8.2 Éléments allogènes . . . 154
8.2.1 Approche qualitative . . . 154
8.2.1.1 Typologie . . . 154
8.2.1.2 Composition des éléments allogènes . . . 157
8.2.1.3 Fonctions des mélanges de langues . . . 159
8.2.1.4 Conclusions sur les éléments allogènes d'après l'analyse qualitative . . . 164

8.2.2 Variation diastratique . . . 166
8.2.2.1 Belgique . . . 166
8.2.2.2 La Réunion . . . 169
8.2.2.3 Québec . . . 172
8.2.2.4 Suisse . . . 175
8.2.2.5 Corpus global : langue de l'élément allogène . . . 178
8.2.3 Conclusions . . . 182
8.3 Régionalismes lexicaux . . . 183
8.3.1 Régionalismes de Belgique . . . 185
8.3.1.1 Approche qualitative . . . 186
8.3.1.2 Variation diastratique . . . 190
8.3.2 Régionalismes de la Réunion . . . 193
8.3.2.1 Approche qualitative . . . 194
8.3.2.2 Variation diastratique . . . 196
8.3.3 Régionalismes du Québec . . . 197
8.3.3.1 Approche qualitative . . . 198
8.3.3.2 Variation diastratique . . . 200
8.3.4 Régionalismes de Suisse . . . 202
8.3.4.1 Approche qualitative . . . 203
8.3.4.2 Variation diastratique . . . 206
8.3.5 Conclusions . . . 209
8.4 Conclusions . . . 212

9 Variation orthographique 215
9.1 Introduction . . . 215
9.2 Variantes à dominantes morphogrammiques . . . 219
9.3 Variation diastratique . . . 224
9.4 Conclusions . . . 227

10 Variation syntaxique 229
10.1 Interrogatives . . . 232
10.1.1 Introduction . . . 232
10.1.2 Variation interne . . . 234
10.1.3 Variation externe . . . 240
10.1.4 Conclusions . . . 247
10.2 Négation . . . 249
10.2.1 Introduction . . . 249
10.2.2 Variation interne . . . 249
10.2.3 Variation externe . . . 259
10.2.4 Conclusions . . . 265
10.3 Conclusions . . . 267

IV Conclusions **269**

11 Conclusions générales de la recherche **271**

12 Limites et perspectives **279**

Bibliographie **281**

A Annexes **307**
A.1 Annexes au Chapitre 1 307
A.2 Annexes au Chapitre 2 320
A.2.1 Questionnaire belge 320
A.2.2 Questionnaire réunionnais 321
A.2.3 Questionnaire québécois 323
A.2.4 Questionnaire suisse 326
A.3 Annexes au Chapitre 3 328
A.3.1 Lexicométrie 328
A.3.2 Néologismes 332
A.3.3 Emprunts 342
A.3.4 Régionalismes 358
A.4 Annexes au Chapitre 4 371
A.4.1 Morphosyntaxe 371
A.4.2 Interrogatives 379
A.4.3 Négatives 384

Première partie

Introduction

Chapitre 1

Problématique

L'essor des technologies de l'information et de la communication (TIC) a marqué de façon décisive le XX^{e} siècle : par la radiodiffusion, la télévision, la téléphonie ou Internet, tant l'information de masse que la communication interpersonnelle ont subi une évolution – voire une mutation – sans précédent. Le développement de ces technologies, notamment dans le domaine de l'informatique, a donné naissance à de nouvelles formes de communication entre locuteurs, par le biais des machines [1].

La notion de *communication médiatisée* (ou *médiée*) *par ordinateur*, abrégée *CMO*, est apparue au début des années 80, au sein du milieu scientifique international [2], pour qualifier les nouvelles formes de communication. Ce milieu considérait alors l'ordinateur comme un simple véhicule ou « médium » et ne distinguait pas les différents types de la CMO [3]. Depuis, il est devenu presque absurde d'omettre la singularité de chaque type et d'ignorer l'influence du médium sur la nature même de la communication (Ko, 1996 ; Panckhurst, 1999). On sait aujourd'hui que l'évolution continuelle du traitement de l'information entraine une mutation dans les genres d'écriture autant que dans les modes de lecture (Lòpez Alonso, 2003).

Certaines des nouvelles formes de communication usent essentiellement de l'écrit ; nous nommons cet ensemble la *communication écrite médiée par ordinateur*, ou *Cémo* [4]. Le *sms*, sigle pour *Short Message System*, s'inscrit pleinement dans cet ensemble. Il est défini comme étant « un service permettant d'envoyer et de recevoir de brefs messages écrits sur un téléphone mobile », mais il s'agit aussi du « message ainsi échangé » (Rey *et al.*, 2009). Le *langage sms* que nous préférerons appeler *écrit sms* (dorénavant *eSMS*, cf. sous-chapitre 5.2.1), est une pratique relativement récente [5], raison pour laquelle, contrairement aux autres types de la Cémo, tels que le courrier électronique, le chat ou les forums, elle ne fait l'objet d'études approfondies que depuis une dizaine d'années. Ces études, comme nous le

1. Notons que ce développement n'est pas géographiquement ni socialement uniforme. Comme l'explique Thurlow (2003, page 1) : « which is not to say that this technology is properly global ; worldwide patterns of [media development] necessarily follow the socioeconomic contours of which distinguish the 'media rich' and 'media poor' generally ».

2. En anglais on dira *CMC* ou *Computer Mediated Communication*.

3. À l'époque, chat, courriel électronique et forums de discussion existaient déjà, en version commerciale et en usage plutôt restreint.

4. Voir le chapitre 5.1.

5. 3 décembre 1992 pour le premier sms d'ordinateur à machine, 1999 pour une utilisation grand public auprès des compagnies de téléphonie (`http://news.bbc.co.uk/2/hi/uk_news/2538083.stm` et `http://timesofindia.indiatimes.com/articleshow/30216466.cms`).

verrons dans la partie 5.2, se sont souvent efforcées de décrire et de comprendre la graphie des sms, portant peu d'intérêt aux autres phénomènes linguistiques, comme pourraient le faire des lexicologues et des syntacticiens. Très récemment, on a vu émerger quelques recherches, principalement en anglais, en japonais ou en finlandais, dont les points centraux se sont révélés originaux : étude des canaux d'ouverture et de fermeture, structures questions-réponses, anaphores, etc. Ces études se basent principalement sur des corpus restreints (Laursen, 2005), récoltés en vue d'un objectif de recherche particulier (Hård Segerstad, 2005 ; Thurlow, 2003). En outre, et comme nous venons de le faire remarquer, peu de ces études s'intéressent à des données francophones[6].

La Cémo est devenue, en une décennie à peine, un phénomène incontournable du quotidien, et en particulier au sein des classes d'âge les plus jeunes, pour qui la pratique de l'écrit est plus importante aujourd'hui qu'elle ne l'a jamais été (Höflich et Gebhardt, 2005). Il serait difficile, dans ce contexte, d'ignorer le nombre d'heures que chacun passe à « chater » sur un ordinateur ou à « tapoter » sur un petit clavier. Dans une société où la distance prise par rapport aux livres et aux cahiers croît chaque jour, la pratique de l'écrit est pourtant désormais chronophage.

La fréquence élevée de cette pratique pose question, notamment quant aux probables répercussions de cet usage sur les compétences linguistiques générales des plus jeunes, qui maitrisent souvent les outils de la Cémo avant même d'avoir assimilé leur langue première. Certains chercheurs vont jusqu'à penser que les usagers de la Cémo réinventent considérablement le langage par leurs pratiques (Shafie *et al.*, 2010).

Afin de mesurer scientifiquement de tels impacts, il est nécessaire d'observer la réalité des productions « sur le terrain ». Pour y parvenir, les chercheurs se doivent de collecter des données et de les analyser aussi objectivement que possible. Cette tâche s'avère complexe, par la difficulté de récolter des données valables ou de définir des tests qui permettent de quantifier scientifiquement les impacts. Devant cette complexité, nous distinguons cinq phases essentielles pour la recherche à venir :

1. La constitution de corpus de Cémo, scientifiquement valables et légalement exploitables. L'enregistrement de données de chat, de courrier électronique, de forum de discussion est techniquement réalisable, mais l'exploitation de ces données n'est pas toujours rendue possible pour des raisons éthiques et juridiques. La collecte de corpus de sms représente encore un degré de difficulté supplémentaire, les opérateurs de téléphonie ne donnant pas accès à ceux-ci ;
2. L'exploitation linguistique des corpus en vue d'en donner les propriétés et la définition. Cette définition concerne tous les versants linguistiques de la Cémo :

6. On citera par exemple David et Goncalves (2007), l'étude de Fairon *et al.* (2006b) basée sur le corpus *sms pour la science* (Fairon *et al.*, 2006a), et celle de Stark (2011), basée sur les données suisses que nous exploitons également ici.

autant la morphologie, que la syntaxe, le lexique, les aspects dialogaux, les propriétés particulières de l'interaction, etc. ;

3. La constitution de corpus larges et hétérogènes de productions écrites hors Cémo (devoirs scolaires, journaux intimes, correspondance traditionnelle, etc.) en tant que base de l'évaluation des impacts ;
4. La mise sur pied d'une méthodologie de quantification des impacts (mettre en lien telle variante orthographique avec la pratique de la Cémo et non avec une limitation liée à l'enseignement scolaire des règles, par exemple) [7] ;
5. La réitération de cette évaluation à travers le temps en vue de valider des résultats qui pourraient être interprétés comme un « effet de mode ».

7. Cette étape n'est pas triviale étant donné les méthodes abordées jusqu'ici dans les quelques rares études sur le sujet qui visent plutôt à trouver des variantes orthographiques dans les écrits traditionnels qui correspondraient à des procédés graphiques typiques de la Cémo : par exemple les études de Lunsford (Université de Stansford), « l'auteure n'a pas trouvé un seul procédé graphique associé au texto dans les travaux écrits des étudiants » et Varnhagen (Université d'Alberta), qui n'a « trouvé aucune relation entre l'emploi des variantes orthographiques dans cette situation de communication et l'habileté à orthographier correctement en anglais », toutes deux citées dans Tatossian (2011).

Chapitre 2

Objectifs généraux de la thèse

2.1 Objectifs méthodologiques

L'apport de notre dissertation doctorale se situe aux étapes (1) et (2) de la méthodologie générale d'évaluation de l'impact des nouvelles pratiques de l'écrit. Nous avons entrepris de (1) constituer des corpus de sms authentiques (non « fabriqués » pour l'occasion) et exploitables pour la recherche (stockés, harmonisés, transcrits en graphie standard). Pour relever les différences et similarités dans nos différents corpus de sms, nous avons colligé des corpus dits « comparables » (cf. sous-chapitre 6.1.4). La méthodologie de collecte, précisée dans la partie 6.1, met en avant, malgré toutes les précautions prises, deux problèmes majeurs de la linguistique de corpus : le biais de l'observateur et la représentativité d'un corpus. Le biais engendré par la présence de l'observateur se situe dans la sélection faite par les usagers des messages à nous faire parvenir. Nous avons constaté que nous recevions un grand nombre de messages à caractère cocasse (qu'ils aient été réellement échangés ou créés pour l'occasion), entrainant une espèce d'exhibitionnisme ; d'un autre côté, nous ne pouvons qu'admettre un tri dans les messages envoyés, mais ce tri concerne davantage les thématiques que le lexique et la syntaxe, ce qui est dit et non la façon de le dire, qui est l'objet de notre recherche. La représentativité des données fait également l'objet d'un questionnement : nous démontrons (partie 6.2.3) la difficulté d'obtenir une représentativité de chaque corpus de sms par rapport à la population effective correspondante (qu'il est tout aussi difficile de définir). Les quatre corpus francophones de Belgique, de La Réunion, de Suisse et de Québec que nous possédons présentent des données non homogènes dans leur représentativité de la population de la région concernée[1] (cf. Figure 2.1). En Suisse, le corpus de 3.889 sms est censé représenter 5.400.000 francophones[2] ; ces francophones sont sous-représentés. La Réunion a récolté 12.660 sms pour une population de 808.000 habitants (en 2008). En Belgique, 30.000 sms ont été colligés, pour une population de 4.139.310 millions de francophones[3] ; les belges sont surreprésentés dans le corpus général. Enfin, au Québec, le corpus comprend 4.731 sms, pour une

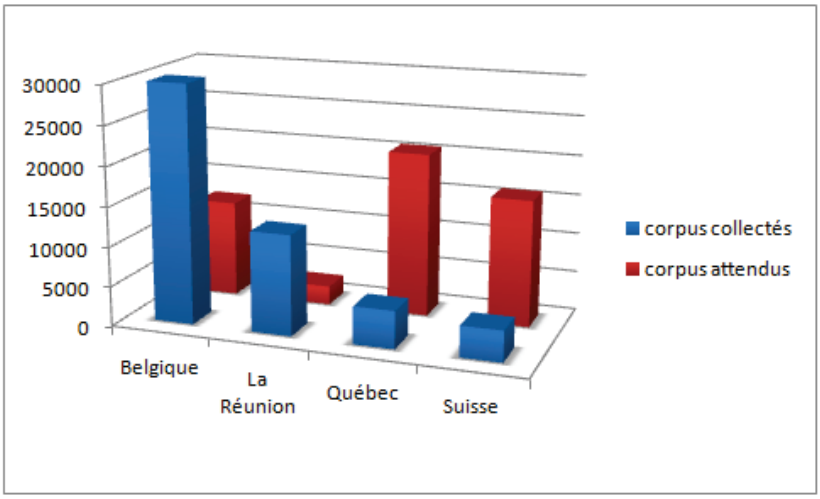
1. Ainsi que dans la représentativité en termes de nombre de sms.

2. Données pour 2010 issues de Wolff (2010) cité sur le site du *Forum francophonie* : http://www.forumfrancophonie2012.org/wordpress/wp-content/uploads/2012/01/Tableau-des-locuteurs-du-fran%C3%A7ais-pays-membres-de-lOIF.pdf.

3. Données pour 2004 issues du CIA World Fact Book : http://en.wikisource.org/wiki/CIA_World_Fact_Book,_2004/Belgium

population de 6.980.000 millions de francophones [4]. Nous verrons comment, malgré ce problème de représentativité, nos données s'avèrent tout à fait exploitables.

FIGURE 2.1 – *Comparatif des corpus colligés par rapport aux corpus attendus en fonction de la population globale de la région échantillonnée*



Notre objectif de recherche (2) s'est ensuite centré sur la définition des versants lexical et syntaxique de l'eSMS. Notre approche emploie les techniques et les fondements de la linguistique de corpus et de la sociolinguistique. La linguistique de corpus nous apporte des techniques de comptage (type, tokens, etc.) et d'analyse (concordances, collocations, etc.). L'approche sociolinguistique, quant à elle, permet d'étudier les traits linguistiques de ce mode en fonction d'informations extralinguistiques (région, profil et pratiques du locuteur), disponibles grâce à la collecte de profils. En effet, la constitution des corpus de sms s'est accompagnée d'un recueil de données sociodémographiques concernant les auteurs des messages. L'analyse sociolinguistique des traits lexicaux et syntaxiques de l'eSMS est donc fondée sur cette double base de données ainsi que sur des outils statistiques. Les statistiques ne guident jamais seules l'analyse d'un phénomène linguistique, elles viennent la compléter.

2.2 Objectifs théoriques

Nous tentons tout d'abord de circonscrire l'eSMS. Une comparaison détaillée de ce type avec les autres types de la Cémo est menée. Une définition de cet écrit est ensuite proposée. L'eSMS se rapproche, comme nous le verrons, par son cadre moins formel, d'un écrit « de proximité » (Koch et Oesterreicher, 1985). Certains comparent cet écrit à l'oral, d'autres au spontané. Nous évitons ces rapprochements,

4. Données pour 2010 issues de Wolff (2010) cité sur le site du Forum francophonie : http://www.forumfrancophonie2012.org/wordpress/wp-content/uploads/2012/01/Tableau-des-locuteurs-du-fran%C3%A7ais-pays-membres-de-lOIF.pdf.

mais cela ne nous empêche pas de comparer nos données à des corpus du français parlé et écrit. On sait par exemple que la production orale est très variable d'une région à l'autre : nous regarderons si cela se manifeste également d'un corpus de sms à l'autre.

Nous examinons les spécificités du lexique et de la syntaxe présents dans ces corpus. À partir de deux listes d'unités lexicales connues et inconnues d'un dictionnaire de référence, nous décrivons les spécificités de cette production : créations lexicales, régionales et mélange de langues. Notre objectif est de répertorier et quantifier ces spécificités, d'en donner la nature afin d'ouvrir des voies encore peu empruntées par la science : quel lexique emploie-t-on dans cet écrit ? La partie syntaxique est, pour sa part, dirigée par des hypothèses émises dans la littérature (principalement celles opposant les traits de l'écrit à ceux de l'oral) et des idées reçues que l'on retrouve, par exemple, dans la presse. Nous tentons de répondre aux questions alarmantes relatives aux « fautes d'orthographe » et de comparer les données aux résultats basés sur la formulation des interrogatives et des négatives dans le français parlé et écrit.

Dans l'approche de nos données francophones, nous nous situons dans la lignée de l'approche renouvelée de la variation (Hambye, 2005) : en ce sens, l'eSMS n'est pas la simple conséquence d'une langue standard abrégée, ludique et simplifiée, mais peut également être le résultat de volontés individuelles (transmettre de l'expressivité, faire valoir son statut, sa position sociale, etc.), de propriétés dialogales (partager des émotions, la solidarité, impressionner, surprendre, etc.) et de critères sociaux (renvoyer à des croyances et des pratiques sociales, faire valoir une appartenance ethnique, etc.).

2.3 Objectifs empiriques

Le but de la méthodologie que nous avons suivie est de permettre la découverte de résultats quantitativement valables, basés sur des phénomènes récurrents. Pour ce faire, il fallait donc rassembler un corpus conséquent en nombre : nous avons rendu possible la collecte de 51.280 messages. De ce fait, même si certaines unités lexicales ou structures syntaxiques sont peu fréquentes, la plupart de nos conclusions se basent sur un nombre d'occurrences élevé.

Nous avons également entrepris de nous aider de l'outil statistique : ce choix, critiquable sous certains aspects[5], permet notamment de dépasser les limitations qu'induit une étude basée sur les fréquences et les pourcentages. En effet, et c'est là notre cas, un phénomène est souvent davantage représenté par la classe d'âge la plus jeune parce que celle-ci a envoyé plus de messages aux organisateurs du projet. Les

5. « On fait dire ce que l'on veut aux chiffres ».

études statistiques permettent d'étudier le lien d'un phénomène linguistique avec des paramètres extralinguistiques en fonction de la distribution de ces paramètres pour le corpus total observé. Ainsi, nous avons pour objectif de confronter des résultats énoncés dans la littérature à propos des pratiques de la Cémo propres à un sexe, à une région, à un âge (Van Compernolle et Williams, 2007), etc. avec nos propres résultats, qui viendront les confirmer, les infirmer, voire les compléter.

2.4 Questions de recherche

Pour résumer les trois objectifs méthodologique, théorique et empirique que nous venons de développer et qui ont dirigé notre travail doctoral, nous présentons ici les questions de recherches principales de cette dissertation :

1. Quelles sont les spécificités linguistiques et extralinguistiques de l'eSMS par rapport à la Cémo en général ?
2. Quels sont les enjeux et les limites d'une collecte de sms et de profils sociodémographiques ?
3. Dans quelle mesure les assertions connues au sujet de l'écrit sms (abréviation systématique, tendance à l'informel, « destruction » du langage ou de la syntaxe) sont-elles vérifiées par l'exploration d'une large quantité de données sms ?
4. Quel lexique emploie-t-on dans les sms ?
5. Dans quelle mesure l'eSMS, en tant que médium de communication écrit, partage-t-il des traits du français ordinaire ?
6. L'écrit sms témoigne-t-il de variations régionales et de variations tributaires du profil sociolinguistique de l'auteur ?
7. Existe-t-il, au contraire, des constantes de l'écrit sms que l'on retrouve dans tous les sous-corpus (quels que soient la région, le profil de l'auteur, etc.) ?

Chapitre 3

Plan de la thèse

Cette dissertation débute, avec les chapitres 5 et 6, par une partie théorique et méthodologique.

Le sous-chapitre 5.1 propose un panorama de la recherche consacrée à la Cémo, principalement en sociolinguistique, mais également dans les domaines de la psycholinguistique, de la pédagogie et du traitement automatique des langues. Les différents types de Cémo sont définis, décrits et comparés en fonction d'un certain nombre de critères inspirés de Herring (2007). L'eSMS est détaillé en fonction de ces critères afin de cerner le plus précisément possible ses spécificités.

La définition de l'eSMS fait l'objet du sous-chapitre 5.2. Après avoir mené un tour d'horizon des diverses études et corpus liés à notre sujet de recherche, nous détaillons les phénomènes graphiques (abréviations, émoticônes, ponctuation expressive, etc.) liés à l'eSMS et qui sont très régulièrement abordés dans la littérature. Nous entamons ensuite une évaluation des notions de français ordinaire, de code, d'oralité et de spontanéité souvent employées pour caractériser l'eSMS ou la Cémo en général.

Le sous-chapitre 6.1 présente les particularités de la méthodologie d'obtention des données, en mettant la lumière sur les corpus de sms : les processus de collecte, d'anonymisation et de transcription y seront détaillés. Nous confrontons notre démarche à la réalité de terrain en abordant la question de la représentativité des données. Nous précisons également les concepts d'alignement et de comparabilité dans le cadre de ces corpus de textes.

L'exploration des données sociolinguistiques fait l'objet du sous-chapitre 6.2 : nous accordons une attention particulière à la nature des formulaires d'enquête ainsi qu'à leurs limites, principalement dans la visée d'un méga corpus de données comparables. Nous détaillons les profils sociolinguistiques recueillis dans chaque région et nous déterminons les types de données comparables, ou « variables », qu'il sera possible d'exploiter dans le cadre d'une étude sur la variation à travers ces régions.

La partie dédiée à l'analyse de données comprend les chapitres 7, 8, 9 et 10.

Le chapitre 7 a pour objectif d'introduire les angles d'approche analytiques. Nous définissons notre emploi de la notion de variation (diatopique et diastratique).

Ensuite, nous distinguons les techniques d'études fondées sur corpus de celles guidées par corpus ; cette distinction explique la différence d'approche entre l'analyse lexicale et l'analyse syntaxique que nous menons. Enfin, nous présentons la nature, la fonction et les limites des tests statistiques employés.

Le chapitre 8 est dédié à l'analyse de la variation lexicale dans les corpus de sms. Les points abordés touchent à la lexicologie générale (lexicométrie, collocations, etc.), mais aussi, en particulier, aux néologismes, aux éléments allogènes et aux régionalismes lexicaux. Nous tentons de dégager une typologie de ces phénomènes propre à l'eSMS.

Une partie plus restreinte, le chapitre 9, se penche sur la variation orthographique dans l'eSMS. La question de l'orthographe est envisagée en suivant la grille de Catach (1980) ; certaines variations sont passées à la loupe, notamment afin d'envisager le taux de variantes et la distribution de ces variantes au sein de la population.

L'étude de la variation syntaxique, développée dans le chapitre 10, porte sur des questions récurrentes de la syntaxe que sont la rétention du *ne* de négation et la structure des interrogatives. Nous reprenons des typologies émises par ailleurs dans la littérature et nous observons leur application au corpus de sms.

Pour chaque phénomène analysé dans cette dernière partie analytique, nous employons d'abord une approche qualitative, que nous complétons à l'aide d'un regard quantitatif, révélant les chiffres liés à chaque phénomène et les tendances mises en avant par les tests statistiques, au regard des profils des auteurs et des régions enquêtées.

Chapitre 4

Conventions

Tout au long de ce tapuscrit, nous avons utilisé un certain nombre de conventions d'écriture et de mise en forme que nous avons choisi d'expliciter clairement dès cet endroit.

4.1 Traductions

Toute citation dans une langue étrangère est traduite dans le texte afin de ne pas perturber la lecture et de favoriser une compréhension aisée. Les citations en note de bas de page ne sont pas traduites. Les passages dans la langue originale sont reportés en note de bas de page, comme ceci [1]. Lorsque l'auteur de la traduction n'est pas indiqué, cela signifie qu'il s'agit de notre propre traduction.

4.2 Transcriptions

Tout exemple provenant d'un corpus de sms est donné dans sa version transcrite en graphie standardisée excepté lorsque les études sur la morphologie demandent la présentation de la version brute. Dans ce dernier cas, nous indiquerons l'exemple en rouge et sa version transcrite sera donnée en note de bas de page. Tous les exemples de sms dans leur version brute sont repris, précédés de leur numéro, en annexe. Quelques exemples de sms présentent des passages en langue étrangère ; ces passages seront traduits dans le texte pour faciliter la lecture et mis en italique, excepté lorsque les études sur l'alternance de langues demandent la présentation des passages en langue originale. Dans ce dernier cas, nous indiquerons le passage en vert et la version traduite du passage sera précisée en note de bas de page.

4.3 Etiquettes

Pour un exposé plus simple et pourtant précis, tout exemple provenant du corpus de sms est suivi d'une étiquette de type

(1) message [smsFX] [2]

1. « Citation originale » (traduit par auteur de la traduction).

2. Le style de ces étiquettes est inspiré des codes utilisés par le centre de recherche Valibel (http ://www.uclouvain.be/valibel.html), qui gère une base de données textuelles comprenant 373 heures d'enregistrement transcrites et encodées sur support informatique.

qui précise la provenance exacte de l'exemple : sms rappelle le type de données, F le caractère français du corpus et X, seul élément commutable, peut être décliné en :
- B pour français de Belgique ;
- S pour français de Suisse romande ;
- R pour français de La Réunion ;
- Q pour français du Québec.

Deuxième partie

De la théorie au recueil de données

Chapitre 5

sms et communication écrite médiée par ordinateur

La communication médiatisée par ordinateur (CMO) est en train de devenir tellement omniprésente qu'elle passerait presqu'inaperçue.
George et Totsching (2001)

Sommaire

5.1 Communication, médiation et informatisation 18
5.1.1 Un état de l'art 20
5.1.2 Traits de la communication médiée 27
5.1.3 Principaux types de la Cémo 28
5.2 L'écrit sms et ses phénomènes graphiques 41
5.2.1 Un nouveau code graphique 42
5.2.2 Phénomènes graphiques et lexicaux 46
5.2.3 Un écrit oralisé, spontané ou simplement un langage ordinaire ? 56
5.2.4 Conclusions 60
5.3 Conclusions 61

Présentation du chapitre

La Cémo est devenue en quelques années tellement populaire, autant au niveau de sa pratique qu'à celui des controverses qu'elle déclenche dans les médias, la population et le monde scientifique, qu'il est devenu pratiquement impossible, en tant que linguiste, de l'ignorer ; un *Laboratoire de communication médiatisée par ordinateur* (ou *LabCMO*[1]) a par exemple été créé en 2002. Malgré cet engouement qui présente souvent la Cémo comme une entité homogène, il existe des types bien distincts de nouvelles productions écrites. Dans une première partie, nous livrerons quelques données historiques relatives à la Cémo. Nous présenterons ensuite l'évolution des approches et évoquerons les diverses disciplines qui ont saisi la Cémo comme objet d'étude. Enfin, nous présenterons les types qui composent la Cémo et, dans le but de situer précisément l'eSMS au sein de ces nouvelles productions, nous établirons une liste - que nous souhaitons la plus exhaustive possible - de critères descriptifs.

1. Pour en savoir plus : `http://cmo.uqam.ca/`.

5.1 Communication, médiation et informatisation

Introduction

La CMO, communication (originellement *médiatisée* puis) *médiée*[2] par ordinateur[3], dont le pendant anglais est CMC (« computer-mediated communication »[4]), est définie dans la littérature de différentes façons, qui nous semblent parfois incomplètes ou imprécises. Herring (2007) propose la définition succincte suivante : « Interactions écrites d'humain à humain via des ordinateurs branchés en réseau ou via la téléphonie mobile »[5]. Cette définition pourrait convenir pour notre objet d'étude, mais elle est pourtant inadaptée puisque qu'elle intègre l'écrit (« text-based ») dans un sigle qui n'y fait pourtant pas allusion et qui renvoie à tout type de communication médiée par ordinateur, dont font partie les téléconférences, les chats vocaux dans les jeux en réseau, etc. Nous notons le même problème avec Anis (2002), qui décrit la CMO comme des « échanges dont les messages, affranchis des supports matériels habituels de l'écriture grâce à des codages numériques, sont véhiculés par des réseaux télématiques ». Pour leur part, Véronis et Guimier De Neef (2006) refusent la dénomination *CMO* au profit de *NFCE* (ou *nouvelles formes de communication écrite*), qui couvre l'ensemble des écrits diffusés sur les vecteurs numériques de communication (sites web, e-mails, forums, messageries instantanées, sms, blogs, etc.). Leur choix est triplement motivé : premièrement, l'appellation *CMO* n'offrirait pas de référence directe à l'écrit et engloberait donc tout ce qui est de l'ordre de la communication orale. Ensuite, elle n'inclurait pas le domaine de la téléphonie, et donc des sms. Enfin, l'appellation, qui renvoie originellement à *communication médiatisée par ordinateur*, pour laquelle *médiatisée* a bien le sens de « servir d'intermédiaire », est en réalité une utilisation très rare du participe, qui entrerait trop souvent en concurrence avec la signification plus commune « popularisé à travers les médias ». Les auteurs précisent bien que

2. Appellation initiée par Panckhurst (1997). L'auteure utilise l'adjectif *médié* au sens de la « médiation » proposée par Vygotski (1997) (cité dans Panckhurst 1997, p. 57), « un outil vecteur de discours, et qui le modifie indirectement ». Notons que Panckhurst a également préféré *médié* à *médiatisé* parce que ce dernier est déjà fortement connoté, renvoyant initialement à « diffusé par les médias ».

3. D'autres chercheurs ont proposé des variantes à la CMO, telles que *communication électronique* (Anis *et al.*, 2004), ou *communication électronique scripturale* (Anis *et al.*, 2004), *nouvelles technologies de l'information et de la communication* (NTIC, Caron et Caronia 2005), Netspeak (Crystal, 2001), *cyberlangage* (Dejond, 2002), *nouvelles formes de communication écrite* (NFCE, Guimier de Neef et Veronis 2004) ou encore la *communication médiatisée par les technologies de l'information et de la communication* (CMTIC, Vold Lexander 2009), que nous trouvons tous trop imprécises. Panckhurst (2007) propose encore le *discours électronique médié* (DEM), qui nous semble pour sa part trop précis, trop orienté « discours » justement.

4. On notera que les Anglo-saxons emploient aussi bien la notion de *Netspeak* pour signifier la même réalité, caractérisée par exemple par les abréviations, les émoticons et l'orthographe non standard (Crystal, 2001). Nous ne retenons pas cette étiquette qui est évidemment hors de propos pour tous les types de communication qui, comme le sms, ne sont pas véhiculées par le réseau Internet.

5. « Text-based human-human interaction mediated by networked computers or mobile telephony ».

Panckhurst (1997) a résolu ce problème en remplaçant *médiatisée* par *médiée*, mais n'en concluent par pour autant qu'il faille adopter l'appellation corrigée. Nous trouvons leur argumentaire intéressant pour ce qui est de la réalisation écrite. Pour ce qui est de l'exclusion du sms du domaine de l'informatique, le téléphone portable, y compris lors de l'emploi de sms, est une technologie qui présente des composantes électroniques, permettant des fonctions semblables à celles des ordinateurs ; le téléphone portable de base [6] est en quelque sorte un mini ordinateur qui permet la rédaction et l'envoi du sms, la suggestion orthographique, l'introduction de signes de ponctuation, voire même de smileys-images [7]. Nous pensons donc, à la suite de Bieswanger (2007) [8], que l'ordinateur en est bien le médium. Pour ce qui est du troisième point relevé par Véronis et Guimier De Neef (2006), nous pensons que l'alternative *médiée* proposée et justifiée par Panckhurst (1997) convient parfaitement.

Personnellement, nous refusons l'adoption du sigle *NFCE* pour trois raisons majeures : premièrement, l'adjectif *nouvelles*, représenté par le N, ne conviendra qu'un temps ; nous pensons qu'il ne convient déjà plus pour un type tel que le courrier électronique qui existe depuis une quarantaine d'années et qui est entré dans l'usage quotidien il y a vingt ans. Ensuite, NFCE ne fait pas référence à l'informatique ou à l'électronique ; il ne fait d'ailleurs pas du tout allusion au médium de communication. Or, comme l'expliquent Höflich et Gebhardt (2005, p. 14-15), le véhicule influe toujours, d'une façon ou d'une autre, sur le langage : il « n'est pas un simple véhicule (neutre) pour la transmission de messages. [Il] montre toujours un sens méta-communicatif qui a un effet sur le contenu de la communication ». La notion de véhicule du langage est essentielle en ce qu'elle peut autant se référer à une machine (le médium ou canal de communication) qu'à un humain ; tous les types de la Cémo sont médiés par une machine (un ordinateur ou un mini ordinateur). Pour certains types, tels que le courrier électronique et le sms, il s'agit là de la seule médiation effective ; mais pour d'autres (chat, forums, sites de « réseautage », etc.), il existe deux médiations successives : la médiation par ordinateur et la médiation humaine (on pensera aux « modérateurs » qui, par la sélection ou la censure notamment, modifient le contenu des messages).

Nous avons désiré solutionner ce problème terminologique en proposant la *Cémo*, ou *communication écrite médiée par ordinateur*, qui, non seulement, met en adéquation le sigle – qui devient ici acronyme – et sa définition, mais aussi, qui conserve les éléments pertinents relevés dans la littérature : il s'agit bien d'une forme de commu-

6. Laissons de côté les téléphones plus performants, dit « intelligents » (*smartphones*), qui coïncident pour leur part totalement avec des mini ordinateurs.

7. Il s'agit de quelques émoticônes qui ne sont pas une suite de signes de ponctuation, mais d'images représentant le faciès en question.

8. « Although not strictly "computer-mediated", text messaging as a form of technologically-mediated [...] is now frequently counted among the modes of computer-mediated communication (CMC) and computer-mediated discourse » (Bieswanger, 2007, p. 2).

nication dont la réalisation est écrite et dont le véhicule « influant » est l'ordinateur, accepté dans son sens le plus large.

5.1.1 Un état de l'art

5.1.1.1 Communication écrite médiée par ordinateur

L'étude de la communication médiée par ordinateur voit le jour dans les années 1970 avec les travaux précurseurs de Turoff et Hiltz : les auteurs s'intéressent alors à l'impact des conférences via ordinateur sur les rapports sociaux (Turoff et Hiltz, 1977). Les recherches sur la Cémo s'intensifient dans les années 80, au moment où le courrier électronique et le chat intègrent massivement le milieu de travail. Les études menées alors (par exemple Goodman et Sproull 1990 et Rice *et al.* 1989) sont encore particulièrement liées au milieu professionnel. En effet, premièrement, ces études se révèlent souvent financées et contrôlées par les entreprises mêmes où les données communicationnelles sont prélevées. On s'intéresse aux usages de l'ordinateur à des fins organisationnelles, pour améliorer la prise de décision en groupe, par exemple. **L'approche** apparait donc principalement **sociologique**, visant le milieu de travail. Deuxièmement, du point de vue des résultats, ces études concluent souvent à l'influence négative de la communication médiatisée par ordinateur en termes de relations humaines : la Cémo remet en cause les frontières et les hiérarchies [9]. Il faut attendre les années 90 pour que les recherches – toujours en sociologie – deviennent plus objectives : les travaux de Rheingold (1993) par exemple démontrent la possibilité pour les usagers des « communautés virtuelles » de choisir leur communauté d'appartenance avec moins d'exigences ou de discrimination que dans la vie réelle. Troisièmement, puisque ces recherches portent sur des contextes « contrôlés » dans le milieu professionnel, elles ne peuvent pas être réutilisées pour expliquer les nouvelles pratiques, lorsque les nouvelles technologies arrivent dans les foyers et permettent une communication plus « spontanée ». Comme l'expliquent George et Totsching (2001) : « on ne peut pas appliquer les mêmes procédures d'analyse aux formes de CMO actuelles, qui s'insèrent dans des sphères sociales beaucoup plus hétérogènes, tel l'espace domestique ». Aujourd'hui, la sociologie s'est adaptée à la diffusion de la Cémo dans toute la société et analyse l'impact de la Cémo sur les rapports sociaux (familiaux, intergénérationnels, amoureux, etc.) et la construction des identités (enfants et adolescents) (Devisme et Dussarps, 2010 ; Paragas, 2003 ; Rivière, 2002). Des auteurs tels que Wei *et al.* (2010), Moynihan *et al.* (2010) et Bamba et Barnes (2007) étudient les comportements humains au contact de la Cémo, par exemple par rapport à l'envoi de publicités par sms. Certains auteurs, dont Leung (2007), se sont plutôt orientés vers l'utilisation ou non des nouveaux médias par des personnes anxieuses socialement [10].

9. Carey (1993) appelle ça « border crossing in technology ».

10. Il en conclut : « People who were socially anxious and were unwilling-to-communicate face-to-face and were put off by the confusing acronyms used in mobile messaging appeared to be those who spent less time, and not more, using SMS despite the fact that SMS could help overcome student's shyness »(Leung, 2007, p. 115).

C'est à la fin des années 90 que la Cémo commence à intéresser la linguistique. Mondada (1999), Gains (1999) et Herring (1998) s'attachent à définir les traits linguistiques associés à ces nouvelles pratiques : les questions de style et d'interaction se retrouvent majoritairement au centre de leurs études. Avec l'avènement d'Internet, la communication interpersonnelle s'entremêle à la communication de groupe, voire de masse[11] : on voit alors apparaitre des formes plus hétérogènes de la Cémo, comme les forums ou les sites de « réseautage » social. Cette diversification des traits de la Cémo mène les chercheurs à réévaluer leur vision d'un ensemble de caractéristiques communes, et à y voir plutôt un ensemble de nouvelles formes de communication, montrant chacune ses caractéristiques propres. Il s'agit alors de spécifier ces caractéristiques (Fairon *et al.*, 2006b ; Frehner, 2008 ; Panckhurst, 2009), notamment le code graphique de la Cémo, en particulier la simplification graphique. Logiquement, ce processus de simplification tend à inquiéter les médias, les parents et le milieu de l'enseignement, qui s'interrogent sur l'impact des nouvelles pratiques écrites sur la maitrise de l'orthographe. Cette vague d'inquiétude entraine un nouvel élan et de nouvelles pistes de recherche qui dépassent finalement tout à fait les simples réponses aux exigences médiatiques ou pédagogiques, et qui touchent simplement à la linguistique (David et Goncalves, 2007 ; Thurlow, 2003 ; Volckaert-Legrier *et al.*, 2006). À partir de ce moment, il est question de savoir si la pratique quotidienne de la Cémo modifie profondément l'orthographe, voire la langue. Bouillaud *et al.* (2007), par exemple, ont étudié l'impact de la Cémo sur l'orthographe en général, et l'orthographe des plus jeunes dans les copies scolaires, en particulier. Leur recherche, basée sur les résultats en dictées de 3 classes de niveaux différents, laissent penser, contrairement à l'affirmation partagée concernant les effets négatifs des nouveaux médias sur la maitrise de l'orthographe, qu'il y a effectivement une corrélation entre maitrise de la Cémo et maitrise de l'orthographe, mais que cette corrélation est positive. Il s'agirait donc d'une compétence supplémentaire, une *digraphie* (Jaffré, 2010). Panckhurst (1998) s'applique quant à elle étudier l'orthographe dans le courrier électronique : elle tente de relever et classifier les types d'erreurs qui apparaissent dans le courrier électronique, notamment en les comparant aux erreurs pouvant survenir lors du traitement de texte et de l'écrit manuscrit. L'auteure prouve que les usagers peuvent facilement mélanger variation consciente et faute d'orthographe, se permettre des déviations soi-disant contrôlées par rapport à la norme, qui peuvent en réalité cacher des lacunes importantes au niveau des connaissances. Fairon *et al.* (2006c), pour leur part, sont plus prudents et concluent, à partir de l'analyse du corpus de sms issu de Fairon *et al.* (2006a) : « presque tout semble permis dans les limites imposées par l'intelligibilité, par le destinataire, des formes employées. Seules certaines tendances qualifiées de "confusions" (volontaires ou non) ont pu être dégagées. Si l'on veut se garder de toute extrapolation hâtive, **il faut donc éviter**, de façon générale, **de parler d'erreurs** ». Dans un précédent article

11. Ce que Rice (1989) appelle « a blend of interpersonal and mass media ».

que nous avons rédigé en 2010[12], nous sommes arrivée aux mêmes conclusions : dans la majorité des variantes, on ne peut conclure s'il s'agit d'une « erreur » proprement dite ou d'une stratégie propre à la Cémo, surtout lorsque la stratégie abrège les formes, ce **qui répond simplement aux exigences du médium de communication**.

Plus récemment, les études linguistiques se sont diversifiées et touchent à des sujets de recherche très hétérogènes :
- les aspects dialogiques, en se focalisant sur les forums de discussion (Rivens Mompean, 2007), des conférences en ligne (Angeli *et al.*, 2003) ou des discussions par chat (Holmer, 2008) ;
- les langues minoritaires (Berruto, 2005 ; Vold Lexander, 2007, 2009) ;
- le contrastif, en comparant par exemple les techniques d'abréviation entre deux langues (Compagnone, 2011) ;
- les fonctions de ce type de communication (Goumi *et al.*, 2011) ;
- l'alternance codique, par exemple sur les forums diasporiques (Atifi, 2007) ;
- les diacritiques (Van Compernolle, 2011), les adverbes (Guimier, 2009), etc.

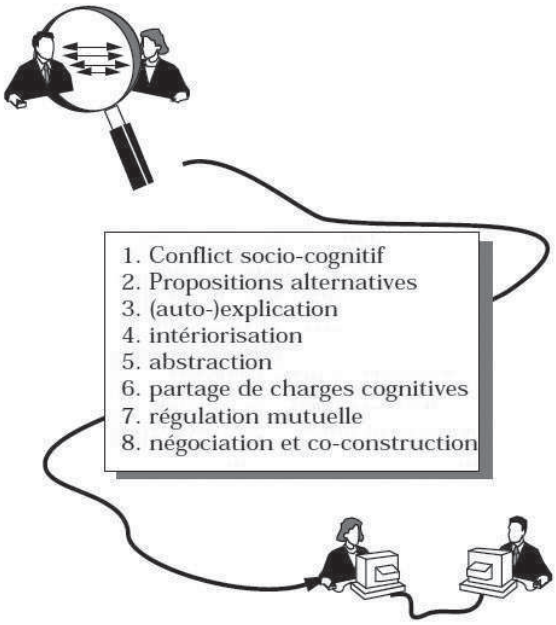
La sociolinguistique, pour sa part, étudie surtout les phénomènes linguistiques de la Cémo en fonction de l'âge (Hashimoto, 2002 ; Kasesniemi et Rautiainen, 2010 ; Laursen, 2005 ; Ling *et al.*, 2005 ; Thurlow, 2003), du sexe (Cochran *et al.*, 1990 ; Moore, 1990 ; Moyal, 1992) ou encore de l'origine sociale des usagers (Panckhurst, 1998). La sociolinguistique s'intéresse également aux langues et pratiques minoritaires. En effet, un très large courant de la sociolinguistique s'est intéressé dès la fin des années 1990 à l'étude des patois, des langues minoritaires et surtout aux pratiques écrites des sourds et muets (Adams, 2000 ; Nagarajan, 2003 ; Power et Power, 2004 ; Rubin, 2001). Guitteny (2007) par exemple s'est penché sur l'usage de la Cémo par la communauté sourde et sur la place de ce nouveau type de communication dans leur accès au savoir, dans l'échange d'information et leur accès à une plus grande autonomie.

La pédagogie, la psychologie et le traitement automatique des langues ont également investi le sujet de la Cémo. La pédagogie, tout d'abord, s'intéresse particulièrement aux effets des nouvelles pratiques communicationnelles sur l'apprentissage coopératif et l'enseignement, en particulier en vue d'améliorer les techniques pédagogiques (Abbey, 2000 ; Bourguin *et al.*, 2000 ; Crinon *et al.*, 2001 ; Curtis et Lawson, 2001 ; Mangenot et Zourou, 2007). Dans les années 1990, on a vu se développer une sous-discipline, l'Apprentissage Coopératif Assisté par Ordinateur (ACAO)[13] (cf. Figure 5.1), pour laquelle l'ordinateur est perçu tantôt comme un outil, tantôt comme un médium de communication, tantôt comme partie prenante d'un système global ; dans tous les cas, son emploi est réfléchi, ses effets, calculés.

12. Pour information, l'article est le suivant : Louise-Amélie Cougnon (2010). « Orthographe et langue dans les SMS ». Dans *Ela. Études de linguistique appliquée*. 4/2010 (n°160), p. 397-410.

13. L'ACAO trouve ses racines dans la théorie de l'activité (psychologie du développement), l'ethnométhodologie et la cognition partagée.

FIGURE 5.1 – *L'éducation collaborative médiée par ordinateur (Schneider, 1997, p. 9)*



De nombreux travaux en pédagogie (par exemple Resnyansky 2002) se posent la question de la meilleure utilisation possible de la Cémo : doit-on laisser l'apprenant libre avec la machine ou doit-il être guidé ? Comment gérer l'interaction entre l'enseignant–machine et les apprenants ? Quelle part du travail doit être prise en charge par la machine ? [14] Même le sms y est envisagé comme outil en vue d'un apprentissage plus efficace des langues (Kennedy et Levy, 2008).

La psychologie s'intéresse à la façon dont les nouvelles pratiques communicationnelles modifient totalement notre manière de lire, d'écrire et jusqu'à notre façon de penser. Certains psychologues (Bremer et Rauch, 1998 ; Brenner, 1997 ; Coulombe, 2003 ; Garneau, 1999) analysent par exemple les effets contradictoires de la Cémo, notamment le repli sur soi (cyberdépendance, etc.) et l'ouverture de soi (réduction des distances, des frontières entre les individus, effacement des différences sociales dans le monde virtuel, etc.).

En marge de l'exploration linguistique de ce nouveau langage, le traitement automatique des langues s'intéresse également de près à ce phénomène. Le sms en particulier semble constituer un sujet de prédilection pour le traitement automatique du langage car il pose de nouveaux défis, les applications TAL ayant été

14. Depuis une vingtaine d'années, de très nombreux logiciels d'apprentissage médiés par ordinateur [15], appelés collecticiels, des forums d'apprentissage collaboratif, tel que *Prodageo*. Pour plus de renseignement sur Prodageo : http://prodageo.wordpress.com/

conçues pour du texte en graphie « standardisée ». Certains chercheurs (How et Kan, 2005 ; Schneider-Hufschmidt, 2005) travaillent sur l'amélioration des systèmes de prédiction (dont nous parlerons plus en détail au sous-chapitre 5.2), d'autres, sur la normalisation des sms en graphie standard (Beaufort *et al.*, 2008 ; Guimier De Neef et Fessard, 2007 ; Kobus *et al.*, 2008). Dans la lignée de la normalisation, l'objectif de certains chercheurs est d'aller jusqu'à la vocalisation du message normalisé (Beaufort *et al.*, 2010). Parmi d'autres traitements et applications du TAL à partir de la Cémo, nous citerons les applications « agendables » qui visent à extraire automatiquement les informations temporelles de messages privés ou (semi-)publics (par exemple dans le courrier électronique, le forum ou le sms) et à le transformer en un événement inséré dans un agenda. De telles applications existent déjà depuis de nombreuses années, le challenge du TAL est de les adapter au langage de la Cémo, c'est-à-dire à une syntaxe spécifique (Weiser, 2010), à des interactions de proximité (Black et Ranjan, 2004) et à des formes lexicales modifiées (Weiser *et al.*, 2011). En outre, on rencontre encore des développements sur base de la Cémo relatifs à la classification de texte (Liu et Wang, 2010), entre autres.

5.1.1.2 Écrit sms

Les recherches pionnières en matière de sms proviennent majoritairement des **pays scandinaves** où les nouvelles technologies mobiles ont été popularisées, notamment auprès des classes d'âge les plus jeunes : à titre d'exemple, près d'un tiers des enfants et adolescents de plus de 10 ans possédait un téléphone portable en Finlande, en 2002 (cf. Kasesniemi et Rautiainen 2010). Ces études, comme nous l'avons vu pour la Cémo en général, concernaient principalement la sociologie et se focalisaient sur l' « outil sms » dans les relations entre jeunes ou l'importance de cet outil dans un processus de construction identitaire.

On observe une diffusion de l'intérêt scientifique au sujet des sms dans les années 2000. D'abord, les études se sont focalisées sur des exemples de message sans s'appuyer sur des corpus (Anis, 2001 ; Cortelazzo, 2000, 2001 ; Davis, 1991). Il existe alors une tendance à la confusion entre eSMS et autres types de la Cémo [16]. Depuis, les recherches sur le sms ont excité l'intérêt de la linguistique de corpus. Nous présentons, par ordre chronologique, un résumé concis des études principales de la linguistique de corpus au sujet des sms (cf. tableaux 5.2 et 5.3). Une liste générale des corpus de sms constitués dans le monde est présente dans Chen et Kan (2011). Nous ne reprenons pas la même liste car, d'une part, nous nous focalisons sur les études qui sont comparables à la nôtre, soit par la méthode de collecte, soit par la taille du corpus ou encore par l'approche, et, d'autre part, parce que nous nous intéressons

16. Anis (2001), par exemple, conclut de manière trop généralisante et confond souvent écrit sms et autres types de la Cémo. On remarque notamment que ses conclusions sur les sms se basent sur un corpus de type différent : « le 'langage texto' s'inspire largement du 'langage Internet' » (Anis, 2001, p. 31)

également aux études qui se basent sur le profil des auteurs des messages, sans avoir collecté de message proprement dit (ce que Chen et Kan (2011) n'abordent pas). Nous détaillons dans le tableau 5.2 quelques traits majeurs de l'étude en question (domaine, taille du ou des corpus, méthode d'obtention) ainsi que la ou les distinction(s) principale(s) que l'on entrevoit avec notre propre méthode et approche.

FIGURE 5.2 – *Principales études de corpus de sms en linguistique/sociolinguistique (tri décroissant par date)*

Référence	Domaine	Méthode de collecte	Corpus	Distinction
Goumi et al. (2011)	sociolinguistique / linguistique	méthode sms4science	1.131 sms ; 100 profils de 13 à 18	population plus ciblée; phénomènes étudiés
Compagnone (2011)	linguistique / comparatiste	SMS Monitor Studies : recopiage manuel des sms sur Internet	1.354 sms ; NC	ampleur du corpus ; approche
Shafie et al. (2010)	linguistique	recopiage manuel des sms dans la banque de données	2.640 sms ; 264 profils de 18 à 24 ans	ampleur du corpus de sms et des participants ; variation du médium de communication
Walkowska (2009)	linguistique/ TAL/ conversationnel	méthode sms4science, pas de média mais un site Internet + participation volontaire avec recopiage manuel des sms dans la banque de données	1.700 sms ; 200 profils de 15 à 40 ans	ampleur du corpus de sms et des participants ; approche
Vold Lexander (2009)	sociolinguistique	NC	NC ; NC	ampleur du corpus de sms et des participants ; méthodologie ; approche
Panckhurst (2009)	linguistique	non applicable	exemples épars de sms + corpus sms4science	pas de versant sociolinguistique/ statistique
Tagg (2009)	linguistique	entourage du chercheur + Internet (forums AOL)	11.067 sms ; 235 profils de 19 à 68	pas de versant sociolinguistique/ statistique
Bakken (2005)	sociologie/ sourds et muets	entretiens	0 sms ; 13 profils de 15 à 27 ans	pas d'étude de l'écrit sms
Hård Segerstad (2005)	sociolinguistique	sélection des sms par les participants	1.152 sms ; 20 usagers	ampleur du corpus de sms et des participants ; pas de variation entre régions
Hoeflich et Gebhardt (2005)	sociolinguistique	aucune collecte de sms	0 sms ; 367 usagers	pas d'étude de l'écrit sms
Laursen (2005)	linguistique/ conversationnel	sans sélection : accord avec compagnie de téléphonie	511 sms ; 6 usagers	ampleur du corpus de sms et des participants ; approche
Ling et al. (2005)	étude des genres/ comparatiste	collecte, sélection et transcription : NC	865 sms ; 25 usagers	ampleur du corpus de sms et des participants ; approche

FIGURE 5.3

Rivière et Licoppe (2005)	sociologie/ comparatiste	entretiens	0 sms ; 70 usagers	pas d'étude de l'écrit sms
How (2004)	linguistique/TAL	6.167 sms : origine inconnue, méthode de collecte NC. 602 sms : Yahoo sms chat. 3.348 sms : étudiants 18-25 ans avec recopiage manuel sur un site dédié	10.117 sms ; nb d'usagers flou	pas de variation entre régions ; méthodologie
Fairon et al. (2004)	linguistique	méthode sms4science	30.000 sms ; 2.436 profils de 12 à 72 ans	pas de variation entre régions
Thurlow (2003)	sociolinguistique	recopiage manuel des sms dans la banque de données	544 sms ; 135 profils d'étudiants de 19 ans	ampleur du corpus de sms et des participants ; approche

Ce tableau comparatif souligne clairement les différences entre notre étude et les études antérieures (colonne *Distinction*) :

- la taille (souvent restreinte) des corpus de sms et de profils sociodémographiques des usagers ; on distinguera les corpus avec *usagers* (qui ont participé à l'enquête mais dont on ne sait rien), des corpus avec *profils* (d'usagers, qui ont participé à l'enquête et répondu à un questionnaire) ;
- la population concernée, souvent ciblée : adolescents, étudiants, etc. ;
- les méthodes de collecte : recopiage manuel des sms, entourage du chercheur, plateforme Internet, etc. ;
- l'approche : quelquefois plus généralisante que la nôtre, souvent plus spécifique ; si quelques unes offrent un panel diversifié d'éléments de syntaxe et de lexique, seuls Goumi *et al.* (2011) abordent la variation à l'aide de méthodes statistiques (pour une population très ciblée).

5.1.2 Traits de la communication médiée

Nous avons vu au point précédent l'influence certaine du médium sur le contenu de la communication. Il semble donc tout à fait pertinent de définir les traits de la Cémo qui sont autant de variables potentielles au sein de cette influence. Herring (2001) pense que le langage est nécessairement affecté par des variables technologiques telles que la synchronie, la granularité (la longueur possible d'un texte), etc. C'est la raison pour laquelle l'auteure propose un *modèle de classification à facettes*[17] (Herring, 2007) des différents types de la Cémo en fonction de leurs

17. « faceted classification scheme ».

contextes technique et social, c'est-à-dire en fonction du médium et de la situation de communication [18]. Nous complétons ce système de classification (notamment en résumant les types de réponse possibles, cf. tableau de l'Annexe A.1, p. 308).

Ainsi, un premier lot de paramètres comprend des caractéristiques technologiques tels que les protocoles, les serveurs et les clients d'un service de Cémo, ou encore le matériel, les logiciels et les interfaces présentés aux utilisateurs. Ce n'est pas pour autant que l'auteure admet une forte influence du médium sur la communication effective, au contraire, mais elle insiste sur l'importance d'inclure toutes les caractéristiques qui permettent de définir au mieux les circonstances spécifiques qui donnent lieu à tel ou tel type de communication. Le second lot de facettes comprend quant à lui des caractéristiques sociales telles que des informations à propos des participants, le type de relation que les participants entretiennent entre eux, le but de la communication, le thème de la communication et le type de langage qu'ils utilisent.

Lorsque Herring (2007) aborde les notions de survivance, de synchronie, de taille, de structure et de format, elle aborde en réalité les notions de temps et d'espace. Ces deux notions sont capitales dans l'étude de la Cémo puisque le temps et l'espace y sont totalement remis en question, voire « transmués » [19], ou encore « déspatialisés » (Caron et Caronia, 2005, p. 27). L'adjectif « déspatialisé » est adéquat puisqu'il insiste sur le fait que temps et espace sont à la fois transformés et fusionnés. Le lieu est transformé parce que le seul espace qui prévaut est celui de l'écran (tout autre type de distance est neutralisé). Quant au temps, il doit être économisé et tendre vers l'immédiateté. La spatiotemporalité de la Cémo tend vers un « ici et maintenant ». Même les médias dits « asynchrones », dont font partie sms et courriers électroniques, répondent à ce besoin : le destinataire du message peut théoriquement le lire quelques secondes après qu'il a été envoyé. On remarquera toutefois que le caractère instantané de la Cémo ne constitue pas une nouveauté : le téléphone et le télégramme y répondaient déjà. Cependant, dans ce contexte, l'instantanéité est propre à l'oral comme à l'écrit et les nouveaux médias de communication deviennent portables, donc accessibles (potentiellement) en tous lieux (téléphone portable, ordinateur portable, baladeurs numériques, téléphones intelligents, etc.).

5.1.3 Principaux types de la Cémo

Concrètement, nous regroupons sous l'appellation Cémo [20] la communication par sms, le courrier électronique, la messagerie instantanée, le chat, le forum de discussion, le Mud (*Mud* est un acronyme pour *Multi-User Dungeons* - ou *Dimensions*

18. Il s'agit d'une méthode analytico-sémantique influencée par la grille étique hymesienne.
19. « Transmués » est à comprendre comme transformés en altérant profondément leur nature.
20. Communication écrite médiée par ordinateur.

-, qui signifie « donjons - ou dimension - à utilisateurs multiples ») et le site de réseautage social.

5.1.3.1 Messagerie instantanée et chat

Une large partie de la population et des médias, ainsi que de nombreux chercheurs ont tendance à confondre *chat* et *messagerie instantanée*. Le chat (encore écrit *tchat*) est un service de communication textuelle collective en direct via Internet. Le service pionnier dans le domaine du chat est IRC [21], conçu en 1998. C'est un système de clavardage [22] en groupe [23], chaque groupe pouvant être constitué sur la base d'une thématique ou d'un intérêt commun. Il s'agit d'un espace public, généralement un « canal » auquel on accède soit librement soit moyennant une inscription. Dans tous les cas, l'usager du chat doit choisir un pseudonyme ou « surnom » qui sera affiché dans les niches de conversation, où les commandes (générant des actions, telles que *join*, *invite* ou *disconnect*) sont à distinguer des messages textuels, objets des conversations. Les messages ne s'affichent et ne défilent pas tout à fait en temps réel sur l'écran : en effet, la saisie du message se fait dans une zone séparée et son affichage dans l'espace public ne se fait que lorsqu'il y a validation de l'intégralité du message. Les messages sont majoritairement limités en nombre de caractères (normalement entre 2000 et 5000).

La messagerie instantanée, MI, en anglais *instant messaging*, pour sa part, est un service de communication textuelle interpersonnelle. Chaque usager dialogue avec des personnes de son choix moyennant l'envoi d'une demande sur l'adresse de messagerie de l'interlocuteur et son acceptation par ce dernier. Le service pionnier qui a diffusé la pratique de la messagerie instantanée est ICQ (de l'anglais *I Seek You*, « je te cherche ») en 1996. Des services antérieurs, moins répandus, proposaient une rédaction de message en temps réel : chaque lettre apparaissait chez le destinataire dès qu'elle était tapée, ce qui signifie que les corrections et retours en arrière étaient perçus par le destinataire. Les principales caractéristiques de la messagerie instantanée et du chat sont résumées dans le tableau de l'annexe A.3, p. 311.

Malgré les différences significatives que montrent chat et messagerie instantanée, ces deux types de la Cémo partagent un nombre élevé d'éléments techniques et situationnels communs (but, canaux, délai, filtres, survivance, synchronie, taille et ton), qui permettent de les regrouper en vue de notre analyse. Parmi les auteurs qui se sont penchés sur les chat et la MI (Anis, 1999 ; Pierozak, 2003 ; Tatossian, 2008 ; Werry, 1996), nombreux sont ceux qui traitent des deux, parfois en les confondant.

21. Sigle pour *Internet Relay Chat*, « discussion relayée par Internet ». Les concepteurs d'IRC sont Jarkko Oikarinen et Darren Reed.

22. Terme québécois, mot-valise pour *clavier* et *bavardage*, d'où la signification de « bavardage via clavier ».

23. Notons que l'option d'une conversation privée est très souvent proposée, en plus des discussions en groupe.

Notons que la terminologie dans le domaine est parfois complexe ; par exemple, la dénomination *chat* désigne autant le protocole informatique, le secteur Internet, le dispositif sociotechnique, que le produit (c'est-à-dire les messages). Pierozak (2003) en particulier s'est intéressée aux phénomènes graphiques et syntaxiques du *chat*.

Les études sur la MI et le chat concernent autant leur nature (leur rapport avec la lettre, le télégramme, etc.) que leur technique (étude du produit communicatif, de l'espace informatisé délimité par l'écran, etc.). La linguistique interactionnelle s'intéresse à leurs aspects conversationnels (organisation énonciative entre émetteur et récepteur). On retrouve également des études sociolinguistiques : (Tatossian, 2008), par exemple, démontre que la réduction des digrammes et trigrammes se retrouve plus souvent (16,9 %) au sein d'une population jeune que chez les adultes (9,2 %).

Certains (Baron et Ling, 2003) rapprochent la messagerie instantanée des sms en ce que ce sont des textes composés tous deux rapidement, avec un style informel, qui réduit l'attention portée à l'orthographe, à la ponctuation et à la grammaire. Mais ce qui fait la spécificité du chat et de la MI par rapport aux autres types de la Cémo, c'est la synchronie et l'immédiateté des échanges. En effet, excepté les Muds, on ne retrouve ces deux critères dans aucun autre type. Or, ces deux critères entrainent des traits importants : contrairement aux autres types qui ne nécessitent pas la connexion simultanée des différents acteurs, la communication chat et MI ne peut avoir lieu à tout moment et entraine donc une moins grande spontanéité dans le démarrage de la conversation. A contrario, l'immédiateté des réponses accroit la spontanéité des thématiques et des réactions. Ces deux lots de caractéristiques entrainent des conséquences tant au niveau des rapports sociaux, qu'au niveau de la langue.

5.1.3.2 Courrier électronique

Le courrier électronique (dorénavant *courriel*), encore nommé e-mail, email, mél, texte électronique (Lòpez Alonso, 2003) ou courriel au Québec, est le plus ancien type de la Cémo et il reste aujourd'hui un des plus utilisés. Il s'agit d'un système d'échanges de messages entre usagers disposant d'une boite aux lettres électronique. Lòpez Alonso (2003, p. 109) propose une définition linguistique : « le courrier peut être défini comme une correspondance individuelle ou collective, réelle ou virtuelle, d'échange de textes digitalisés avec une forme, une fonction, un contenu, une structure et un langage propres ». Les messages sont stockés soit en ligne, sur un serveur, soit sur le disque dur d'un ordinateur. L'histoire du courriel débute en 1965 mais son usage est alors limité à deux personnes liées à un même serveur [24]. Cet usage s'intensifie avec l'arrivée d'ARPAnet (le futur Internet [25]) en 1969 ; sa

24. Les deux ordinateurs étaient situés à l'Université de Los Angeles et à l'Université de Stanford.
25. L'Advanced Research Project Agency se développe dans le cadre d'un projet de recherche mis sur pied par le Ministère de la défense américain, en pleine guerre froide. Il est donc initialement

diffusion véritable, avec une pratique attestée en dehors du milieu professionnel ou universitaire, date du début des années 90. En outre, le courriel se présente comme la forme dominante d'utilisation d'Internet [26]. Le tableau de l'annexe A.4, p. 312, reprend les caractéristiques principales du courriel.

Le courriel, initialement très proche du format et du style de la lettre « traditionnelle », s'en écarte à présent par des variétés stylistiques et registrales en fonction du rapport destinateur – destinataire, du ton du message, des pratiques personnelles du destinateur. Ainsi, on trouvera dans le courriel un continuum allant d'une lettre digitalisée, respectant les titres, formules de politesse et autres conventions, à un message constitué de quelques mots, sans ponctuation, accent, ni signature, proche du style du sms ou du chat [27].

Panckhurst (1998, 1999), Baron (1998) et Baron et Ling (2003) se sont particulièrement penchés sur l'identification des traits linguistiques et sociaux du courriel : contrairement aux autres types de la Cémo, il y a une importance du péritexte (date, destinataires multiples, objet, etc.). Il s'agit d'échanges plutôt lents (délai plus long que dans les autres types) mais quel que soit le message initial, la réponse, s'il y en a, est toujours attendue dans un délai relativement court. L'importance de ce délai est prouvée, puisque s'il est dépassé (après un jour, une semaine ou un moins), s'installe chez le destinateur ce que Panckhurst (1998) nomme « une angoisse de la communication ».

Outre les recherches sur les traits intrinsèques du courriel, les études linguistiques concernent deux traits du courriel en particulier : tout d'abord, les chercheurs s'intéressent particulièrement au genre auquel appartient le courriel (Cusin-Berche, 1999 ; Marcoccia, 2004a), puisque, comme nous venons de l'expliquer brièvement, le courriel montre des pratiques de communication pouvant l'assimiler tantôt au courrier traditionnel, tantôt aux formes de communication modernes plus brèves. Ensuite, le courriel intéresse les spécialistes de l'interaction (Herring, 1998 ; Marcoccia, 2005a), qui remarquent notamment une évolution dans les échanges, qu'elle soit linguistique ou sociolinguistique, notamment dans les relations professionnelles. Enfin, Maurais (2003), dans son étude sur 4.003 courriels majoritairement québé-

prévu pour un usage militaire mais se propage assez rapidement dans les milieux scientifiques et universitaires.

26. En effet, dans l'Union européenne (27 pays), en 2009, 57 % des particuliers âgés de 16 à 74 ans utilisaient Internet pour envoyer/recevoir des messages électroniques, contre 51 % pour rechercher des informations sur des biens et services, 32 % pour une consultation dans un but d'apprentissage, 31 % pour lire/télécharger des journaux/magazines d'information en ligne, 28 % pour télécharger/écouter/regarder de la musique et/ou des films et 15 % pour rechercher un emploi. Données issues de Eurostat, site de statistique européenne de la Commission européenne (http ://epp.eurostat.ec.europa.eu). Notons que les chiffres concernant l'accès aux films et à la musique sur Internet datent de 2008.

27. Ces messages sont usuellement le fait d'usagers du courriel qui envoient un grand nombre de messages quotidiennement ou qui disposent d'un système de courriel sur leur téléphone portable, doté d'un plus petit clavier et disponible dans des contextes d'utilisation plus « pressés ».

cois, comprenant au total 405 714 mots, a pour objectif l'analyse de la qualité de la langue, en fonction, notamment, de l'âge et du sexe des auteurs des courriels.

Enfin, le courriel a fait l'objet de nombreuses critiques, par exemple par Melançon, qui voit dans le courriel une déshumanisation des échanges : il « est immatériel, il n'occupe aucun lieu véritable, alors que la lettre traditionnelle n'a d'existence que concrète et offerte à l'autre [...]. Sur le plan temporel, ce message [...] peut être reproduit *ad infinitum* s'il a été archivé, cela tant à l'émission qu'à la réception » (Melançon, 1996). Notons qu'il s'agit d'une opposition valable pour tout objet numérique par rapport à son équivalent analogique (photographie, album musical, etc.) et qui n'enlève en rien l'existence dudit objet et les échanges interpersonnels possibles de cet objet.

5.1.3.3 Forums de discussion

Les forums de discussion[28], en anglais *Internet forums* ou *message boards*, sont apparus à la fin des années 90. Il s'agit de sites publics, sortes de plateformes interactives accessibles via une adresse web, sur lesquels les visiteurs peuvent contribuer à des « sujets de discussion » en envoyant des messages souvent après approbation d'un modérateur, plus généralement après une inscription plus détaillée. Les forums de discussions concernent souvent des thématiques (l'écologie, les jeux vidéos, etc.) ou des publics précis (jeunes parents, adolescents « emo », etc.). Les forums peuvent aussi être des outils intégrés à des sites web plus larges ; le forum constitue alors le lieu d'échange d'informations ou d'opinions sur des thématiques proches du site en question (par exemple le forum du site du journal Le Monde aborde les questions d'actualité).

Quel que soit le type, un forum de discussion se présente généralement comme une page web sur laquelle apparait d'abord une liste de thématiques (des sous-catégories de sujets plus précis peuvent exister à l'intérieur de chaque thématique). Une fois la thématique choisie, se dresse la liste de toutes les contributions existantes. Les contributions sont hiérarchisées : un premier message lance un thème et d'autres y répondent. La contribution d'un Internaute est parfois différée car modérée avant d'être affichée sur la page, parfois instantanée. Les forums de discussion qui, comme nous l'avons expliqué, se présentent comme une page web, accessible par un simple navigateur, sont à distinguer des serveurs de nouvelles (ou groupes de nouvelles), en anglais *newsgroups* apparus dans les années 80[29]. Il s'agit également de listes de messages publics accessibles à tous et classés selon des thématiques. À la différence des forums, pour consulter les messages, il faut utiliser un logiciel spécifique au serveur de nouvelles en question ou s'inscrire pour une réception des messages par courrier électronique. Le tableau de l'annexe A.5, p. 314 présente les principales

28. On notera le pléonasme de l'expression, *forum* désignant déjà « lieu de débats publics ».
29. On citera par exemple *Usenet* fondé en 1979 par des étudiants de Caroline du Nord au États-Unis.

caractéristiques du forum de discussion.

Une importante série de travaux utilise les forums de discussion comme corpus de pratiques écrites pour les recherches conversationnelles : dans le monde francophone, c'est principalement Marcoccia (2004b, 2006) qui s'y intéresse. Le sujet syntaxique de cet écrit « spontané » intéresse particulièrement le monde scientifique (Mondada, 1999 ; Von Münchow, 2004). Puisque les forums traitent de sujets précis, la terminologie et la lexicologie (Desquinabo, 2008 ; Fabrot, 2001) se sont également penchées sur les mots en usage dans tel ou tel domaine, principalement en politique. On notera également une lignée de recherches en sociolinguistique (Kling, 1996 ; Marcoccia, 2006), les forums se présentant comme de véritables réservoirs à tons, registres, styles et personnalités diverses.

5.1.3.4 Sites de réseautage social

Les *sites de réseautage social* (SRSs), ou *Social Network(ing) Sites* (SNSs) en anglais, sont des services Internet permettant à un individu de créer un profil (semi-) public, d'établir des liens avec d'autres individus (ou « connexions »), inscrits dans le même système, avec qui il partage un lien et de parcourir le profil de ces connexions, ainsi que leur propre liste de connexions[30]. Il s'agit d'un service de plus en plus populaire, surtout auprès des adolescents, mais pas uniquement. La raison pour laquelle le phénomène des réseaux sociaux, qui ne montre pas de fonction première de « communication par messages », mais bien de bâtir des « relations sociales », nous intéresse particulièrement, réside dans l'évolution des services que proposent ces sites : en effet, ils encouragent désormais, dans la majeure partie des cas (excepté quelques cas limites), des échanges conversationnels et montrent en moyenne une pratique de l'écrit quotidienne. La Figure 5.4 livre un panorama des médias sociaux en 2012 dans lequel s'inscrivent deux portions qui nous intéressent parce qu'elles mettent en jeu davantage d'écrit quotidien.
Il nous a paru intéressant d'étudier ce type de Cémo parce qu'il mélange des traits de chaque autre type et, comme nous allons le détailler, parce qu'il varie en fonction de la variété de SRS elle-même, ce qui permet de nous poser la question : existe-t-il un type d'écrit SRS ?

Selon la définition précédente, le premier SRS est apparu en 1997 sous le nom de

30. Cette définition conduit à considérer les réseaux sociaux comme des espaces à nombre de connexions infini. Cette considération est d'ailleurs à l'origine même de la notion. C'est notamment Milgram qui, en 1967, a initié les recherches sur ce sujet, dans le cadre du « phénomène du petit monde » (en anglais small world phenomenon), encore appelé « paradoxe de Milgram », dont l'hypothèse était que chaque individu pouvait être relié à n'importe quel autre individu par une courte chaine de relations sociales, et qui a prouvé que tous les individus étaient reliés entre eux par une chaine de six degrés de séparation. Pour plus d'information à ce sujet, voir Chung et Lu (2004). Notons encore que selon une étude de 2011 réalisée par le réseau Facebook, il n'existait qu'un degré de séparation de 4,74 (http://www.rue89.com/2011/11/22/oubliez-les-6-degres-de-separation-il-y-en-474-226785).

FIGURE 5.4 – *Panorama des médias sociaux en 2012 (Cavazza, 2012)*



SixDegrees.com : il permettait à l'inscrit de créer un profil, de lister ses amis et de parcourir la liste de ses amis. La deuxième vague importante de SRSs fut initiée par *Ryze.com*, en 2001, le premier service de réseautage dont le but était de tirer profit du réseau professionnel de chaque inscrit (*Ryze.com* était principalement focalisé sur le monde des affaires). À l'heure actuelle, il existe des centaines de SRSs, montrant des services, des pratiques et des intérêts différents. Si on met de côté les réseaux professionnels tels que *LinkedIn*, *Ryze*, *Tribe*, *Academia.edu*, *MySpace*, etc. où la norme écrite est davantage de rigueur [31] puisque les champs sont visibles par les acteurs du monde professionnel de l'inscrit, les autres réseaux fonctionnent plus ou moins sur le mode de la communication rapide, courte et émotionnelle. Chacun de ses traits varie en fonction du site, du contexte et du rapport entre l'inscrit et ses relations ; en fonction du site, puisque l'espace dont dispose l'inscrit dans chaque champ est très différent d'un site à l'autre : *Twitter*, par exemple, ne permet la saisie que de 140 caractères pour remplir le champ du statut, alors que *Facebook* en offre 420 et 1000 pour les messages « postés » sur le profil de l'utilisateur. Cette contrainte d'espace régit déjà, à elle seule, des usages différents d'un site à l'autre ; en fonction du contexte de vie, de la situation de communication, puisque

31. Si l'on ne prend pas en compte les interactions des visiteurs qui réagissent au contenu des pages, par exemple sur les forums de *MySpace*.

un commentaire déclaratif et sérieux sera rédigé de façon beaucoup plus normée qu'un commentaire humoristique ou hargneux, par exemple ; enfin, en fonction du rapport de l'inscrit avec chacune de ses relations, car les inscrits peuvent choisir dans leur réseau des individus proches uniquement, ou au contraire donner accès à une population plus vaste, comprenant des membres du réseau professionnel, de la famille, etc [32].

La plupart des SRSs offrent **deux champs** principaux [33] pour l'écrit : un champ de « statut » de l'inscrit au site lui-même (régulièrement mis à jour) et/ou des champs de « commentaires » sur les statuts, profils (encore appelés « murs »), activités et photos de ses relations. On voit donc aussi apparaitre deux variétés de contenu de messages et de pratiques : des **messages différés**, en réaction à des commentaires, photos, etc. d'autres individus et des **statuts**, ou « humeurs » du quotidien, postés plus régulièrement [34]. Les statuts semblent plus proche du type d'écrit sms : en une poignée (limitée) de mots, l'usager transmet une information succincte à un destinataire collectif. Les messages différés, souvent des commentaires, quelquefois très fournis, liés à des images ou renvoyant à des sites Internet en dehors du SRS, partagent davantage les caractéristiques du forum de discussion. Le tableau de l'annexe A.6, p. 315 reprend les traits principaux des SRSs.

Le but des SRS est de proposer une *vitrine* pour chaque individu. D'un point de vue linguistique ou sociologique, le type d'écrit des SRS est particulièrement intéressant puisque le « paraitre » devient primordial et la langue en est (entre autres choses [35]) l'outil. Pour les adolescents cela peut se manifester par une tendance plus nette à l'emploi de smiley, de langue familière, par une recherche plus subtile d'effets de style, etc. qui auront tendance à écarter leurs écrits de la norme. Des SRSs tels que *Facebook* et *Twitter*, laissent place à la variation, à l'écrit spontané, voire, dans certains cas à des pratiques proches de l'écriture automatique [36]. Au contraire, sur les sites professionnels, l'inscrit s'efforcera de respecter une norme qui le définit – au moins – comme compétent à l'écrit. Un SRS tel que *MySpace*, encourageant le partage de données artistiques (principalement la musique), montre des usages également différents : en fonction de l'inscrit (un groupe de musique professionnel ou amateur, un simple usager, un acteur de l'industrie du disque, etc.) et du type d'information qu'il diffuse sur sa page, les écrits varieront sur un continuum (normé – non normé) mais qui ne laissera pas la place à variation.

32. Certains politiciens par exemple ouvrent leur réseau aux adhérents de leur parti, aux journalistes, etc. autant d'individus qui vont « normer » leur façon de communiquer sur leur SRS.

33. Si on laisse de côté les services de messagerie privée inclus dans les SRSs qui coïncident plus avec du courrier électronique.

34. On voit même apparaitre des sites, tel que *Twitter*, dont la fonction est d'encourager la communication en temps réel.

35. Notons toutefois que les messages semblent constituer l'outil principalde communication, en termes de pratiques réelles, dans les réseaux sociaux (voir annexe 3).

36. En ce qu'elle est alors rapide et sans contrôle, qu'il soit grammatical, lexical, esthétique ou moral.

Pour ce qui est de la recherche scientifique, puisque les SRSs collectent autant des informations sociologiques que des données textuelles, de nombreuses études sociolinguistiques (en fonction de l'âge, de l'origine ethnique, des intérêts, etc.) sont possibles ; pourtant, comparés aux autres types de la Cémo, les SRSs ne font pas encore l'objet de nombreuses recherches, surtout dans le monde scientifique francophone. C'est étonnamment la didactique qui s'y est le plus intéressée : on citera par exemple Perez (2009) et McBride (2009). Dans le monde anglo-saxon, les chercheurs abordent une perspective majoritairement sociologique (Boyd et Ellison, 2007) ou TAL (Thelwall *et al.*, 2009). Ce dernier point nous permet de revenir à la question initiale : existe-t-il un type d'écrit SRS ? Nous avons montré qu'il existe une diversité étonnante dans les types de SRSs ; les messages peuvent également varier : statut, commentaires, etc. ; enfin, les relations entre les individus d'un même réseau montrent aussi des variations importantes sur un continuum allant de la proximité amicale à la distance professionnelle.

Contrairement à d'autres types de Cémo, les SRS ne semblent pas former un tout cohérent. En effet, alors que le courrier électronique, le forum, le chat et le sms montrent une variation de style ou de registre uniquement, dans les SRS, le seul trait commun est la volonté d'initier, d'entretenir ou de renforcer un réseau social (professionnel, amical, etc.). Pour le reste, on remarque que ni l'espace de communication, ni les canaux, ni le format, ni la synchronicité, ni le délai de réponse, ni la survivance du texte ne se manifestent de manière homogène. Au sein des SRS, il existe donc une variation de registre, de traits de communication et de style. Les **SRS ne constituent pas une variété bien délimitée de la Cémo** avec ses propres caractéristiques conversationnelles, textuelles, etc.

5.1.3.5 Muds ou communication dans les univers virtuels

Les *Muds* constituent un contexte propice à un type moins connu de la Cémo, pourtant ils occupent une place prépondérante dans la communication globale des internautes, surtout des adeptes des « jeux de rôle en réseau », depuis 1978. Comme nous l'avons dit précédemment, *mud* est un acronyme pour *Multi-User Dungeons* (ou *Dimensions*), qui signifie « donjons (ou dimension) à utilisateurs multiples » [37]. Il s'agit de serveurs sur Internet permettant à des internautes de dialoguer entre eux a travers un « monde » où ils créent et/ou collaborent à des activités en commun (le plus souvent un jeu), pour accomplir une tâche, un but, une « quête ». Les participants aux Muds incarnent les plus souvent des héros fantastiques qui évoluent dans un monde avec lequel ils peuvent interagir. Le tableau de l'annexe A.7, p. 318 résume les principales caractéristiques des MUDs.

Les Muds ne doivent pas être confondus avec les *Moos*, apparus en 1990, acronyme

37. L'utilisation de « donjon » dans l'appellation provient d'un jeu de plateau, *Donjons et dragons*, qui propose un univers fantastique dans lequel les personnages doivent accomplir des tâches.

pour *Muds Object-Oriented*, « Muds orientés objet » qui sont une variété spécifique de Muds [38].

Contrairement aux autres types de jeux (jeux de plateau, jeux sur console individuelle), ces plateformes créent des aires de communication où les joueurs interagissent, communiquent entre eux, notamment afin de mener à bien leur mission dans le jeu [39]. D'ailleurs, à l'époque où cette technologie ne permettait que des échanges textuels, les personnages, leurs attributs et l'environnement étaient décrits avec des mots. De nos jours, la communication interpersonnelle dans les Muds consiste en des ordres ou des demandes d'aide par rapport au jeu lui-même ou bien en des conversations « ordinaires ». En pratique, cela ressemble à une activité de chat qui a lieu dans un « univers » défini.

La pertinence de ce type de la Cémo pour la recherche en linguistique réside dans le fait que, quel que soit le but de la communication, l'attention est portée sur une tout autre activité que l'acte langagier : ceci étant, la communication est très peu surveillée et, tout comme le sms, le Mud offre un espace de communication limité (le sms en espace, le Mud en temps) qui oblige à écrire le plus rapidement et le plus abrégé possible (notamment pour que la lecture soit plus rapide).

Avec le temps, les Muds se sont diversifiés et il ne s'agit plus toujours de participer à une aventure dans un univers fantastique : la mode s'est par exemple répandue aux univers de la science-fiction. Quant à l'aventure, elle devient aussi pédagogique (on citera par exemple *Moose Crossing* [40]).

Bien que les Muds existent depuis une trentaine d'années, c'est seulement depuis vingt ans que l'intérêt des chercheurs pour cet objet spécifique s'est manifesté. On trouve des études sur les enjeux pédagogiques (Reid, 1994 ; Woodruff et Waldorf, 1995) des Muds, mais c'est principalement l'impact psychologique et social des Muds qui fait l'objet de très nombreuses recherches : ainsi Polin (1993) et Turkle (1992) se sont penchés sur les aspects sociaux, les rapports entre les membres d'une même communauté Mud, alors que Heaton et Lafrance (1994) ainsi que Bruckman (1992) ont analysé la construction des identités. Ces recherches psychosociales ont prin-

38. Les Moos sont des programmes de jeux en réseau à utilisateurs multiples où l'on peut construire un monde totalement virtuel avec un personnage totalement factice que l'on partage avec les autres personnages. L'identité de chaque personnage est très détaillée dès l'inscription. Au sein du monde virtuel, les personnages vivent, se côtoient et parlent les uns avec les autres, ce qui donne lieu à de nombreuses productions écrites et/ou orales, selon le service. Selon Meunier et Peraya (2004, p. 422), le Moo est « un *chat sophistiqué* possédant une topologie de lieux, des avatars et des objets permanents ». Le but du Moo n'est pas forcément d'atteindre un certain objectif, mais simplement de « vivre une autre vie » dans un monde parallèle à la réalité. Un exemple de Moo connu est *LambdaMOO*.

39. L'annexe 4 montre un exemple d'écran de Mud avec les différentes parties dédiées à du texte

40. Moose Crossing est un Mud d'apprentissage pour enfants de 8 à 13 ans. Pour en savoir plus : http://www.cc.gatech.edu/elc/moose-crossing/aboutus.html.

cipalement exploré le comportement des joueurs, la façon dont ils se supportent, s'encouragent, punissent les tricheurs ou encore de quelle manière ils gèrent la neutralité des genres : en effet, un personnage fictif peut être masculin ou féminin, le sexe du personnage ne respectant pas pour autant celui du joueur, les rapports entre individus de sexe différent sont neutralisés. Quant aux études linguistiques, elles sont beaucoup plus rares et anglo-saxonnes ; on citera par exemple Carlstrom (1992) qui s'intéresse aux anciennes versions de Muds, basées uniquement sur du texte, et Cherny (1995a,b) qui se focalise sur les Muds en tant que lieux de conversations.

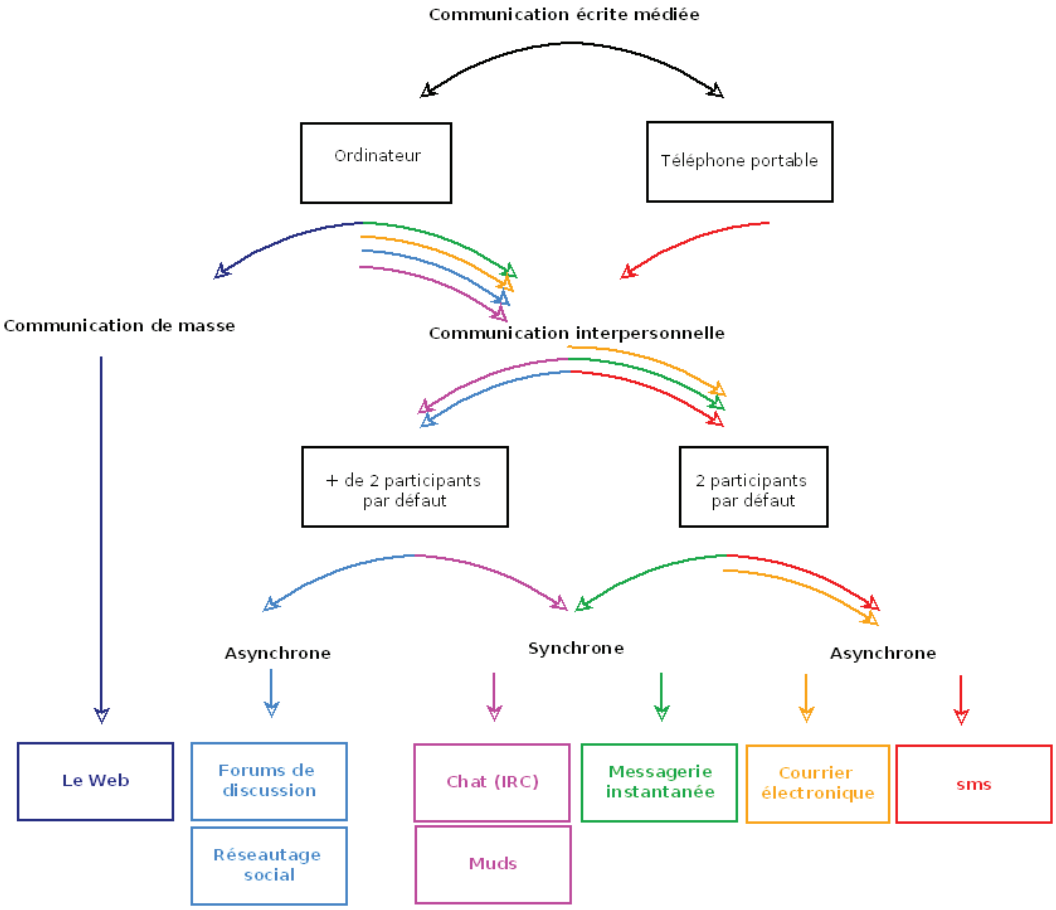
Conclusions

La présentation des types de la Cémo que nous venons de réaliser fait apparaitre des points communs et des divergences. Pour ce qui est des points communs, nous pensons qu'ils constituent la plupart du temps un moyen d'écrit *éphémère*. Comme l'explique Panckhurst (1998), la majeure partie du temps, le message est rédigé rapidement par le destinateur, rarement (ou sommairement) relu, vite envoyé, reçu par le destinataire et promptement lu et détruit. La comparaison des tableaux caractérisant chaque type de la Cémo nous mène également à penser que les traits principaux de la communication médiée, ceux qui permettent la plus courte et pertinente distinction entre les types, sont : - le support (ordinateur, téléphone portable) ; - le but de la communication (de masse, interpersonnelle) ; - les acteurs (paire par défaut, multiacteurs) ; - et la synchronicité. Afin d'illustrer au mieux cette distinction minimale, nous améliorons le schéma de Marcoccia (2005b, p. 5).

Après une analyse détaillée des traits de chaque type de la communication médiée, il apparait aussi clairement que nait une certaine *convergence médiatique*, encore appelée *convergence numérique* (souvent décrite sous ses aspects techniques), c'est-à-dire une fusion des services qu'offrent les différents médias de communication. Caron et Caronia (2005, p. 32) appellent cela une tendance à l'*environnement multitâche*. Il s'agit d'un phénomène ancien, mais qui s'est intensifié ces dernières années grâce à la numérisation des contenus. Il est désormais possible d'envoyer des courriers électroniques et des messages instantanés avec un téléphone portable (notamment avec les modèles de téléphone dit « intelligents », ou *smartphones*), de chater à partir d'hébergeurs de courriers électroniques (par exemple *Gmail*), de téléphoner via des services Internet (par exemple *VoIP*, pour *Voice over IP*, « la voix sur réseau IP »), etc.

Face à cette fusion, nous sommes en droit de nous poser la question suivante : qu'est-ce que ces fusions impliquent par rapport à l'évolution de la communication ? Nous savons que la convergence se fait via la fusion des trois éléments qui forment le numérique : l'information (texte, photo, vidéo, etc.), son support (hardware, software et protocole) et son transport (émission, transmission). L'évolution du langage va donc se faire en fonction de ces trois éléments : par exemple, là où du

FIGURE 5.5 – *Classement de la communication écrite médiée en fonction de ses traits principaux*



texte peut être remplacé par une image ou une vidéo, l'usager de la Cémo n'aura plus autant recours aux mots pour communiquer. En outre, certaines fonctionnalités comme la possibilité d'envoyer du courrier électronique à la place des sms, à l'aide d'un écran azerty, allègeront les contraintes d'écriture, par exemple la limite en nombre de caractères, ce qui influencera indéniablement les usages. Ces hypothèses sont évidemment à corroborer grâce à de futures recherches.

Tant que cette fusion n'est pas aboutie, est-ce qu'il existe des variétés bien délimitées de la Cémo ? Nous répondons à cette question par l'affirmative et ce même, avant d'avoir exploité la notion d'eSMS. Au vu des caractéristiques intrinsèques de chaque type que nous venons d'étudier, **il existe une pratique écrite inhérente à chaque type** ; cette pratique varie au sein d'un type du point de vue du style et du registre. Seuls les SRS ne semblent pas pouvoir être inclus dans cette liste. Ils partagent des caractéristiques tantôt d'un type, tantôt d'un autre, mais ne possèdent pas des traits fixes.

5.2 L'écrit sms et ses phénomènes graphiques

The dream of wireless communication among people has long been around and is a fundamental desire of the human being.
Frehner (2008)

Sommaire

5.1 Communication, médiation et informatisation 18
5.1.1 Un état de l'art 20
5.1.2 Traits de la communication médiée 27
5.1.3 Principaux types de la Cémo 28
5.2 L'écrit sms et ses phénomènes graphiques 41
5.2.1 Un nouveau code graphique 42
5.2.2 Phénomènes graphiques et lexicaux 46
5.2.3 Un écrit oralisé, spontané ou simplement un langage ordinaire ? 56
5.2.4 Conclusions 60
5.3 Conclusions 61

Introduction

Un nombre restreint de termes équivalents à sms est employé dans la littérature : *texto* et *mini-message*[41] sont ceux qui apparaissent le plus souvent. On ne peut plus nier l'omniprésence – et parfois même la nécessité – de la communication par sms. À titre indicatif, 2.300 milliards de sms et mms ont été échangés en 2008 à travers le monde[42]. Selon le rapport Mediappro (2006) basé sur les pratiques de 9 pays européens[43], 95 % des jeunes de 12 à 18 ans possèdent un téléphone portable ; parmi eux, 79 % préfèrent envoyer un sms qu'initier un appel téléphonique.

Comme explicité précédemment, nous préférerons la notion d'*écrit* ou d'*écriture sms* (Panckhurst, 1997)[44], (abrégé *eSMS*), à celle de langage sms. En effet, il ne s'agit pas d'un nouveau langage, ni d'une nouvelle langue ; il s'agit d'une nouvelle pratique écrite possédant ses propres caractéristiques, une nouvelle manière de transcrire la langue : nous aurions ainsi un « code sms » qui utilise, tout comme

41. Le Petit Robert propose, pour texto, une définition semblable à celle de sms, à ceci près qu'il ne reprend pas le sens de « service ». En ce sens, texto serait donc moins ambigu. Mini-message n'apparait pas dans le dictionnaire ; en outre, il est principalement employé en France.

42. Ces données proviennent du Journal du Net. Source : http://www.journaldunet.com/chiffres-cles.shtml

43. Ces 9 pays sont : la Belgique, le Danemark, l'Estonie, la France, la Grèce, l'Italie, la Pologne, le Portugal et le Royaume-Uni.

44. Cette formule est également préférée par d'autres linguistes, tels que Marie-José Béguelin, qui l'a énoncée lors d'un reportage radiophonique pour la chaine Arcinfo. Source : http://www.arcinfo.ch/videos/actualites/video/6960/le_langage_sms_sous_la_loupe_des_linguistes.html

notre langue écrite standard, les lettres de l'alphabet latin et les chiffres arabes, en les employant de manière traditionnelle ou en leur attribuant une valeur différente (phonétisante, symbolique, etc.) ; mais il s'agit également d'une nouvelle pratique comportant ses propres espaces, objectifs, thématiques, etc. En cela, on ne peut la réduire à un code ; c'est pour mettre l'emphase sur le résultat d'une pratique, que nous avons choisi de la définir en tant qu'*écrit sms*.

Qu'est-ce que le sms [45], texto, mini-message, etc. [46] ? Selon Anis (2002), il « s'agit d'une communication interindividuelle entre des partenaires qui se connaissent préalablement et ont un certain niveau d'intimité. Le régime temporel est le différé, mais à échéance rapide : une quasi-immédiateté est visée pour la lecture du message et une réponse est attendue dans de brefs délais ».

Comparé aux autres type de la Cémo, le sms n'a engendré que très tard un intérêt dans le monde scientifique. L'usage du sms n'a d'ailleurs commencé à s'intensifier significativement qu'à la fin des années 90. Afin de cerner au mieux notre sujet d'étude, nous allons proposer dans cette partie un panorama des caractéristiques de l'eSMS. Ensuite, puisque le noyau de notre recherche portera sur les phénomènes lexicaux et syntaxiques, nous aborderons seulement dans cette présentation les divers phénomènes (néo-)graphiques récurrents dans les sms, propres à la Cémo en général, ou aux sms en particulier. Enfin, nous tenterons de qualifier cet écrit selon la terminologie traditionnelle et notre propre raisonnement.

5.2.1 Un nouveau code graphique

Nous aborderons ici la notion de code graphique en ce qu'elle nous parait primordiale pour approcher la communication médiée par ordinateur, un nouveau canal qui modifie le code écrit standard. En effet, et comme l'a expliqué Klinkenberg (1996) : « quand le transcodage a pour fonction de transposer un code d'un canal à un autre, ce code acquiert des caractéristiques nouvelles, qui lui sont fournies par le nouveau canal que ces signes empruntent ». Pourtant, la variation n'est pas toujours d'actualité dans les sms car, comme le montrent Caron et Caronia (2005, p. 187) : « comme une vraie langue, l'écriture en sms prévoit un haut degré de standardisation et, en même temps, elle permet des variations ».

Avant de définir précisément ce code, et pour mieux le cerner, nous allons résumer en premier lieu les différents genres de la communication écrite plus « traditionnelle » de type épistolaire [47]. Abordons tout d'abord la *lettre* : elle existe depuis l'Antiquité

45. Ou Short Message System. Herring (2007, p. 7) le décrit comme étant du « text messaging through mobile phones ».

46. Les Anglo-Saxons utilisent le plus souvent les termes text et texting.

47. Cette partie renvoie aux recherches réalisées par Labbé et Marcoccia (2005), Cusin-Berche (1999), Kerbrat-Orecchioni (1998) et Marcoccia (2005b)

gréco-latine mais elle est, jusqu'au XIXe siècle au moins, le fait d'une minorité lettrée et/ou aisée ; après quoi, elle se diffuse lentement au sein de la population. La lettre demande des formules d'adresse et autres conventions de forme, telles que la date, le lieu, l'adresse du destinateur et du destinataire, la signature manuscrite, etc., dont la disposition sur la page est tout aussi réglementée. L'application de ces règles varie en fonction du destinataire et de la situation de communication, mais certaines règles, telles que la formule d'adresse et la date, se maintiennent dans la plupart des lettres. De la lettre, l'eSMS a surtout hérité des formules d'adresse : comme nous le verrons dans le corps de notre recherche, les salutations qui marquent l'ouverture et la fermeture du canal de communication sont présentes dans les sms de notre corpus.

Une autre forme de correspondance est le *billet* : le billet est intéressant en ce qu'il constitue une version simplifiée de l'échange épistolaire. Il s'agit d'un type d'écrit dont la forme est abrégée et dont le contenu est plus familier que la lettre, tout en gardant une expression assez soignée. Le billet est apparu au XVIe siècle, lorsque les épistoliers ont éprouvé le besoin d'une forme d'échange plus brève, en évitant les formules de politesse traditionnelles souvent pour faire passer un message à valeur informative. Le billet répond donc à un besoin de souplesse dans une société très conventionnelle. Le billet est en outre un exemple pertinent car il permet à l'épistolier de briller par sa capacité à faire un jeu de mot tout en respectant l'exigence de brièveté. Deux traits du billet sont maintenues dans l'eSMS : il doit se soumettre aux exigences de brièveté et cette exigence de brièveté donne lieu à des jeux, des remaniements ludiques du code.

Le *télégramme*, mis au point en 1832 par Samuel Morse, propose une communication où l'exigence de brièveté est encore plus marquée ; à cela s'ajoutent d'autres changements au niveau de la syntaxe, qui vont également servir à abréger le message : absence de sujets, de prépositions, de mots-liens, etc. Enfin, la ponctuation y est remplacée par un « stop » à valeur unique : une pause dans l'énoncé. Avec le télégramme, on assiste aux débuts du principe d'économie dans la communication ; comme l'explique Jaffré (2003, p. 43), l'écriture a toujours « tendance à privilégier des associations aussi économiques que possible entre des segments phoniques et des segments graphiques. Et cela, comme si la nécessité cognitive du code l'emportait sur d'autres contraintes, qu'elles soient sociales, culturelles ou linguistiques ». Ces deux derniers traits (modification de la syntaxe ainsi que neutralisation de la ponctuation) rapprochent particulièrement le télégramme de l'eSMS : en effet, une part importante est faite, dans les deux cas, à l'interprétation du lecteur. Ce qui nous mène à livrer un quatrième trait du code sms : son caractère très ambigu pour lequel le rôle de désambigüisation par le lecteur-destinataire (en fonction du contexte et de ses propres connaissances)[48] est plus important que lors des échanges écrits traditionnels.

48. Nous reviendrons plus en profondeur sur cette notion de désambiguïsation dans la partie dévolue à l'usage de régionalismes (lexicaux et syntaxiques), notamment.

Enfin, la *carte postale*, un carton portant le message à transmettre ainsi que l'adresse du destinataire, sans enveloppe, apparait en Autriche en 1869 et s'étend rapidement à l'Europe tout entière. Au départ, totalement textuelle, la carte postale se pare ensuite d'une illustration et propose, comme le billet, un cadre moins formel pour un contenu plus familier. Au contraire du billet, la carte devient vite très populaire, car disponible pour une modique somme. À la différence du billet également, l'exigence de brièveté n'est pas influencée par la volonté de l'homme, mais imposée par le support de communication. C'est ce dernier trait qui clôt notre définition du code sms par la comparaison avec l'écrit traditionnel.

Par rapport à ces plus anciennes formes de communication, la pratique du sms apparait donc comme héritière et singulière à la fois. Elle partage avec la pratique de la lettre ses formules d'adresse, mais elle diffère de la lettre en de nombreux points : le sms se libère totalement des conventions d'usage qu'impose la correspondance traditionnelle, il se joue du temps étendu de rédaction et de relecture. Pour ce qui est du billet, on a constaté qu'il partage avec l'eSMS l'utilisation d'un code plus abrégé, le recours à des jeux stylistiques et graphiques ainsi qu'une visée informative et pratique. Par contre, son exigence de brièveté n'est pas imposée par le médium de communication dans le cas du billet, mais par son contexte d'utilisation par les usagers. L'eSMS partage avec le télégramme une syntaxe et une ponctuation simplifiées (ce que l'on appelle couramment un « style télégraphique »), mais s'en distingue par son usage beaucoup plus répandu ainsi que l'absence de marqueurs interphrastiques (les « stops »). Enfin, tout comme la carte postale, le sms présente la caractéristique d'être employé par une très large population, indifféremment de l'âge et de l'origine socioculturelle. Notons enfin que c'est la rapidité des échanges qui distingue l'eSMS de toutes les plus anciennes formes : chaque lettre, billet ou carte postale demande plusieurs jours de transmission. Le télégramme est en réalité beaucoup plus rapide (quelques minutes), mais il n'est pas réceptionné directement par le destinataire, contrairement au sms.

Pour résumer, grâce à cette première étape de recentrage du sujet par comparaison avec les plus anciennes formes, nous remarquons, comme le souligne Anis (2002), l'émergence d'un nouveau code qui se caractérise par :

- l'immédiateté des échanges, ce qui entraine une rapidité de réaction, et donc un style brut (c'est-à-dire sans relecture) ;
- une quasi absence d'exigences formelles (conventions), qui amène à un style plus familier (alors qu'on associe habituellement écrit et formalisme) ;
- l'exigence d'un médium limité, qui oblige le scripteur à faire appel au ludique (s'exprimant par la néographie, le jeu de mot, le smiley).

Les différences de contexte d'utilisation, les variations lexicales, morphologiques et syntaxiques qui ont lieu dans les sms proposent donc une autre manière de considérer l'écrit ; on aboutit presque à deux codes différents, avec leurs règles

propres. Cette pratique, parallèlement à la pratique de l'écrit traditionnel (en milieu professionnel, scolaire, etc.) nous fait penser qu'il existe une **double compétence** inhérente aux usagers du sms qui manient ces deux codes.

Identifions à présent l'eSMS dans le concert international des autres types de la Cémo, en recourant, comme nous l'avons fait précédemment aux caractéristiques fondamentales selon Herring (2007) (cf. Tableau de l'annexe A.8, p. 319).

À la lecture de ce tableau, l'eSMS semble se démarquer des autres types de la Cémo par trois de ces principaux traits : l'**asynchronie**, **la limite de 160 caractères** [49] et **la norme d'organisation** (c'est l'unique service de la Cémo qui ne demande pas d'inscription préalable et qui ne filtre pas les messages). À cela, il faut ajouter une quatrième et une cinquième caractéristiques fondamentale qui ne sont pas explicitées dans le tableau : la **mobilité** et le **système d'encodage**. L'eSMS permet la mobilité : il est en réalité le seul type de la Cémo [50] qui est pratiqué quels que soient l'endroit, la situation, étant donné qu'il ne nécessite qu'un téléphone mobile pour être employé [51]. L'eSMS dépend du système d'encodage : l'encodage des lettre dans un sms peut se faire de deux façons : soit par la technique des pressions multiples (en anglais *multi-tap*) soit par un système de prédiction de texte (en anglais *predictive texte entry*).

FIGURE 5.6 – *Exemple de clavier numérique d'un téléphone portable traditionnel*



La technique des pressions multiples repose sur un principe simple : chaque touche du clavier numérique standard [52] peut représenter, outre un chiffre, une série de lettres

49. Ces 160 caractères dépassent l'affichage de la plupart des écrans de téléphone portable, ainsi il est très fréquent qu'il faille faire défiler de haut en bas le message pour en apercevoir la totalité.

50. Ce trait tend évidemment à disparaitre avec l'arrivée massive des Smartphones, ou « téléphones intelligents », qui permettent l'envoi de courrier électronique à l'aide du téléphone mobile. Ils nécessitent toutefois une connexion à Internet qui est souvent incluse, mais pas toujours, dans les abonnements des opérateurs de téléphonie.

51. Notons que cette spécificité de l'eSMS est en voie de disparition étant donné la mobilité des ordinateurs et la disponibilité croissante des connexions au réseau internet, y compris à partir d'un téléphone portable.

52. Les sms peuvent également être envoyés à partir d'un ordinateur, mais la grande majorité est effectivement envoyée par téléphone mobile. On ajoutera à cette remarque que les nouveaux téléphones « intelligents » offrent un clavier tactile qui ressemble au clavier d'un ordinateur (même

et de signes. Pour encoder la lettre ou le symbole dont il a besoin lors de la rédaction d'un message, l'usager du sms doit presser la touche où apparait cette lettre ou ce signe une ou plusieurs fois, en fonction de la position qu'occupe cette lettre ou ce signe sur (ou plutôt derrière[53]) la touche en question[54]. La deuxième technique d'encodage des caractères dans un sms est le système de prédiction de texte : ce système nécessite une technologie linguistique supplémentaire. L'usager presse une seule fois sur la touche derrière laquelle se trouve la lettre qu'il désire encoder ; au fur et à mesure qu'il entre les différentes séries de lettres qui vont constituer son mot, le système va essayer d'identifier le mot, sur base d'un dictionnaire de mots installé dans le téléphone et de la combinaison la plus probable de ces séries de lettre. Le système va ensuite proposer à l'usager les différentes combinaisons de lettres formant un mot[55] ; l'usager n'aura plus qu'à choisir parmi les différentes propositions. MacKenzie et Soukoreff (2002) nomment encore cette technique « one-key with disambiguation ». Un exemple connu de ce type de système est le T9 lancé par l'entreprise Tegic Communications, Inc. Le désavantage d'un système de prédiction de texte est qu'il ne reconnait aucune forme s'écartant de la norme, telle que l'argot, les mots étrangers et surtout l'abréviation graphique des mots. Connaitre le mode d'encodage des caractères d'un message permet ainsi de l'étudier avec plus de précision et d'évaluer des phénomènes que l'on pourrait interpréter différemment si un seul système d'encodage était utilisé.

5.2.2 Phénomènes graphiques et lexicaux

Les phénomènes graphiques[56] et lexicaux que nous allons aborder dans cette partie se regroupent en quatre catégories principales : abréviations graphiques avec ou sans incidence phonique, ponctuation expressive et blancs. Ces phénomènes revêtent plusieurs fonctions ; nous empruntons à Klinkenberg (1996) les fonctions en question, à partir de ses catégories concernant l'écrit en général.

Premièrement, les fonctions graphémologiques visent à « représenter la langue »

si les touches sont effectivement plus petites) en ce qu'ils permettent un mode de saisie semblable à celui du clavier d'ordinateur. Ceci pourrait effectivement jouer sur la pratique de l'eSMS si ces téléphones étaient employés par une large majorité de la population. Ceci n'est pas valable pour les nouveaux types de téléphones présentant un clavier du même type que celui présent sur les ordinateurs traditionnels.

53. En effet, certaines lettres (comme les lettres accentuées) et les signes de ponctuation ne sont pas affichés tels quels sur les touches du téléphone mais sont toutefois encodées derrière une des touches et révélées uniquement lors des pressions multiples.

54. Dans notre exemple de clavier de la Figure 13, la lettre « k » peut être révélée en pressant deux fois la touche 5.

55. Selon le service, l'ordre de ces propositions de mots dépend d'une fréquence des mots les plus utilisés, soit préétablie, soit établie en fonction des pratiques personnelles de l'usager.

56. Certains des phénomènes que nous allons traiter sont valables pour le français uniquement. En effet, des phénomènes tels que la neutralisation des e muets finaux se retrouvent en tant que tels dans un corpus anglais, mais où on retrouvera aussi couramment la neutralisation du g final (ex. *darlin* pour *darling*).

(Klinkenberg, 1996, p. 47) ; on compte parmi ces phénomènes [57] :

- les signes phonographiques [58], qui renvoient à des phonèmes et des syllabes représentés par des lettres, des chiffres ou des signes. Nous comptons également dans cette catégorie les squelettes consonantiques ;
- les signes logographiques, qui sont des graphèmes qui renvoient à des signifiés linguistiques. Un exemple pertinent de cette fonction est le x final qui signifie « bisou ».

Deuxièmement, les fonctions grammatologiques (terme dérivé du grec γραμμα, « la lettre, le caractère ») concernent l'espace qu'occupe l'écriture ; on dénombre quatre techniques parmi ces fonctions dans l'écrit sms (seule la fonction indicielle étant exclue de cette pratique) :

- la technique indexicale, lorsque certaines règles conventionnelles permettent de lier un signe à un objet contigu au signe. Il nous est permis de penser que certains émoticônes, par exemple, participent de cette fonction. En effet, lorsque l'émoticône suit une proposition ou un message dont il altère totalement, ou du moins modifie, le ton et souvent donc le sens, sa fonction peut alors être indexicale ;
- la technique taxonomique, lorsque les signes sont disposés dans un **ordre** différent de celui attendu. Dans le contexte du sms, on s'attend à un message continu. Rares sont les phénomènes dérivés de la technique taxonomique. Une fonction taxonomique apparait par exemple lorsque des (groupes de) signes sont séparés par des retours à la ligne pour dresser des listes. Ce phénomène est plutôt rare dans l'écrit sms, mais n'est pas inexistant ;
- la technique symbolique, lorsque le choix de la typographie ajoute une valeur au contenu des signes graphémologiques. C'est par exemple le cas avec l'usage de majuscules pour signifier que la lettre doit être prononcée de manière alphabétique ou en criant (mais non de manière régulière) ;
- la technique iconique, lorsque le tracé du signe ou d'une série de signes renvoie figurativement à un objet ou à un autre signe. Dans le cadre du code SMS, ce phénomène se retrouve aussi bien dans les dessins ASCII que dans l'emploi de certains signes tels que @ pour signifier « oreille ».

Avant d'aborder ces phénomènes en détail, rappelons encore que certains de ces phénomènes apparaissent dans les autres types de la Cémo, d'autres dans tous les types de la Cémo, mais que nous les appliquerons ici aux seuls contexte et usage du sms.

57. L'eSMS ne montre pas d'exemple des deux autres fonctions graphémologiques, c'est-à-dire la fonction morphosyntaxique et la fonction thématique.

58. C'est d'ailleurs une des fonctions primordiales de l'écriture car, comme le rappelle Jaffré (2006) : « sans phonographie, l'écriture n'existerait pas, du moins sous la forme que nous lui connaissons ».

5.2.2.1 Abréviations graphiques

À la suite de Eco (2002), nous pensons vivre « à une époque où le diminutif, le bref et le simple ont une valeur inestimable dans la communication [59] » [60]. Cette tendance à l'abréviation s'illustre dans la plupart des types de la Cémo. En effet, l'exigence de rapidité que propose le contexte du sms (accorder à l'encodage d'un message quelques minutes est particulièrement contraignant d'un point de vue pratique dans le quotidien d'un usager [61]), ajoutée à la difficulté d'encodage à partir d'un clavier [62] (pour les non initiés surtout), ne semble pas constituer une spécificité de l'écrit sms. On retrouve la première caractéristique dans le courrier électronique et les MUDs, quant à la seconde, elle est évidemment de rigueur dans tous les types de la Cémo, excepté lorsque les usagers encodent leur message via un procédé de reconnaissance vocale et non via un clavier. Ces deux traits encouragent un procédé d'écriture bref, au moins en temps, sinon également en nombre de caractères. Pourtant, les sms montrent deux caractéristiques supplémentaires qui participent aussi au phénomène d'abréviation : la contrainte d'espace et de cout. La contrainte d'espace existe dans la plupart des autres types de Cémo (réseautage social, messagerie instantanée, etc.) mais elle ne se profile pas comme une véritable contrainte dans la majorité des cas, puisque cette limitation, comme nous l'avons vu, est moins stricte [63], les usagers ne la dépassant pratiquement jamais ; le sms pour sa part, avec ses 160 caractères par message, propose une réelle limite d'espace qui contraint souvent à réduire le nombre de caractères d'un message. Licoppe (2002, p. 196) explique : « leurs contraintes de taille font qu'ils ne peuvent en effet prêter à de longs développements, et ils vont d'emblée constituer une ressource orientée vers le répertoire des usages en mode "connecté" ». La contrainte de cout quant à elle n'existe dans aucun autre type de Cémo, si l'on ne prend pas en compte le cout du forfait Internet des usagers (excepté lorsque les messages sont envoyés via un téléphone portable). En ce sens, il s'agit bien d'un trait spécifique du sms, mais variable d'une région à l'autre [64].

Les quatre traits que nous venons d'énoncer (rapidité d'échange, difficulté d'en-

59. « We live in a an age where the diminutive, the brief and the simple are highly prized in communication » (notre traduction).

60. Relativisons cette tendance dite « contemporaine » que nous interprétons simplement comme une intensification de l'application de « la loi du moindre effort », comme la nomme Bauche (1920).

61. De plus, comme le fait remarquer Véronis et Guimier De Neef (2006), théoriquement, « les sms autorisent un temps d'écriture plus long [que le chat] [mais ils sont] très souvent utilisés sous une forme d'échange rapide qui les place dans le même cas que les chats ».

62. Notons que si pratiquement tous les Cémo nécessitent le recours à un clavier, seul le sms montre un clavier si petit, et donc encore plus contraignant.

63. Excepté dans l'espace dédié au « statut » pour le réseautage social (souvent de l'ordre de 140 caractères).

64. Le cout d'envoi d'un sms entre la Belgique ou la France (en moyenne 0,12 euros en 2012) et le Canada (en moyenne 0,10 euros en 2012), par exemple, ne semble pas varier énormément ; c'est bien le système d'abonnement qui varie : les Français et les Canadiens fonctionnent davantage avec l'achat de packs de sms, avec des abonnements sms illimités, etc. qu'en Belgique, où cette procédure est rendue possible mais peu utilisée.

codage due au clavier, limite d'espace et cout) encouragent particulièrement l'abréviation, la « simplification » dans les sms, qu'elle soit syntaxique, lexicale ou morphologique. Le concept de simplification est à comprendre selon l'optique empruntée par Gadet (1991) ; « plus une langue accumule de traits relevant de ces trois caractères plus elle tend à être simple » : économie (réduction, absence, invariabilité...), analycité (schéma syntaxique fixe, relation invariante entre forme et fonction...) et motivation.

Pour catégoriser les phénomènes qui touchent à l'abréviation, nous nous baserons sur la typologie énoncée par Tatossian (2008) et par Berruto (2005), que nous complétons et adaptons au contexte particulier du sms.

Abréviations graphiques sans incidence phonique Premièrement, nous regroupons sous l'appellation *abréviations graphiques sans incidence phonique*[65] les phénomènes suivants[66] :

1. L'apocope (*poss* pour *possible*) :
   (a) L'apocope peut entrainer ce que Panckhurst appelle une *abréviation sémantisée*, c'est-à-dire une abréviation réduite à l'initiale, une initiale qui n'est alors pas lue de manière alphabétique mais en fonction du sens et de sa place dans la proposition. Ainsi une lettre *t* peut tantôt remplacer *tu* (dans *t peux venir*), tantôt remplacer *te* (dans *je t donne*)[67] ;
   (b) L'apocope peut également user de la ponctuation et être suivie d'un point qui indique la chute. Ainsi *auj.* et *dem.* calqués sur les modèles plus connus *etc.*, *env.*, *ex.*, etc.
   (c) L'apocope peut se réaliser sous la forme d'une neutralisation de graphèmes muets en finale absolue (*ils port* pour *ils portent*) ; Anis (2002) nomme ces pratiques des *mutogrammes*[68].
2. L'aphérèse (*lut* pour *salut*) ;
3. La syncope (*aple* pour *appelle*). Deux cas spécifiques de la syncope sont :
   (a) La simplification des consonnes doubles (*come* pour *comme*) ;
   (b) l'effacement du schwa à l'intérieur du mot (*ramner* pour *ramener*).

La multiplication des syncopes ramène quelquefois l'écriture d'un mot à ce qu'on appelle couramment un ***squelette consonantique*** ; il s'agit de « l'abréviation d'un mot commun charpenté quasi-exclusivement autour de ses consonnes » (Prochasson

65. Les abrègements lexicaux sont courants dans l'eSMS, par ex. *anniv* pour *anniversaire*. Ils sont pourtant non spécifiques de l'eSMS, raison pour laquelle ils ne sont pas repris dans cette partie.

66. Nous illustrons chaque phénomène entre parenthèses.

67. Dans ce cas elle répond à une logique plus complexe, qui fait jouer à la fois la logique morphologique et phonétique du scripteur et du lecteur, comme c'est le cas pour les autres phénomènes de la phonétisation, mais aussi syntaxique et sémantique.

68. Sautot (2000) : « Sous le terme de 'mutogramme' seront rangés les graphèmes dont la valeur phonique est égale à zéro [...] : des lettres muettes donc. »

et al., 2007). Comme l'explique Anis (2002) : « on sait depuis longtemps grâce à la théorie de l'information que les consonnes ont une valeur informative plus forte que les voyelles [...] Les consonnes retenues comportent toujours la première et la dernière ». Les consonnes médianes doubles peuvent d'ailleurs être simplifiées sans altérer la reconnaissance du mot.

Les squelettes consonantiques sont de rigueur dans les écritures hébraïque et arabe : en effet, dans de nombreux contextes (notamment officiels ou journalistiques), seules les consonnes apparaissent.

FIGURE 5.7 – *Racine arabe KTB sans voyellation*



La Figure 5.7 montre une racine arabe KTB sans voyellation (ce qui signifie que les voyelles courtes manquent). Cette racine pourrait autant se référer à *kutub* (« des livres ») qu'à *kataba* (« écrire »). C'est le contexte qui désambigüise la graphie. En réalité, le processus de reconnaissance des écritures sémitiques en fonction de la racine et du contexte est plus complexe que le processus de reconnaissance de l'écriture française à partir des squelettes consonantiques dans les sms : premièrement, tous les mots dans l'eSMS ne sont pas touchés par les syncopes multiples et certains mots sont plus concernés (*bjr*, *bsr*, *stdrd*, etc.) que d'autres ; enfin, les mots d'un même message sont rarement tous réduits à leur squelette, alors que dans un texte hébraïque ou arabe, c'est tout à fait le cas.

Deuxièmement, nous regroupons sous l'étiquette ***spécialisation de caractères*** :

1. La siglaison[69] (*tlm* pour *tout le monde*, *asv* pour *âge sexe ville*, *irl* pour *in the real life*). Cette siglaison peut devenir extrême et représenter toute une phrase. C'est ce que Caron et Caronia (2005) appellent l'*acrostiche*[70] ;
2. L'acronymie (*lol* pour *laughing out loud*).

Abréviations graphiques avec incidence phonique L'ensemble des *abréviations graphiques avec incidence phonique* comprend ce qui a trait à la ***phonétisation de caractères***[71], c'est-à-dire toute modification qui n'altère pas la prononciation de l'unité :

69. Nous n'avons pas inclus siglaison et acronymie dans les simplifications morphologiques, bien qu'il s'agisse d'une suite d'apocopes réduites à l'initiale, puisque ce phénomène dépasse le cadre du mot, pour en représenter plusieurs.

70. Les sms italiens montrent plus d'exemples de ce type que les francophones : tvtrb pour ti voglio troppo bene.

71. Notons que la combinaison des phénomènes suivants est souvent appelée rébus dans la littérature (voir par exemple Dejond 2006). Nous avons rejeté cette appellation parce qu'elle ne

1. Phonétisation [72] par la lettre : substitution d'une lettre à un son avec la valeur sonore qu'on confère à la lettre lorsqu'on la prononce (*r* pour *air*). Un cas spécial de cette phonétisation est la réduction de digrammes et trigrammes [73] (*sel* pour *seul*, *mè* pour *mais*). Panckhurst précise que ce phénomène de réduction, qu'elle nomme plus ou moins *substitution*, peut être entier (lorsque l'orthographe du lexème est totalement modifiée, ex. *o* pour *eau*) ou partiel (l'orthographe du lexème étant partiellement modifiée, ex. *bo* pour beau) ;
2. Phonétisation par le chiffre : substitution d'un chiffre à un son avec la valeur sonore qu'on confère à ce chiffre lorsqu'on le prononce (*9* pour *neuf*) ; Anis (2002) explique à ce propos : « quand les chiffres sont utilisés pour transcrire un mot entier, on est proche du logogramme ou signe-mot » ;
3. Phonétisation par le signe : substitution d'un signe à un son avec la valeur sonore qu'on confère à ce signe lorsqu'on le prononce (*pl@* pour *plate*) ;
4. Phonétisation par des graphies supposées plus proches du phonétisme réel, plus proche d'un accent régional ou d'un style particulier (/ca/, /co/, /cu/, /qu/, /k/ par /k/) [74].

Nous ajouterons un dernier phénomène qui joue considérablement sur la morphologie en contexte sms : la gestion des ***blancs graphiques***, encore appelés *espaces blanches*. En effet, l'écrit sms montre deux tendances :

1. Une tendance à supprimer les blancs : on parle alors de soudure. Anis (2002) les nomme « réductions avec compactage, lequel dissout les frontières de mots et évoque le mot phonique » (ex. *vérépa* pour *verrai pas*). Deux méthodes prévalent : recourir aux lettres capitales pour marquer la séparation entre les mots ou supposer des capacités de décodage et de reconnaissance lexicale suffisantes chez le destinataire. On remarquera que certaines langues telles que le grec ancien n'utilisaient que très peu de blancs graphiques, et que cette tendance du sms vers une *scriptio continua* n'a rien de nouveau, d'original, même l'Antiquité latine a longtemps procédé de la sorte [75]. Certes, comme l'explique Pombo (2001, p. 41-42), « La *scriptura continua* [...] exigeait davantage de temps de déchiffrement que ne nous l'exige aujourd'hui notre écriture ; en outre, elle demandait une lecture vocalisée, une prononciation à haute voix », mais le destinateur du sms fait prévaloir les habitudes linguistiques inhérentes

correspond pas exactement à la réalité du phénomène : on préférera à celle-ci combinaison de phonétisation.

72. Anis (2000, p. 60) la nomme « graphie à connotation oralisante ».

73. Il s'agit de la succession de deux ou trois graphèmes représentant un phonème unique.

74. Ou encore la substitutionle remplacement des variantes /voy + sa/, /voy + se/, /voy + si/ en /z/. Un cas très récurrent : *biz* pour *bise*.

75. « Certains textes théoriques latins traitent les problèmes de lecture que cette scriptio continua entrainait. La lecture fluide d'un texte jamais lu auparavant était difficile, car le lecteur devait d'abord déchiffrer l'écrit avant de procéder à la réalisation vocale des énoncés. Cette prononciation était assistée par le souvenir du texte, lorsque le lecteur le connaissait déjà. Les Latins suivaient un entrainement à la lecture rapide ; cet exercice faisait partie des enseignements du grammairien » (Pombo, 2001, p. 41-42).

à la communication par sms – et pour lesquels il sait implicitement son interlocuteur compétent – sur les contraintes que pose la lecture de son message ;

2. Une tendance à ramener les signes de ponctuations, tels que l'apostrophe et le trait d'union, à des blancs graphiques (*est ce que* pour *est-ce que*) [76]. C'est également le cas avec la ponctuation syntaxique (virgules, points, double points, points-virgules, etc.).

5.2.2.2 Emoticônes

Un émoticône [77] consiste en une combinaison de caractères, le plus souvent des signes de ponctuation, permettant de représenter, si on les incline à 90 degrés dans le sens des aiguilles d'une montre, des expressions faciales, des émotions, telles que des éclats de rire, des clins d'œil, des moues de tristesse, etc. Le concept même du petit visage souriant, initialement jaune, date de l'époque des hippies de San Francisco, dont l'émoticône était la mascotte ; il a aussi été réutilisé, plus récemment, en tant que badge représentant le style de musique *house*. Par conséquent, au moment où les informaticiens ont voulu désambiguïser le ton de leurs premiers messages échangés sur des serveurs privés [78], comme le dit Dejond (2006, p. 28), lorsqu'ils ont voulu « matérialiser les ressentis », ils ont immédiatement pensé au petit visage et à la manière la plus pratique de le schématiser à l'aide d'un clavier.

Comment caractériser la nature même des émoticônes ? Nous retenons de Marcoccia (2004a, p. 2), ces quelques caractéristiques : les émoticônes sont en quelque sorte des pictogrammes (utilisation iconique de caractères graphiques) ; « on peut aussi noter que la combinaison de signes graphiques pour former un symbole figuratif de mimique faciale repose sur un principe comparable à la combinaison de kinèmes en kinémorphèmes, telle qu'elle est exposée dans la théorie de Birdswhistell (voir Cosnier and Brossard 1984). Selon Mourlhon-Dallies et Colin (1999) cette fois, les émotiônes font partie « des éléments méta-énonciatifs qui s'apparentent aux *didascalies* du théâtre » [79], tout comme l'étirement graphique et l'usage de lettres capitales. Ce sont des « sortes d'indices de l'énonciation intégrés à l'énoncé qui s'en trouvent en quelque sorte modalisé » (Pierozak, 2007 : 76).

La position de l'émoticône au sein de l'énoncé n'est pas aléatoire : il se situe généralement à la suite d'un mot, en fin de phrase ou de message et peut revêtir alors plusieurs fonctions. Les fonctions de l'émoticône sont inspirées par Wilson (1993) ou Mourlhon-Dallies & Colin (1995), tel que l'a fait Marcoccia (2000). Un émoticône peut revêtir une fonction [80] :

76. De la même manière, nous noterons que les signes diacritiques sont très souvent neutralisés (voila pour voilà, ca pour ça et element pour élément).
77. Encore appelés *émoticons*, *smileys* (de l'anglais *to smile*) ou *binettes*. Le terme émoticône est un mot-valise créé à partir des mots émotion et icône.
78. Voir le sous-chapitre sur la messagerie instantanée et le chat.
79. didascalies
80. Ces quatre fonctions ne sont pas exclusives et aucun des divers types d'émoticônes n'est figé

1. *Expressive*, en vue d'exposer l'état d'esprit du destinateur (tristesse, moquerie, colère, tendresse, etc.) ;
2. *Interprétative*, en vue d'apporter une aide au destinataire qui comprendra mieux le sms reçu. L'émoticône en forme de clin d'œil permet par exemple de marquer l'ironie d'un énoncé qui ne doit pas être pris au premier degré. C'est donc en quelque sorte une fonction de désambiguïsation, ce que Cosnier appelle la « mimogestualité connotative », qui participe à l'interprétation des énoncés ;
3. *Relationnelle*, en vue d'indiquer au destinateur la relation qu'il désire instaurer avec lui, en vue de synchroniser les émotions des deux parties ;
4. *De politesse*, un *softener* dans la théorie de Brown et Levinson (1987), un moyen d'adoucir le caractère offensant d'un message. En effet, la Cémo montre très souvent des messages dont l'interprétation sera plus offensante que le sens réel, puisque proposant moins de contexte ; l'émoticône permet de préciser ou réorienter le ton du message.

L'émoticône peut également revêtir une fonction *lexicale* et prendre la place d'une entité au sein de la phrase, représentant un nom, un adjectif, etc. Deux exemples de contextes d'utilisation illustrent parfaitement cette tendance.

- Adjectif (:-) pour *content*) :

(2) Ma petite crevette, ça m'a fait plaisir de t'entendre, et suis super :-) que tu puisses passer 2 minutes à la maison. [smsFB]

- Substantif : (:-) pour *sourire*) :

(3) Alors ? Vous êtes toujours à Han ? C'est bien ? Tu as deviné mon :-) ? [smsFB]

Comme les unités linguistiques, la forme des émoticônes est variable ; nous pouvons dresser une typologie des émoticônes selon trois émotions récurrentes :

1. Le marquage d'humeur : signifiant sourire, contentement, rire, fous-rire, ou, à l'opposé, mécontentement, tristesse, malheur, déception, etc. qui sont représentés par :-) , :o) , :-D , :(, :-(, :,-([81] ;
2. Le marquage de complicité représenté par ;) , ;-) , ;o) ;
3. Le marquage d'ironie représenté par :p ou :P .

Ces types d'émoticônes différents peuvent être appréhendés selon leur fonction dans le message ; Pierozak (2007 : 82) a cherché à classer les smileys les plus fréquents sur un continuum allant de la coénonciation à la colocution[82]. Le résumé de son

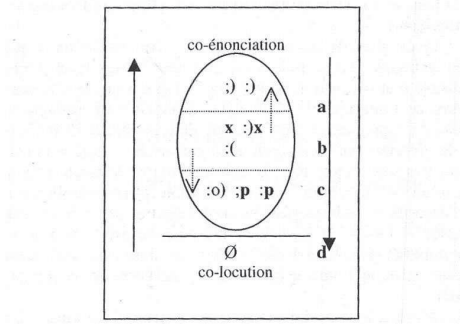
dans une fonction précise.

81. On ne notera en exemple que les émoticônes les plus fréquents.

82. Bouvet et Morel (2002 : 12-13) définissent la coénonciation (« La façon dont celui qui parle (le parleur) envisage la réception de son discours par celui auquel il s'adresse (l'écouteur) [...] L'articulation à la pensée prêtée à l'autre et l'anticipation de ses réactions relèvent de la coénonciation ») et la colocution (« La prise en compte du droit à la parole de chacune des parties prenantes du dialogue et l'anticipation d'une éventuelle prise de parole du colocuteur (écouteur) sont gérées par les variations d'intensité. Le colocuteur n'est pas la personne physique à qui l'on parle, mais la façon dont on envisage son droit à parler. »).

approche est illustré dans la Figure 5.8.

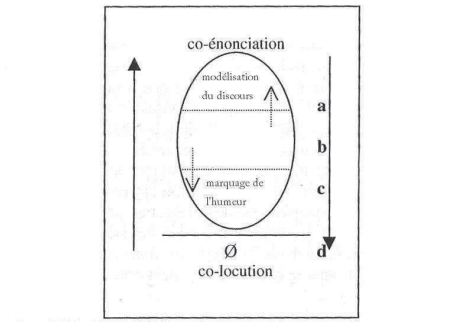
FIGURE 5.8 – *Les émoticônes sous l'angle de la coénonciation/colocution selon Pierozak*



Comme l'expose assez clairement cette figure, aucun émoticône n'est, pour Pierozak, purement colocutoire ; tous les émoticônes revêtissent une fonction coénonciative, à différents degrés.

À côté de cette classification émique[83], nous proposons une classification fonctionnelle globale ; comme nous l'avons brièvement énoncé plus haut, nous distinguons 4 grandes catégories fonctionnelles que nous regroupons en **deux fonctions principales** : les émoticônes qui **modélisent une lexie**, une proposition ou un message, qui y ajoutent une tonalité ou du sens (fonctions interprétative et de politesse) et les émoticônes qui **expriment l'humeur** du destinateur et sa relation avec le destinataire, sans altérer d'aucune façon le message en lui-même (fonction relationnelle et expressive). Bien que les deux catégories se situent en effet dans la coénonciation (le destinateur prend en compte la perception de son message par le destinataire), la première est toutefois plus proche de l'acte coénonciatif pur que la seconde. Selon cette approche, aucun smiley n'appartient en lui-même à une des deux catégories, c'est son utilisation en contexte qui définit sa fonction. Ainsi, nous proposerons une adaptation du schéma de Pierozak en fonction de nos catégories (cf. Figure 5.9).

FIGURE 5.9 – *Les émoticônes sous l'angle de la coénonciation/colocution selon nous*



83. « Qui s'appuie sur les concepts et le système de pensée propres aux acteurs sociaux étudiés » (Aymes *et al.*, 2005)

5.2.2.3 Ponctuation et casse expressives, caractères échos

Le contexte du sms, au même titre que d'autres types de la Cémo, montre des emplois particuliers des signes de ponctuation ; nous avons vu précédemment qu'ils pouvaient servir à reproduire une expression faciale dans les émoticônes. Ils peuvent également revêtir une fonction expressive plus intense que dans un contexte écrit traditionnel [84], ajoutant ainsi à l'écrit une information paralinguistique. C'est le cas lorsqu'il y a démultiplication d'un même signe. Nous avons vu lors du sous-chapitre 5.2.2.1 que dans l'écrit de la Cémo la tendance était à ramener la ponctuation syntaxique [85] à des blancs graphiques ; on remarque également un sous-emploi de la ponctuation à fonction rythmique. Les scripteurs se concentrent uniquement sur la fonction expressive et intonative (points d'exclamation, d'interrogation, points de suspension, etc.) [86]. Comme l'explique Marcoccia (2004a, p. 2) : « le redoublement du point d'exclamation est un procédé très utilisé dans la [Cémo] ; il repose sur des conventions d'écriture habituellement utilisées dans la mise en scène de l'oral à l'écrit (dans la littérature, par exemple) ». Ce redoublement sert à apporter de l'expressivité, de l'intensité, mais également du son. On notera également dans ce contexte un suremploi des points de suspension, tel que le montre l'exemple suivant :

(4) Malgré le fait que tu n'es pas gentil avec moi (!!!!!), je t'aime de tout mon cœur mon bébé... et tu me manques chaque minute où tu n'es pas près de moi... [smsFB]

Il s'agit d'un phénomène d'extension (de duplication, de répétition) graphique que l'on trouvait déjà à l'époque des premiers modes de communication médiée, tels que le Minitel en France, comme le montrent les travaux de Anis (1999) ou Luzzati (1991). L'extension ne concerne pas uniquement la ponctuation et touche aussi aux caractères latins : à la suite de Tatossian (2008), nous nommons ce phénomène ***caractères écho***. Ce phénomène consiste en une répétition de caractères à valeur d'expressivité, d'intensité, mais également en vue d'apporter du son (comme lors du redoublement de ponctuation). C'est souvent le cas avec les interjections (*oooooh !*, *mdrrrrrr* ou *nooooooon*), mais nous pensons pouvoir également inclure dans cette catégorie les cas de variantes orthographiques de type bisoo au lieu de bisou. En effet, le but de la variante s'avère le même que celui cité précédemment : simuler un accent d'insistance [87].

(5) J'ai réussiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii (bon, ça suffit, là) !!!! Bisous. Si tu veux, on va à la quinzaine de la bière ce soir [smsFB]

84. Par exemple dans les plus anciennes formes de communication écrite.

85. Veronis et Guimier de Neef (2006) livrent à ce sujet les chiffres suivants : seuls 40 % des sms montrent un point final. Nous nous posons toutefois la question de la pertinence de tels chiffres, notamment en raison de la présence de signatures finales (par exemple à l'aide d'hypocoristique de type : *Ta doudou*) dans les sms. On ne ponctue généralement pas après une signature ; or, ces auteurs ne précisent pas dans l'article s'ils ont pris ce phénomène en considération.

86. La terminologie concernant les fonctions provient de la Banque de dépannage linguistique de l'Office québécois de la langue française :`http://www.oqlf.gouv.qc.ca/ressources/bdl.html`.

87. Notons que pour cet exemple précis, on pourrait aussi y voir une insistance sur la forme de la bouche qui imiterait le baiser lui-même.

L'emploi de lettres capitales peut également marquer la prononciation alphabétique (dans la logique de phonétisation par la lettre, telle que décrite précédemment).

Enfin, les lettres capitales peuvent participer à l'expressivité ; en effet, certaines conventions implicites, communes aux différents types de la Cémo (mais pas uniquement [88]), proposent de considérer les parties du discours en lettres capitales comme des parties exprimées en criant. Cette convention offre donc une information paralinguistique de taille. Comme l'explique Marcoccia (2004a, pages 2-3) : « On retrouve dans ce procédé la fonction d'amplification du verbal par l'accentuation paralinguistique », et, comme le précise Tatossian (2008), les lettres capitales refléteraient alors « l'exacerbation d'une émotion – négative comme un sentiment de colère, ou positive comme la joie ».

5.2.3 Un écrit oralisé, spontané ou simplement un langage ordinaire ?

Comme l'expriment Höflich et Gebhardt (2005, p. 23) : « un e-mail ou un sms peut ressembler quelquefois plus à une lettre et d'autres fois plus à un appel téléphonique » [89]. En effet, certaines caractéristiques qui transparaissent dans la production sms semblent quelquefois relever de **l'oral**, montrent les traits d'une communication spontanée, voire ordinaire. Du point de vue de la théorie des genres, il est donc difficile de définir l'eSMS. À la suite de Charaudeau et Maingueneau (2002, p. 280), nous pensons que pour parler de « genre », on peut prendre en compte « l'ancrage social du discours, [...] sa nature communicationnelle, [...] les régularités compositionnelles des textes [et] les caractéristiques formelles des textes produits ». Nous présentons dans cette partie les divers points qui, selon la littérature, rapprochent cet écrit du langage oral, spontané et ordinaire [90] ; nous nous positionnerons ensuite par rapport à ce rapprochement.

Les recherches qui ont exposé les ressemblances entre la Cémo et l'oralité sont nombreuses. Pour le français, on retiendra par exemple Anis (1998) ; Luzzati (1991) ; Marcoccia (2004a) ; Panckhurst (1998, 2007) et Collot et Belmore (1996) ; Hård Segerstad (2005) ; Yates (1996), pour l'anglais. Certains, dont Ko (1996), ont conclu à un rapprochement plus clair de la Cémo et de l'oral, en se basant sur le fait qu'il s'agit dans les deux cas de communication (pratiquement) en

88. On retrouve ce phénomène dans la bande dessinée également.

89. Notre traduction. En effet, du point de vue de la synchronicité, par exemple, certains sms requièrent une réponse rapide, d'autres sont purement informatifs et ne nécessitent pas de réponse de l'interlocuteur.

90. Ce que Dejond (2006) et Anis (1999) nomment du *parlécrit*. Nous n'adapterons pas cette appellation, à propos de laquelle Bilger et Blanche-Benveniste soulignent : « Beaucoup d'auteurs en sont venus à manipuler des notions comme 'le parlé parlé' et le 'parlé écrit', ou encore 'l'oral dans l'écrit'. Il est aisé de faire la critique d'une telle pratique et de montrer comment, avec ces jeux de vocabulaire, on s'enferme dans un cercle » (Blanche-Benveniste et Bilger, 1999, p. 21).

temps réel, contrairement aux plus anciennes formes de communication écrite[91]. D'un point de vue discursif, il est vrai que le fonctionnement de la communication par sms « à l'implicite, sur présupposés partagés, sous-entendus et inférences conversationnelles » (Gadet, 2004, p. 36), comparable au fonctionnement oral (opposé à l'écrit, qui tend à décontextualiser, qui impose l'explicite), est manifeste, récurrent, mais non régulier, omniprésent ; cela n'en est donc pas un trait.

D'autres (Collot et Belmore, 1996 ; Yates, 1996) ont préféré situer les nouvelles pratiques de l'écrit dans un intermédiaire entre l'oral et l'écrit, considérant un continuum dont les extrémités seraient l'écrit (et les types asynchrones de la Cémo) et l'oral (et les types synchrones de la Cémo). Anis (1998) suggère pour sa part de considérer la communication électronique comme « un hybride » entre l'écrit et l'oral.

Gadet (1989, p. 3) a défini le français ordinaire comme suit : « Ce n'est bien sûr pas le français soutenu, ni recherché, ni littéraire, ni puriste. Mais ce n'est pas non plus (pas seulement) le français oral ou parlé, puisqu'il peut s'écrire. Pas davantage le français populaire, ramené à un ensemble social. C'est davantage le français familier, celui dont chacun est porteur dans son fonctionnement quotidien, dans le minimum de surveillance sociale : la langue de tous les jours ». Il serait donc tentant de qualifier la Cémo de « français ordinaire » (Blanche-Benveniste, 1997) ; les traits qui rapprochent la Cémo du **français ordinaire** peuvent être regroupées comme suit[92] :

- les phénomènes[93] de bruit, telles que les tournures moins formelles (*à qui je peux parler ?*), les phénomènes de dislocation, telle que l'absence de liaison (de ponctuation) ou l'ajout d'un morceau en fin d'énoncé introduit par et (*le bus passe-t-il à 16h et tous les combiens ?*) et les phénomènes d'organisation, tels que *pour* et *comme* employés comme *appuis du discours* (*j'ai comme l'impression que tu ne viendras pas*) ;
- les énoncés sont conçus dans le fil de leur énonciation et les échanges s'exposent à ce que nous avons nommé précédemment l'angoisse de la communication (du silence, de la non réponse) ;
- l'emploi prépondérant du présent de l'indicatif[94] ;
- l'emploi prépondérant du pronom *on* pour *nous* et *ça* pour *cela* ;
- la suppression des pronoms personnels, des prépositions, des conjonctions et des articles (*arrive dans deux heures*) ;

91. Et de compléter : « We may thus see computer-mediated discourse as a written (typed) medium, but with a different language element, the electronic element » (Ko, 1996)

92. Cette liste de présentation n'est pas exhaustive. En outre, nous resterons prudente quant à la validité de tous les phénomènes énoncés, des phénomènes qui sont parfois invalidés par des études basées sur des mesures de fréquence.

93. Cette première classe de phénomènes est empruntée à Luzzati (1991).

94. Pour Cusin-Berche (1999), l'emploi du présent de l'indicatif est tout à fait fonctionnel puisqu'il vise à fusionner le temps de l'énonciation (écriture) avec celui de la réception (lecture).

- les ellipses de segments entiers (*te vois ce soir, pas eu le temps avant*) ;
- l'absence du pronom impersonnel il (*y a pas*) et de l'adverbe de négation *ne* ;
- l'évitement de l'interversion du sujet et du verbe par l'introduction des questions à l'aide de *est-ce que* et par le point d'interrogation en unique marqueur (remplacé à l'oral par l'intonation) ;
- l'absence de schwas (*jtador*) ;
- l'assimilation consonantique (*kèk chose*) ;
- les syncopes et apocopes (*'tain kestufé ?*).

Bien que la présence de ces traits dans la Cémo (et également dans nos propres corpus de sms) soit évidente, nous n'avons pas choisi de rapprocher les nouvelles formes de productions écrites de la production ordinaire, ni de celle de l'oral. En effet, il est trop dichotomique de vouloir scinder la communication en « oral » et « écrit »[95]. D'autre part, il est dangereux de confondre médium de la communication et propriétés de langage. Nous suivons ici le constat de Gadet (2004, p. 33) : « Il faut avant tout opposer le médium [...] et la conception [...]. Le médium relève de la dichotomie, la conception du continuum. Aussi un énoncé d'oral médial peut-il avoir des caractéristiques discursives d'écrit, ou l'inverse ». Nous nous situons donc également dans la continuation des conclusions de Véronis et Guimier De Neef (2006), s'appuyant sur les travaux de Blanche-Benveniste, qui réfutent l'argument selon lequel ces nouvelles formes relèvent d'un « écrit oralisé » ou d'un « oral écrit » ; pour eux, tout est une question de *registres* et de *fréquences* ; ils s'expliquent : « lorsque l'oral devient formel, on y retrouve les tournures caractéristiques de l'écrit [...], et à l'inverse l'écrit informel utilise les tournures fréquentes à l'oral (double marquage, clivées, etc.) ». D'ailleurs, comme le montrent les études de Panckhurst (1998, 1999) sur le courrier électronique, on trouve dans ces écrits des marques informelles alternant avec des marques formelles, comme le « ne » de la négation, qui peut être omis autant que maintenu. De plus, certains traits spécifiques de l'écrit y sont toujours tout à fait présentes ; c'est le cas par exemple de l'interruption de communication, qui est toujours possible sans accord du destinataire et sans notification préalable. En outre, quelques traits fondamentaux de l'oral y sont absents ; on citera par exemple les hésitation et les répétitions (non fonctionnelles), tout ce qui, en réalité, peut être gommé à l'écrit. Nous ne nous trouvons donc définitivement pas devant un cas d'écrit oralisé, ni de français informel ou ordinaire.

Mis à part l'oral, la Cémo a également été assimilée à un **écrit spontané**. En effet, la fréquence parfois élevée des traits moins formels dans cet écrit, associée à d'autres tendances (volonté de transmettre une certaine forme d'expressivité,

95. D'autant que l'usage délicat de l'une ou l'autre expression est renforcé dans notre domaine ; en effet, comme le fait remarquer Anis, « le fonctionnement et le statut de l'écrit sont modifiés en profondeur par la communication électronique ; celle-ci introduit une nouvelle gestion de l'espace et du temps et développe la variation dans les systèmes et les usages graphiques ; les relations entre l'oral et l'écrit en deviennent beaucoup plus complexes » (Anis, 2000, p. 60).

contexte dialogique, absence d'autorité normative, volonté d'appartenance à un groupe socioculturel, etc.), a fait émerger cette notion. Or, même si nous l'employons pour expliquer certains phénomènes précis, cette désignation est trop imprécise pour définir intégralement l'eSMS : le qualificatif *spontané* renverrait au fait qu'il n'existe aucune contrainte linguistique ou extralinguistique qui influencerait la langue et que l'usager du sms s'exprimerait en toute liberté. Or il existe au moins les contraintes d'adaptation à l'interlocuteur[96] : on voit en ce sens une nette différence entre des messages formels adressés à des personnes plus âgées ou d'un statut différent du locuteur, et les messages adressés à des collègues de classe et autres individus de statut identique.

Pour conclure, nous proposons donc d'adopter, comme explicité précédemment, la terminologie *écrit sms* car, médialement parlant (opposition phonique/graphique), il s'agit indéniablement d'un écrit qui est « vu »[97] et interprété. En effet, nous préférons désigner de façon neutre le médium (écrit) et son lieu de création (sms, SRS, etc.) sans présager de son caractère formel ou informel.

En termes de paramètres communicatifs, notre hypothèse pose que ces écrits se situent majoritairement vers le pôle de la *Nähesprache* (Koch et Oesterreicher, 2001), la proximité ou immédiateté, qui s'oppose à la distance. Cette hypothèse est également soutenue par Weininger et Shield (2004) lors de leurs analyses des échanges électroniques simultanés. Nous nous inscrivons donc pleinement dans la théorie d'un continuum conceptionnel entre la « distance communicative » et « l'immédiat communicatif ». Remarquons dès cette réflexion qu'une des questions de recherche qui guident les analyses réalisées dans cette étude consiste à identifier les phénomènes qui se situent clairement vers le pôle de la *Nähesprache* (proximité communicative) de ceux qui se positionnent plutôt vers la *Distanzsprache* (distance communicative). Notre hypothèse sera donc la suivante : un certain nombre de phénomènes linguistiques de l'eSMS sont assimilables à la distance, d'autres à la proximité, d'autres encore se situent entre la distance et la proximité, d'où l'importance d'un continuum afin de positionner les phénomènes intermédiaires. Notre deuxième hypothèse pose l'impossibilité de cataloguer la Cémo en tant que tout unifié, étiqueté « langue spontanée », « langage oral » ou « ordinaire », « immédiat » ou « distance communicative », étant donné l'immensité des variations qui existent sous la réalité de la Cémo et que nous essayons de mettre en avant. Selon le tableau des paramètres pour caractériser le comportement communicatif des interlocuteurs, proposé par Koch et Oesterreicher (2001) (cf. figure 5.10), on remarque par exemple que si le sms est clairement de l'ordre d'une communication privée (1), l'interlocuteur (2) n'est pas forcément intime et l'émotionnalité (3) n'est pas constamment forte (sms informatif,

96. Un usager du SMS réunionnais nous a fait remarquer, à ce sujet, lors de la collecte, « et bien cela dépend des personnes... Par exemple les amis bn jvé ekrir koma, pi la famy je vais écrire comme sa pour bien qu'ils comprennent... » [smsLR] (‘*...ben je vais écrire comme ça, puis la famille je vais écrire comme ça pour bien qu'ils comprennent...*’).

97. Et non de l'oral qui est perçu par notre ouïe.

de type *j'arrive au train de 15h50*).

FIGURE 5.10 – *Paramètres pour caractériser le comportement communicatif des interlocuteurs par rappport aux déterminants situationnels et contextuels (Koch et Oesterreicher, 2001 ; p. 586)*

	Immédiat		Distance
1.	communication privée	1.	communication publique
2.	interlocuteur intime	2.	interlocuteur inconnu
3.	émotionnalité forte	3.	émotionnalité faible
4.	ancrage actionnel et situationnel	4.	détachement actionnel et situationnel
5.	ancrage référentiel dans la situation	5.	détachement référentiel de la situation
6.	co-présence spatio-temporelle	6.	séparation spatio-temporelle
7.	coopération communicative intense	7.	coopération communicative minime
8.	dialogue	8.	monologue
9.	communication spontanée	9.	communication préparée
10.	liberté thématique	10.	fixation thématique
	etc.		etc.

Notre troisième hypothèse consiste en la définition de l'eSMS en tant que *genre*, ou *registre* (Biber et Conrad, 2009), réunissant des textes au contexte situationnel (production écrite échangée par le biais d'un téléphone portable), aux traits linguistiques et aux relations fonctionnelles entre traits et contexte qui sont communs ou partagent des traits communs (Biber, 2011) [98].

5.2.4 Conclusions

Nous avons présenté un panorama des phénomènes graphiques communs à l'eSMS et autres types de la Cémo. Aucun des procédés graphiques exposés dans cette partie n'apparait comme marginal dans l'histoire de l'écrit. La plupart des phénomènes morphologiques apparaissent également dans les jeux sur la langue (par exemple les rébus), la littérature (par exemple l'OuLiPo), la chanson (notamment le slam), etc. L'usage de blancs graphiques n'est pas nouveau : comme le présentent Véronis et Guimier De Neef (2006) : « Certaines langues (le chinois par exemple) s'écrivent toujours sans espaces et avec très peu de ponctuations ». On pourrait alors penser que c'est dans leur utilisation à des fins de brièveté et d'efficacité dans la communication que réside leur originalité. Or, là encore, ces phénomènes ne sont pas pionniers puisqu'on en retrouve la plupart en contexte scolaire (notamment dans les prises de notes) ou encore, comme nous l'avons vu dans la communication par télégramme. La sténographie [99] en a d'ailleurs montré un usage beaucoup plus intense que dans le contexte de la Cémo. Ce qui fait la nouveauté des abréviations dans la Cémo c'est sans conteste leur fréquence d'utilisation (au quotidien) et la population visée (pas d'élite concernée en particulier [100]).

98. Notons que dans Biber (2011), l'auteur définit lui-même le courriel en tant que genre distinct.

99. « Capable d'écrire des dizaines de mots à la minute (parfois plus de cent), de décortiquer les mots en sons, syllabes et consonnes pour en faire de véritables signes à part entière, le sténographe manie l'art de rendre le langage abréviatif et réapprend la langue... par la suppression de l'orthographe » (Dejond, 2006, p. 32).

100. On a vu que même si l'usage d'un ordinateur ne touchait pas toutes les populations dans le monde, l'emploi d'un téléphone portable est par contre beaucoup plus répandu et concerne même

Une fois le travail de présentation des phénomènes graphiques réalisé, il est intéressant de constater que les caractéristiques peuvent entrer en contradiction : ainsi les étirements graphiques et les émoticônes offrent un contraste évident avec les abréviations. Les abréviations ne sont pas toujours de rigueur, contrairement à ce qu'une considération rapide des nouveaux écrits pourrait laisser croire. Tel que l'explique Panckhurst (2009), les phénomènes de réduction et de suppression « correspondent souvent (mais pas nécessairement) à un souhait de *simplification* du message. Le cas de *l'augmentation/ajout* montre que le sms ne vise pas toujours l'économie. La dimension expressive, du point de vue para et non-verbal [,] est essentielle ici. Fayada (2007), dans sa recherche portant sur l'utilisation de l'eSMS par les jeunes (15-18 ans), montre que l'aspect stylistique parait plus important que l'économie de caractères, ce qui résulte parfois en des augmentations de longueurs de messages ». Le besoin de brièveté et de rapidité semble moins la conséquence de contraintes technologiques (limite à 160 caractères, petit clavier, etc.) que de demandes discursives telles que la fluidité dans les tours de parole et dans l'interaction sociale en général [101].

En ce qui concerne les traits de proximité, distance, graphie standard ou altérée, registre standard ou spontané, qui permettent de préciser la nature et le genre d'un corpus de Cémo, nous tenons à préciser qu'ils ne sont pas suffisants. En effet, comme nous le développerons au cours de notre recherche, ces écrits sont aussi particulièrement empreints de traits régionaux, de néologismes, d'éléments allogènes et autres phénomènes qui permettent de définir au mieux le type d'écrit que nous étudions. La Figure 5.11 tente d'illustrer notre approche originale.

5.3 Conclusions

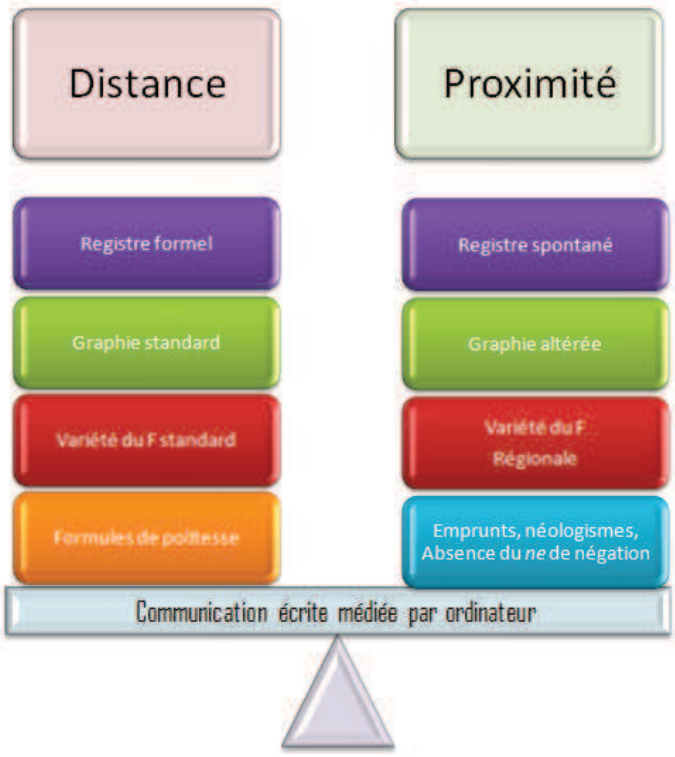
Nous avons montré au travers de ces deux premiers chapitres plusieurs points d'amorce importants pour notre travail de recherche : dans un premier temps, nous avons discuté comment ce qui a déjà été étudié et constitué se distingue de notre propre travail (taille du corpus de sms, informations sur le profil des auteurs, approche de la variation, outils statistiques, etc.).

Dans un deuxième temps, nous avons précisé la nature de tous les types de la Cémo en livrant leurs caractéristiques précises et en les situant les uns par rapport aux autres. Cette démarche nous a permis de placer de façon cohérente l'eSMS au cœur d'un réseau de communication beaucoup plus large et d'en présenter les spécificités. Il nous a donc été possible de définir des variétés bien délimités de la Cémo ; nous avons également mis de côté les sites de réseautage social en tant que variété (puisque les situations de communication et donc leurs traits sont trop hétérogènes).

des populations reculées qui les utilisent pour pratiquer leur langue minoritaire à l'écrit.

101. C'est également l'avis de Thurlow (2003).

FIGURE 5.11 – *Définition du genre de la Cémo en fonction de quelques caractéristiques principales*



Nous avons ensuite réalisé un panorama des phénomènes graphiques qui apparaissent dans les corpus de sms puisque, même s'ils ne constituent pas la matière même de notre travail, ils en fondent les bases et permettront par exemple de ne pas confondre un phénomène d'abréviation du type de la *syncope* avec une marque de régionalisme particulière.

Enfin, nous choisissons de définir l'eSMS comme un type de communication suivant les caractéristiques de Herring (2007) et un écrit dont la graphie, le style, le registre, les traits lexicaux et syntaxiques fluctuent en fonction de la situation de communication, du locuteur, de l'interlocuteur, entre autres variables.

Chapitre 6

Recueil et traitement de données

*"**Chomsky** : The verb* perform *cannot be used with mass word objects : one can* perform a task *but one cannot* perform labour *. **Hatcher** : How do you know, if you don't use a corpus and have not studied the verb* perform *? **Chomsky** : How do I know ? Because I am a native speaker of the English language"*

[...] *Chomsky was, in fact, wrong. One can* perform magic*, for example, as a check of a corpus such as the BNC reveals.*

McEnery et Wilson (2001)

Sommaire

6.1 Constitution de corpus de sms 64
6.1.1 Linguistique de corpus et sms 66
6.1.2 Des sms comme corpus de données textuelles 70
6.1.3 Collecte des données 73
6.1.4 Comparabilité des données 83
6.2 Recueil de données sociodémographiques 101
6.2.1 Constitution et traitement de formulaires 102
6.2.2 Distribution des variables 110
6.2.3 L'échantillon de locuteurs 118
6.3 Conclusions 124

Présentation du chapitre

Ce chapitre présente la méthodologie et les données qui sont à la base des analyses de la deuxième partie de cette thèse. Nous partons du principe, à la suite de Calvet et Dumont (1999, p. 11), que les analyses ne peuvent pas porter uniquement sur l'observation, « il faut donc extraire ces réponses, les révéler [...] et pour cela se donner à la fois des modes d'élaboration du corpus [...] et des principes d'analyse de ces données ». Nous présentons donc d'abord la méthodologie que nous avons suivie. Notre investissement personnel porte sur tous les aspects de la chaine de travail : il mène de la prise de contact individuelle pour former une équipe scientifique prenant en charge les collectes, à l'analyse des données elles-mêmes et à la formulation des conclusions émanant de ces analyses. Nous ne nous attarderons pas

de manière identique sur toutes les étapes de cette chaine, mais nous prendrons soin d'en tenir compte tout le long du travail, notamment lors de la présentation des biais et des limites de notre corpus. Nous détaillerons ensuite les données quantitatives de nos différents corpus : 51.280 messages et 2.563 profils d'auteurs ; nous énoncerons également les singularités de chaque zone, notamment en ce qui concerne la méthodologie, ainsi que les méthodes d'annotation et d'anonymisation. D'un point de vue théorique, et étant donné la nature de notre matériau de base (messages et profils), nous nous devrons de prendre en considération les notions de la linguistique de corpus autant que la notion de corpus pour la sociolinguistique.

6.1 Constitution de corpus de sms

Introduction

Le concept de corpus appartient à cet ensemble de concepts vagues, voire « fourre-tout », qu'il convient de définir au sein d'un cadre théorique précis pour éviter de s'y perdre. En effet, on pourrait définir un corpus en sciences du langage comme un ensemble composé de textes [1] partageant une ou plusieurs propriétés communes : langue, genre, thématique, médium, auteur, etc.

Notre acception de la notion de corpus dans le cadre de notre recherche suit celle de Leech (1991, p. 9) : « un ensemble conséquent de textes exploitables informatiquement » [2]. Nous retenons cette définition, parce qu'elle convient à notre méthode et à notre objectif : collecter un corpus suffisamment conséquent, contenant uniquement du texte et à partir duquel nous travaillons à l'aide d'outils informatiques. L'analyse du discours, par exemple, considère le corpus de manière plus globale, comme un « recueil large [...] de documents ou de données » (Charaudeau et Maingueneau, 2002) et inclut ainsi des données picturales ou mathématiques, ce que nous ne ferons pas ici. D'autres chercheurs colligent et travaillent sur des textes avec une visée commune : pour Kennedy (1998, p. 1), il doit pouvoir servir de base à l'analyse et à la description linguistique [3].

Notre objectif n'est pas d'entreprendre une fois de plus une présentation de la littérature au sujet des corpus. Reprenons et détaillons plutôt la définition de Leech (1991), afin de dépasser le niveau conceptuel en confrontant celui-ci à nos données :

1. Qu'est-ce qu'un « ensemble conséquent » ? Un ensemble de textes qui parait suffisant et satisfaisant aux yeux du chercheur et/ou dans l'optique d'un projet

1. Texte écrit ou transcription de l'oral. On connait également des corpus de gestes en langue des signes : le corpus DEGELS1 (DEfi Geste Langue des Signes, identifiant : crdo000767), pour ne prendre que lui, est le fruit du travail conjoint du Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (Limsi, Orsay) et du Laboratoire parole et langage (LPL, Aix-en-Provence). Ce corpus est constitué de vidéos annotées, qu'une conception du corpus comme celle de Charaudeau et Maingueneau (2002) inclut volontiers.

2. « large bodies of computer-readable text ».

3. « which can serve as a basis for linguistic analysis ».

de recherche donné. Un échantillon contenant suffisamment d'exemples d'un ou plusieurs phénomènes qui suscitent l'intérêt d'une communauté. Nous pensons qu'une prise en compte des deux critères (quantitatif en termes d'exemples et pertinent selon l'optique de la recherche) constitue probablement la meilleure solution pour évaluer un échantillon. Bien que la méthode pour combiner ces deux critères ne soit pas claire, on peut s'efforcer de l'appliquer tout le long de notre démarche. Après avoir collecté un corpus de prime abord *conséquent* tel que celui de sms4science, on peut par exemple évaluer au fil des analyses si les phénomènes étudiés sont *suffisamment représentés* pour qu'ils fassent l'objet d'une étude et qu'ils engendrent des conclusions (c.-à-d. pour éviter de tirer des conclusions à partir d'hapax ou de peu d'occurrences).

2. Qu'entend-on par « texte » dans le cadre de la linguistique de corpus ? Il y a plusieurs manières de définir la composition d'un texte. Pour certains, le focus se fait sur les phrases (par exemple, un texte peut être une « séquence bien formée de phrases liées qui progressent vers une fin », Slakta 1985, p. 138), sur les énoncés (Dubois *et al.*, 1994) ou encore sur les syntagmes (une phrase proverbiale, un dicton même, selon Charaudeau et Maingueneau 2002, p. 571). Pour d'autres, l'unité d'analyse du texte est le lexème (Dister et Moreau, 2006). Selon Dubois *et al.* (1994), est texte tout « échantillon de comportement linguistique qui peut être écrit ou parlé ». Selon cette définition, « texte » serait synonyme de « corpus » ; notre raisonnement deviendrait alors tautologique. À la suite de Halliday et Hasan (1976, p. 293), on s'accordera à dire que, quelle qu'elle soit, l'unité textuelle doit être sémantiquement cohérente ou « signifiante ». Nous préférerons donc entendre la notion de « texte » comme suit : tout discours porteur de signification, fixé à l'écrit ou à l'oral.

3. Qu'entend-on par « exploitable informatiquement » ? Le texte doit être enregistré numériquement (contrairement aux corpus papier), dans un format qui ne prenne pas le risque de perdre des données (XML, CSV, etc.). Une autre attention doit être portée à toute modification des textes, comme ce sera notre cas avec les anonymisations ou, pour d'autres, avec les étiqueteurs syntaxiques, etc. Il faut que les textes soient systématisés (qu'ils montrent le moins de variation possible) et, enfin, exploitables par la machine, par les logiciels que nous prévoyons d'utiliser (dans notre cas : *Unitex*, *Lexico3*, *SPSS*, etc.).

Selon le parcours réflexif que nous venons de mener, un corpus serait donc un échantillon de comportements linguistiques matérialisé sous forme de discours porteurs de signification, partageant (au moins) une propriété commune et servant de base à l'étude scientifique. En ce sens, notre ensemble de sms francophones est bien un corpus : c'est un ensemble de textes (au sens de messages), partageant plusieurs propriétés (rédigés sur téléphone portable, envoyés par ondes et collectés selon la même méthodologie) et servant de base, comme nous l'avons vu précédemment, aux études en (socio-)linguistique, sociologie, psychologie, pédagogie et traitement automatique du langage, notamment.

6.1.1 Linguistique de corpus et sms

La linguistique de corpus ne s'intéresse pas à un aspect du langage lui-même, elle se focalise sur les procédures, les méthodes, qui mènent à l'étude du langage (McEnery et Hardie, 2012). L'activité de constitution et d'étude de corpus n'est pas récente : dès les années 1920, surtout dans le monde anglo-saxon, on a vu se former une tradition de comptage de mots dans des textes donnés, particulièrement dans une visée pédagogique (afin d'enseigner les unités lexicales et les structures grammaticales les plus fréquentes). Ce sont les linguistes de l'école de Prague qui ont systématisé cet élan quantitatif et qui l'ont élargi à d'autres visées. Pourtant, la linguistique de corpus n'en était alors encore qu'à ses prémices, étant donné que le travail se faisait manuellement, et engendrait donc des biais et des limites dans la taille des corpus exploitables. Conséquemment, à travers l'histoire de la linguistique de corpus, ce sont les méthodes de collecte et d'analyse qui ont fondamentalement changé, non ses fondements. Kennedy (1998) explique : « La linguistique de corpus est donc désormais intimement liée à l'informatique, qui a introduit une vitesse incroyable, [...] une reproductibilité exacte, une fiabilité statistique et une capacité à gérer des quantités énormes de données » [4]. Il faut attendre les années 1960 pour voir arriver une telle activité scientifique qui se cantonnait, à ses débuts, à l'étude du lexique. Le Brown Corpus [5] créé par Francis et Kucera de 1961 à 1964 fut le premier « corpus moderne écrit » en anglo-américain. Pour le Royaume-Uni, on citera le LOB Corpus [6], achevé en 1978, par Leech, Johansson et Hofland. La linguistique de corpus des années 80 et 90 est considérée comme une branche de la linguistique qui comprend l'analyse de larges quantités de textes intelligibles par les machines (en utilisant des outils souvent spécialement conçus pour l'analyse textuelle).

Malgré la variété de considérations sur la linguistique de corpus, on s'accordera plus ou moins à dire qu'elle n'est pas une fin en-soi, mais un moyen, un outil visant à aider la compréhension, la description ou l'exploitation de la langue.

La francophonie a suivi de peu le monde anglo-saxon dans le domaine de la linguistique de corpus et a hérité du même engouement. Les indices de cet engouement sont nombreux, en termes de sujets de recherche (on citera par exemple les travaux de B. Habert, A. Salem, F. Rastier, G. Williams ou encore N. Kübler), mais également en termes d'activités scientifiques (par ex. les écoles d'été [7] et les Journées de Linguistique de Corpus [8]), d'activités gouvernementales (par ex.

4. "Corpus linguistics is thus now inextricably linked to the computer, which has introduced incredible speed, [...] accurate replicability, statistical reliability and the ability to handle huge amounts of data".

5. Pour plus d'information : http://icame.uib.no/brown/bcm.html.

6. Pour plus d'information : http ://khnt.hit.uib.no/icame/manuals/lobman/

7. Tel est le cas de l'Ecole d'été Linguistique de Corpus organisée par l'Université de Caen - Basse-Normandie en 2004.

8. Elles sont organisées par l'Université de Bretagne-Sud tous les ans.

le Programme de développement, de mise en réseau et d'exploitation des corpus lexicaux québécois entre 1997 et 2007) et d'initiatives de publications (par ex. la création de la revue *Corpus*[9]). Des corpus importants de langue française ont ainsi vu le jour. On distinguera les corpus oraux des corpus écrits. Parmi les corpus oraux, on citera, à titre d'exemples connus, le Corpus de référence du français parlé de l'Université de Provence (440 000 mots), le Corpus d'Orléans (plus de 900 000 mots) et la base de données orales transcrites ouverte VALIBEL (3 900 000 mots). Parmi les corpus écrits, on recense par exemple la banque de données ouverte Frantext constituée par l'équipe de l'*atilf* à l'université de Nancy (nombre de mots non communiqué), le Corpus français de l'Université de Leipzig (700 millions de mots) et le corpus dynamique GlossaNet[10]. Ces corpus ont été constitués dans le cadre de projets ou de politiques spécifiques, dans une visée disciplinaire souvent précise (sociolinguistique, pédagogique, etc.) ou générale.

Notre recherche sur les sms s'inscrit-elle bien dans la linguistique de corpus ? Pour répondre affirmativement à cette question, nous utilisons la définition élaborée par Laviosa (2002, p. 8) : les grands principes qui guident l'étude du langage selon la linguistique de corpus sont résumables en 8 points. Notre recherche rencontre ces 8 caractéristiques des études de corpus :

1. Notre étude est fondée sur corpus et guidée par corpus. Ses lignes directrices sont principalement descriptives ;
2. Notre étude sociolinguistique envisage la variation dans l'eSMS comme une potentielle volonté individuelle (non systématique) de marquer une identité sociale (régionalismes, anglicismes, argot, etc.) ;
3. Un corpus de sms répond aux exigences communicationnelles ou dialogiques traditionnelles (salutations en début et fin de message, réponses aux questions, etc.)[11] tout en présentant également une créativité importante, tant au niveau de la graphie (néographismes) qu'à celui du lexique (néologismes) ;
4. Le projet sms4science tend à l'hétérogénéité des participants à la collecte : il n'y parvient pas tout à fait, mais le corpus présente une diversité non négligeable en termes d'âge, de classe sociale, etc. La variation régionale, quant à elle, fait l'objet d'une part importante de cette recherche doctorale (par exemple dans la partie 8.3) ;
5. La méthode envisagée par le projet sms4science se focalise sur l'authenticité des messages collectés. De plus, nous adoptons dans ce travail de recherche une attitude résolument basée sur les concordances, les collocations et les fréquences lexicales, auxquelles nous ajoutons des résultats statistiques plus précis encore ;

9. La revue *Corpus* a été fondée par le laboratoire *Bases, Corpus, Langage* de l'Université de Nice Sophia Antipolis en 2001.

10. Pour en savoir plus : `http://glossa.fltr.ucl.ac.be/index.php`.

11. Et lorsqu'il n'y répond pas, la distance par rapport à l'acte communicationnel attendu porte alors une fonction sémantique importante (marqueur d'attitude, etc.).

6. Nous ne traitons pas particulièrement ici de sémantique, mais, lorsque nous le faisons, nous envisageons toujours l'unité analysée dans un contexte plus global (phrase ou message), par exemple via les concordances, voire même en fonction du lieu de collecte (pour les régionalismes) ;
7. Aucune règle n'est à la base de notre démarche, celle-ci est purement exploratoire. Une fois les patterns récurrents de la langue en usage découverts, nous nous efforçons de les décortiquer pour mieux les comprendre et les formaliser ;
8. Nous combinons les deux approches de la science sociale et de la science appliquée en envisageant autant les versants sociaux qui mènent à l'eSMS comme résultat, que les aspects récurrents de l'eSMS (notamment par les statistiques) qui, une fois formalisés, pourront être appliqués aux sciences informatiques, par exemple (mais aussi en pédagogie, en connaissant notamment les types d'erreurs récurrentes que font les élèves dans leurs sms).

Enfin, il est un dernier aspect que nous avons entrepris de définir : il s'agit des « travailleurs de la linguistique de corpus », comme nous choisissons de les appeler. Ceux-ci peuvent être théoriquement scindés en 4 groupes (Kennedy, 1998) :

1. Le premier groupe est constitué de créateurs de corpus (par collecte ou compilation) ;
2. le second groupe renvoie aux créateurs d'outils pour l'analyse de corpus (concepteurs d'algorithmes pour le traitement automatique de la langue) ;
3. Le troisième groupe rassemble les linguistes descriptifs qui basent leur description sur des corpus (et qui peuvent également utiliser les outils élaborés par le groupe 2) ;
4. Enfin, dans le quatrième groupe, les chercheurs exploitent les descriptions linguistiques basées sur corpus (groupe 3) en vue du développement d'applications.

Notre travail s'inscrit dans les groupes 1 (organisation du projet sms4science et collecte de corpus) et 3 (description des formes et des variations linguistiques). Naturellement, nous avons pour cela besoin de nous aider du travail du groupe 2 (par ex. logiciels d'étiquetage syntaxique ou de formalisation et recherche de structures syntaxiques). Le travail ainsi abouti peut servir au groupe 4.

Une fois ces bases définies, nous nous devons de saisir la différence entre un corpus de sms et un corpus plus traditionnel. En quoi la linguistique de corpus dans le cas des sms est-elle distincte de la linguistique de corpus telle que nous venons de la circonscrire ? Tout d'abord, la différence tient à **la définition du corpus en termes quantitatifs**. En effet, pour décrire un corpus, on aura tendance à en donner le nombre d'unités lexicales (cf. Brown Corpus et son million de mots) ou de phrases. Même les corpus oraux, transcrits, présentent une cette description quantitative (cf. Valibel et ses 4 millions de mots, en 2002). Dans le cas d'un corpus de Cémo, l'unité est plus difficilement identifiable. En effet, comme nous l'avons vu dans la partie 5.2.2.1, les unités lexicales au sein des sms sont souvent abrégées,

voire omises, ou agglutinées. La problématique de l'espace blanche séparant et définissant ou non deux unités lexicales est déjà bien connue. Comme le pose Essono (1998, p. 110) : est-ce « une transcription écrite comprise entre deux blancs sur une ligne ? Une portion de la chaine parlée comprise entre deux coupures ? Une unité fondamentale de la lexicologie ? ». Pour ce qui est des sms, le blanc graphique a moins de lien avec la délimitation de l'unité lexicale. Dans le cas qui nous occupe, le même problème se pose quant à la délimitation de la phrase : nous avons vu que la ponctuation à fonction syntaxique est ramenée à des blancs graphiques ou est supprimée. Dans ce cas, quelles sont les possibilités de définition quantitative du corpus ? Premièrement, nous pouvons en donner le nombre de caractères. Ainsi, grâce à l'utilitaire Wordcount de Microsoft Word, nous obtenons les résultats suivants : 2.514.399 caractères en Belgique, 748.484 caractères à la Réunion, 225.256 caractères au Québec et 374.644 caractères en Suisse.

Le corpus global de sms (bruts) à la base de notre étude compte 3.862.783 caractères.

Deuxièmement, nous pouvons nous focaliser sur la définition quantitative du corpus transcrit en graphie standard. Ce nombre indiquerait effectivement la quantité de lexies transcrites. Ainsi, l'outil AntConc [12] nous offre les chiffres suivants : 688.550 tokens [13] en Belgique, 233.285 tokens à la Réunion, 60.809 tokens au Québec et 94.398 tokens en Suisse.

Le corpus global de sms (transcrits) à la base de notre étude compte 1.077.042 mots (tokens).

Ces deux définitions de corpus [14] montrent elles aussi des inconvénients. Dans le cas du nombre de caractères, on aurait tendance à se demander ce qu'illustre ce nombre, surtout lorsqu'on sait que de nombreux caractères sont omis dans le corpus de messages bruts. Dans le cas des transcriptions en graphie standard, nous nous posons la question de l'exactitude terminologique pour un objet précis : le nombre de mots du corpus *transcrit* peut-il légitimement définir le corpus *brut* (plusieurs corpus bruts distincts pourraient donner le même corpus transcrit après normalisation) ? Ne trouve-t-on pas ici un problème de mutation de l'objet même de l'étude ?

La solution est peut-être de ne pas définir la valeur d'un corpus en termes de nombre de mots, mais plutôt par le truchement d'autres critères, tels que la disponibilité des données rares ou de données dont le recueil est couteux (en termes financiers et/ou méthodologiques) [15]. Le corpus sms4science répond à ces deux critères : les

12. Pour en savoir plus : `http://www.antlab.sci.waseda.ac.jp/software.html`

13. Pour la terminologie, nous distinguerons le type (nombre d'unités **différentes** d'un texte) du token (nombre d'unités d'un texte, y compris celles qui se répètent plusieurs fois).

14. Notons qu'il aurait été également possible de définir notre corpus de sms transcrit en termes de nombre de phrases. Toutefois, étant donné le protocole de transcription 6.1.4.2 qui préconise de ne pas réintroduire de nouvelles ponctuations, cette solution n'est pas envisageable.

15. Ceci est notamment proposé par (Dister *et al.*, 2009).

messages issus d'échanges sms sont difficiles à obtenir en grande quantité sans un rayonnement médiatique, une publicité importante qui impose beaucoup en termes financiers et humains. Ces messages sont aussi techniquement plus difficiles à récolter (par exemple par rapport à des corpus de chat) puisqu'ils transitent via le médium téléphone portable et non directement via un ordinateur qui permet un enregistrement immédiat. Enfin et probablement surtout, ces messages sont très difficiles à obtenir par rapport à des corpus littéraires, par exemple, parce qu'ils ne sont pas du domaine public et, contrairement aux corpus littéraires, il n'est pas possible d'en acheter les droits ; ainsi, les aspects de protection de la vie privée entravent le processus de collecte et ajoutent de la valeur au corpus.

On peut également mettre en avant la singularité d'un corpus de sms par sa **position par rapport à la notion de conversation**. Certains types de Cémo tendent clairement vers le pôle de la conversation écrite (ainsi chat et MI). D'autres, tels que les blogs et les sites de réseautage social tendent vers le pôle du monologue public. Un troisième groupe flotte entre ces deux extrêmes et comprend sms et courrier électronique. Selon la situation de communication, nous nous trouvons clairement dans la conversation, dans l'échange, ou clairement dans la communication unidirectionnelle. Il s'agit d'un point important puisque pour définir un corpus de textes, il est bon, comme nous l'avons vu, de circonscrire ses propriétés. Dans ce cas, nous poserons qu'il s'agit d'un corpus de messages écrits asynchrones, et non de conversation. Enfin, notons que le corpus de sms que nous possédons se distingue totalement des corpus des autres types de la Cémo puisque, même à l'intérieur du sous-corpus de véritables conversations par sms, nous possédons uniquement la copie des messages d'un des deux locuteurs. En aucun cas donc, nous ne pouvons définir notre corpus comme étant un corpus de conversations écrites. Nous garderons à l'esprit que certaines propriétés conversationnelles pourraient s'y retrouver çà et là. En ce qui nous concerne, il s'agit simplement d'un **discours** [16], que l'on nommera, à la suite des explications de la partie 5.2.3, « **de proximité** ».

6.1.2 Des sms comme corpus de données textuelles

Le réseau académique du projet sms4science comprend 19 institutions de recherche (universités et fondations) réparties sur 19 zones de collecte appartenant à 10 pays. Nous présentons ce réseau international sous forme de tableau (Figure 6.1 : par ordre alphabétique).

Le projet initial de collecte de sms fut initié en 2004 par le Cental, à l'Université catholique de Louvain. Le projet, alors baptisé *faites don de vos SMS à la science !*, portait sur la récolte des messages et des profils, en Belgique francophone uniquement. L'extension internationale du projet *sms pour la science* a débuté en 2007 grâce au soutien de Belgacom-Proximus. Le projet fut alors rebaptisé *sms4science*

16. Au sens de Benveniste (1966, p. 266) : « la langue en tant qu'assumée par l'homme qui parle, et dans la condition d'intersubjectivité qui seule rend possible la communication linguistique ».

FIGURE 6.1 – *État du projet sms4science en date du 12 octobre 2011*

Pays	Zone de collecte	Université partenaire	Année de collecte
Allemagne	Allemagne	Université de Cologne	
Belgique	Belgique francophone (Wallonie + Bruxelles)	Université catholique de Louvain	2004
Canada	Québec	Université de Montréal	2010
	Canada anglophone	Simon Fraser University	2011
Espagne	Pays Basque	Fondation Elhuyar	
	Catalogne	Université autonome de Barcelone	
France	Hautes Alpes	Université Stendhal Grenoble 3	2010
	Languedoc-Roussillon	Université Paul-Valéry Montpellier 3 Université Montpellier 2	2011
	Poitiers	Université de Poitiers	
	Rennes	Université Haute Bretagne Rennes 2	
	La Réunion	Université de la Réunion	2008
Grèce	Grèce + Chypre	Aristotle University of Thessaloniki	
Italie	Italie	Université de Padoue	
Roumanie	Roumanie	*Université* «Babeç-Bolyai» de Cluj (Roumanie)	
Royaume Uni	Royaume Uni	University of Birmingham	
Suisse	S. Alémanique	Université de Zurich	2009
	S Romande	Université de Neuchâtel	2009
	S Italienne	Université de Berne	2011

pour répondre aux exigences de son internationalisation. Les contacts se sont assez rapidement intensifiés, la renommée du projet augmentant dans le monde de la linguistique (de corpus). Durant les 3 années qui suivirent, nous avons travaillé sur des partenariats avec les pays suivants : Afrique du Sud, Algérie, Allemagne, Argentine, Burkina Faso, Cameroun, Canada, Danemark, Espagne, États-Unis, France, Grèce, Inde, Israël, Italie, Luxembourg, Pays-Bas, République tchèque, Roumanie, Royaume-Uni et Suisse.

Le tableau 6.1 présente le résultat du travail de coopération scientifique. Finalement, près de la moitié (9) des pays avec lesquels nous nous sommes entretenus (21) au sujet du projet ont signé un partenariat avec l'UCL ; ce résultat est plutôt positif. Pourtant, seul un septième des pays contactés a effectivement collecté des sms ; ce

pourcentage est moins encourageant, bien qu'il ait été suffisant pour s'engager dans notre étude sociolinguistique. Cet état de fait témoigne de la charge importante que présente la mise en place d'un tel projet.

Sur le site Internet [17] dédié au projet belge de 2004, on trouve dans les objectifs principaux de la collecte le postulat suivant : « Une difficulté se présente quand on cherche à mener une étude dans ce domaine : c'est l'absence de tout corpus de référence. Un tel corpus est nécessaire pour envisager des études linguistiques de grande envergure et pour permettre de nouveaux développements dans le domaine du TAL (Traitement automatique du Langage). C'est à ce besoin que notre projet tentera de répondre ». Notre attention porte sur cette notion de *corpus de référence*. Cet objectif ambitieux, fixé au démarrage, n'a pas été réellement atteint (Cougnon et François, 2010). Nos collectes ne permettent justement pas d'obtenir des corpus *de référence* (au sens de Habert 1999, pages 14-15 cité dans Dister 2007) puisqu'elles ne garantissent pas la représentativité en termes de diversité et d'équilibre des sous-catégories représentées (le spontané/non spontané, français standard/régional, le vocabulaire technique/scientifique, etc.). Le projet international cherche à assembler une masse de données suffisante, mais ne vise pas la diversité et l'équilibre parfaits, que nous savons quasiment impossibles à atteindre (cf. partie 6.2.3) ; nous cherchons la spontanéité des productions réelles dans les sms et leur vocabulaire spécifique, dont nous ignorons encore les propriétés.

Le projet sms4science a pour ambition la collection de sous-corpus pour la constitution d'un (méga) corpus de sms dont les caractéristiques sont les suivantes :

1. Il s'agit d'un **corpus ouvert** (opposé à *clos*) destiné à être enrichi régulièrement, à chaque nouvelle collaboration, jusqu'à ce que l'étude des sms perde elle-même son sens, en raison de la disparition de cette pratique, ou bien jusqu'à ce que les fonds permettant de mener à bien les collectes viennent à manquer ;
2. Le projet n'est pas régi par les *objectifs d'un chercheur unique*, mais est enrichi par des contributeurs ayant chacun des axes de recherche qui s'inscrivent dans des domaines (partiellement) diversifiés : analyse du discours, communication, psycholinguistique, syntaxe, etc. ;
3. Comme nous l'avons dit, il ne vise pas l'obtention d'un *corpus de référence*, mais d'une masse de données suffisantes pour exemplifier des phénomènes linguistiques divers ;
4. Ce n'est donc pas un corpus *complet* qui vise à l'*exhaustivité* ;
5. Le corpus global *sms4science* est un corpus plurilingue ; celui utilisé dans le cadre de cette dissertation est uniquement **francophone**. Chaque sous-corpus brut est également transcrit en graphie standard, de sorte que les sous-corpus bruts et transcrits sont **alignés** (au message près, cf. partie 6.1.4.2) ;

17. http://www.smspourlascience.be/index.php?page=1

6. Il s'agit donc plus précisément d'un corpus dit **dynamique**, qui ne fonctionne cependant pas comme **corpus de veille** ou « monitor corpus » (Renouf, 1993) (observatoire de la langue en usage dans les sms) puisque l'arrivée des données n'est pas constante ;
7. Parce qu'ils sont le produit d'une méthodologie commune (cf. parties 6.1.3.1 et 6.1.4), les sous-corpus sont dits « **comparables** » ;
8. En raison des différentes dates de collecte (de 2004 à 2010, cf. partie 6.1.3.2), nous considérons que notre corpus est **diachronique** ;
9. Enfin, il s'agit d'un corpus **spécialisé**, au sens de Kennedy (1998), puisqu'il est composé exclusivement de sms et, qui plus est, de sous-corpus de sms régionaux (par exemple, au regard de la francophonie). En tant que corpus spécialisé, et au vu des chiffres analysés (cf. partie 6.1.3.3), il ne vise pas la *représentativité*.

6.1.3 Collecte des données

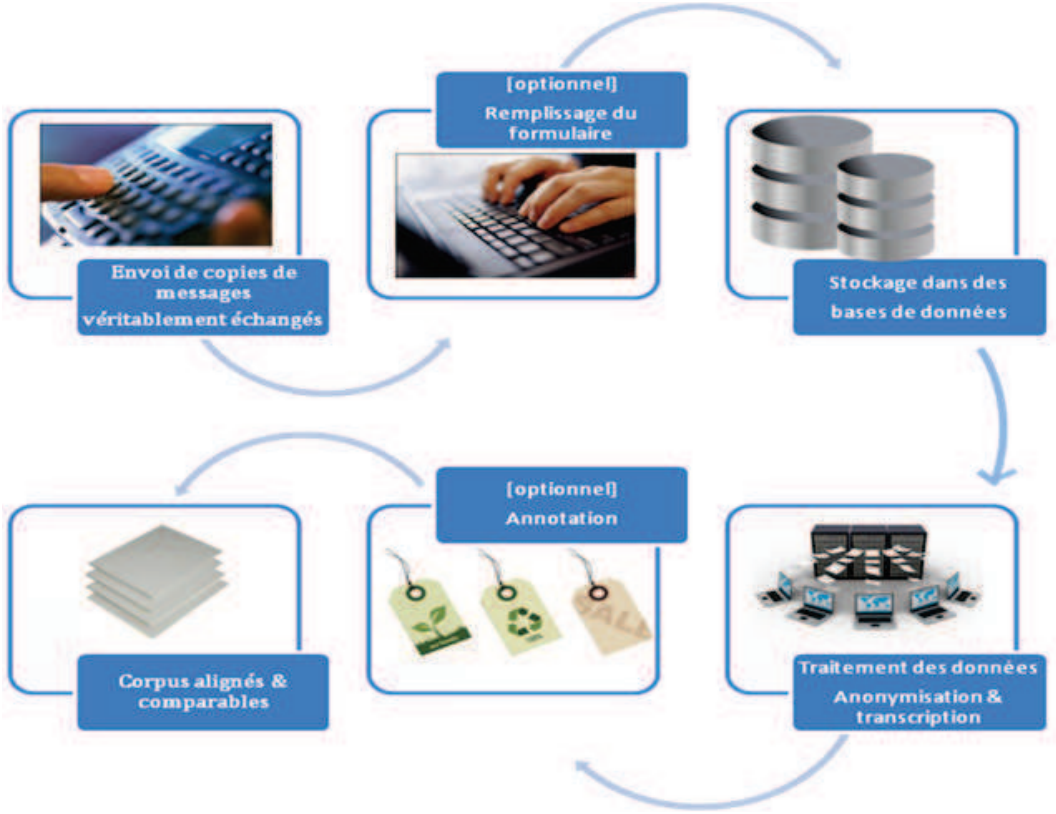
6.1.3.1 Méthodologie de collecte

La méthodologie de collecte du projet sms4science, même si elle a différé en certains aspects (cadeaux, numéro gratuit, etc.), est basée sur des principes généraux correspondant à la méthodologie du projet initial que les différents partenaires ont été invités à suivre. Ces lignes directrices touchent autant à la logistique et à la technique qu'à la communication ou au juridique. Nous énonçons ici les principes de collecte qui ont indubitablement marqué, voire quelquefois biaisé, les données finales. La Figure 6.2 schématise l'ensemble de la méthodologie.

La logistique. Le bon fonctionnement du projet repose sur la constitution dans chaque région d'une équipe suffisante en nombre de personnes et multitâche. En effet, chaque collecte est prévue pour durer approximativement 3 mois à temps plein. Or, les personnes enrôlées dans ce projet appartiennent à des centres de recherche, à des départements, à des organismes qui leur imposent déjà une charge de travail. Le fait de disposer de plus de deux personnes-ressources par équipe locale est donc une nécessité durant la période de collecte. En raison des tâches très différentes que requiert le projet (par ex. communication avec les médias et gestion des bases de données), les personnes-ressources montreront des profils de compétence hétérogènes. Après la collecte proprement dite, un travail de traitement doit également être prévu. Ce traitement des données, par son fondement même, donc par souci d'homogénéité, doit préférablement être réalisé par une seule et même personne.

La technique. La Figure 6.2 présente les diverses étapes où intervient la gestion technique du projet. Les participants au projet sont invités à rediriger des copies des sms qu'ils ont préalablement envoyés à un destinataire réel vers un numéro de téléphone dédié au projet. La gestion de ces données se fait chez un opérateur de téléphonie, mais il est nécessaire que l'équipe scientifique vérifie le bon stockage des

FIGURE 6.2 – *Méthodologie générale de collecte du projet sms4science*



données (encodage de caractères spéciaux, maintien des sauts de lignes, fusion des messages en plusieurs parties, etc.). Les participants sont également encouragés à remplir un formulaire d'information les interrogeant sur leurs données personnelles et leurs pratiques en matière d'utilisation de sms. La récupération des informations sociodémographiques est mise en pratique par le biais d'un formulaire hébergé sur Internet ; ces informations sont ensuite stockées sur des serveurs belges (pour les projets belge, réunionnais et québécois) [18] ou suisses. Le numéro de téléphone portable des participants sert de clé d'identification permettant de lier messages et formulaires a posteriori.

La communication. La qualité du plan de communication participe directement au bon fonctionnement du projet. Ceci concerne la communication avec la population et les médias. Le discours direct des organisateurs envers la population se fait principalement via le **site Internet du projet** (qui sert de vitrine) et via **les réseaux sociaux** [19]. Le contenu de ce discours est tout à fait important et

18. Nous tenons à remercier H. Naets et I. Lecroart pour la gestion des formulaires, des sites des partenaires (Drupal) et autres aspects techniques du projet sms4science en Belgique, à la Réunion et au Québec.

19. Notons que le projet belge, datant de 2004, n'a pas bénéficié de ce vecteur de communication.

doit inclure trois points majeurs : la récompense des participants (via des tirages au sort) qui reçoivent des lots ou « cadeaux » (de type carte prépayée, téléphone portable, cours de cuisine, etc.), la **garantie de l'anonymat** et la **gratuité de la participation** [20]. Ces deux derniers aspects doivent être particulièrement bien mis en avant lors de la présentation du projet à la population ou aux médias. Le succès du projet sms4science dépend en grande partie de sa **promotion par les médias**. En effet, puisqu'il s'agit de faire participer une proportion importante de la population (« le grand public »), les médias jouent un rôle prépondérant de vecteur d'information, que ce soit, comme nous venons de le dire, pour mettre en avant la gratuité ou des lots à gagner, mais également pour transmettre l'intérêt d'une telle collecte pour la science. Comme toutes les classes d'âge de la population sont visées par la collecte, il est nécessaire d'intéresser une partie de celle-ci qui n'est touchée ni par des « cadeaux », ni par les réseaux sociaux. Cela se fait notamment via des articles de vulgarisation scientifique mettant en avant les enjeux d'une telle étude, lors du lancement, et des articles présentant l'évolution des données et les premiers résultats, tout le long de la collecte. Concrètement, il s'agit donc pour chaque équipe d'organiser une conférence de presse et de répondre à des entretiens, que ce soit pour des journaux, des magazines, la radio ou la télévision.

Le juridique. Lorsque l'on cherche à étudier des données personnelles, un certain nombre de considérations éthiques apparaissent très rapidement. Elles touchent autant à la méthodologie qu'aux sms et profils des auteurs collectés, afin de respecter la vie privée. La méthodologie passe notamment par une incitation à la participation par des lots. Ceci peut engendrer des complications dans certaines régions puisque le projet peut ainsi tomber sous le coup de la loi sur les jeux de hasard. Pour parer à ce problème, il est important d'être précis sur le critère d'attribution des lots ; il s'agira notamment, comme cela a été fait en Belgique, d'insister sur la gratuité d'envoi lors de la participation (s'il n'y a pas de mise, il n'y a pas de jeu de hasard). Les messages eux-mêmes nécessitent un travail d'anonymisation de données personnelles ; les données personnelles sont celles qui permettent l'identification d'un individu – quel qu'il soit – dans les messages : nom, prénom, adresse, numéro de téléphone, numéro de carte bancaire, adresse de blog, etc. Les instances juridiques (étatiques [21] ou universitaires) de la plupart des régions requièrent également l'acceptation par les participants d'un consentement de participation spécifiant notamment : « les sms sont transmis volontairement et librement par des utilisateurs témoins qui déclarent être les auteurs des messages transmis, connaitre les objectifs de cette collecte et autoriser, etc. » [22]. Au niveau du repérage des données, les numéros de téléphone qui, comme explicité précédemment, servent dans un premier temps de clé d'identification, doivent ensuite être transformées en clés d'identification neutres. Pour ce qui est des formulaires, un consentement est

20. Ce qui signifie que les auteurs de messages ne payent pas lors de l'envoi des sms aux organisateurs. C'est en général le sponsor qui se charge du cout des messages.

21. On citera par exemple la CNIL pour l'ile de la Réunion : `http://www.cnil.fr/`.

22. Passage provenant de la description détaillée du projet.

également – numériquement – signé et les clés d'identification sont aussi transformées. Les données sociodémographiques elles-mêmes ne sont pas altérées. Un travail manuel de lecture des zones de commentaire libre est réalisé *a posteriori* afin de mettre de côté tout commentaire à caractère injurieux, obscène ou discriminatoire. Chacune de ces précautions peut être prise en charge par l'opérateur ou par l'équipe scientifique[23]. Enfin, le droit des usagers ira jusqu'à la possible consultation de la base de données une fois constituée et la suppression éventuelle de leurs messages.

La mise sur pied des partenariats ainsi que la mise en œuvre des collectes ont montré un certain nombre d'obstacles et de contraintes que nous nous devons d'éclaircir ici. Au niveau des partenariats, il nous a été difficile d'établir des équipes dans les pays d'Afrique du Nord, que nous visions pourtant particulièrement en vue de notre étude sur la francophonie. Dans ce cas, la difficulté se situait au niveau de la situation géopolitique et économique des régions intéressées. Nous avons également rencontré des problèmes au moment de trouver des sponsors : en effet, certaines régions montrent un moindre intérêt pour la recherche scientifique exposée à la population et n'encouragent donc pas la participation de sponsors dans des projets de recherche universitaire. Cette absence de sponsor a causé l'abandon de la collecte auprès d'un partenaire officiel et le faible taux de participation à la collecte d'un autre. Nous avons également rencontré le cas d'un sponsor intéressé par une copie de la base de données de messages et par un financement dans le projet, mais aucune équipe universitaire disponible (manque de temps et/ou de ressources humaines). Au niveau de la mise en place de la collecte, certains comités d'éthique font preuve de plus d'exigences que d'autres, ce qui entraine, d'une part, un ralentissement de la préparation de la collecte, et, d'autre part, un ralentissement dans le traitement a posteriori des données (voir le point 6.1.4.1 sur les traitements particuliers des prénoms par la Suisse). Enfin, au niveau de la collecte elle-même, la faible participation au projet dans deux régions s'explique par une liste finie de facteurs. Tout d'abord, comme nous l'avons énoncé plus haut, la non-gratuité de l'envoi des sms par les participants[24] est un facteur majeur de non-participation. Même si le projet intéresse par sa teneur scientifique ou amuse parce qu'il répond à un effet de mode, cela ne suffit pas à encourager une population à participer, surtout si cette dernière doit s'investir financièrement. La non-gratuité de la participation engendre une difficulté supplémentaire au niveau de la volonté d'échantillonner (aléatoirement) des données. Enfin, on constate également qu'un faible impact du projet dans les médias entraine, comme attendu, peu d'intérêt de la population dans le projet et donc un

23. En Suisse, par exemple, les universités ont reçu les sms de la compagnie Swisscom sans indications de l'expéditeur. Ni les numéros de téléphone, ni les adresses électroniques n'étaient à aucun moment visibles pour les collaborateurs des universités.

24. On peut par exemple lire dans les *Conditions de participation* du Québec « L'utilisation du téléphone portable n'est pas gratuite. Malheureusement, nous ne pouvons compenser tous les participants pour les frais encourus. Aussi, chaque semaine deux gagnants d'une carte prépayée d'une valeur de $10 seront désignés par tirage au sort parmi les participants qui auront envoyé 5 textos dans la semaine et qui auront complété le questionnaire » (`http://texto4science.ca/?q=fr/node/9`).

faible taux de participation.

6.1.3.2 Collectes régionales

Nous allons ci-après présenter les corpus de données textuelles obtenues dans chaque pays.

Belgique Tous les détails concernant cette collecte se trouvent dans Fairon *et al.* (2006b) et le corpus collecté a été publié dans Fairon *et al.* (2006a). Nous ne nous étendons donc pas dans cette partie sur le détail de ces données. Nous en résumons simplement les grandes lignes afin de permettre la comparaison par région.

La collecte belge s'est officiellement tenue du 15 octobre 2004 au 15 décembre 2004 [25] (`http://www.smspourlascience.be/`). Elle a été prise en charge par l'Université catholique de Louvain, en particulier le Cental et Celexrom, deux centres de recherche de l'UCL. L'enquête a été étendue aux populations francophones de Belgique (Wallonie et Bruxelles). Le projet a été soutenu par la société de téléphonie Belgacom/Proximus.

Le corpus de messages belge est monolingue. Ceci découle d'une politique de traitement des données qui a écarté les messages dans d'autres langues que le français. Ce qui signifie que des sms en d'autres langues ont été collectés, mais qu'ils n'ont pas été maintenus dans le corpus final. La suppression a été uniquement appliquée aux messages intégralement rédigés dans une ou plusieurs autre(s) langue(s) que le français. Les messages contenant des formes de code-mixing, code-swiching ou autres phénomènes de contact des langues où le français est présent (partiellement), n'ont donc pas été éliminés (ce qui a permis notamment des études sur ces phénomènes de mélange de langues).

La collecte belge a permis de rassembler 70.000 messages ; dans le cadre de notre recherche, nous n'en utilisons cependant que 30.000, qui ont été transcrits en graphie standard. Ce corpus de 30.000 messages bruts comprend **2.514.399 caractères** espaces non compris [26]. Le corpus de 30.000 messages transcrits présente **688.550 mots occurrences** (unités lexicales simples) encore appelés tokens [27] pour 24.265 mots types [28]. En termes de densité lexicale (Herslund, 2003, p. 37), on obtient donc un rapport type/token de 3,52 % [29].

25. Dans les faits, elle a été prolongée d'un mois.
26. Chiffre basé sur le corpus original (non anonymisé) d'après l'application Wordcount de Microsoft Word.
27. D'après l'outil AntConc.
28. Notons que le nombre de types est légèrement restreint puisque les analyses portent sur un corpus transcrit et anonymisé. Ainsi, tous les prénoms, les noms de rue etc. ont été réduits à un tag identique et donc à un seul type.
29. Nous reviendrons plus en détail sur ces chiffres dans la partie 8.1.1.

La Réunion La collecte réunionnaise a officiellement eu lieu du 15 avril 2008 au 30 juin 2008 (http://www.lareunion4science.org/). La collecte a été prise en charge par l'Université de la Réunion, en particulier le Laboratoire Langues, textes et communication dans les espaces créolophones et francophones, un centre de recherche de l'Université de la Réunion. Le projet a été encouragé financièrement par la société de téléphonie SFR et la Région Réunion et a concerné uniquement le territoire de l'ile de la Réunion.

Le corpus de messages réunionnais est bilingue français-créole. Ceci découle d'une politique de traitement des données qui a accordé aux deux langues le même intérêt scientifique. Pour les besoins de cette dissertation, nous n'analyserons pas les parties créoles. Les messages totalement en langue créole étant très difficilement identifiables (notamment en raison de la similitude de certaines transcriptions du créole avec les pratiques propres à l'eSMS), nous gardons la totalité du corpus comme matériau d'analyse. Notons encore qu'il faudra bien se garder de mélanger la langue créole, qui constitue un système linguistique distinct du français, de la variété régionale du français à la Réunion qui, elle, requiert toute l'attention de notre analyse de sms francophones.

La collecte réunionnaise a permis de rassembler 12.660 messages qui ont été transcrits en graphie standard. Ce corpus de 12.660 messages bruts comprend **748.484 caractères** espaces non comprises[30]. Le corpus de 12.660 messages transcrits présente **233.285 mots occurrences** pour 13.080 mots types. En termes de densité lexicale, on obtient donc un rapport type/token de 5,6 %.

Québec La collecte québécoise a officiellement eu lieu du 23 novembre 2009 au 30 avril 2010 (http://texto4science.ca/). Notons que cette collecte fut officieusement prolongée de quelques mois ; toutefois, notre étude se base sur les données récoltées durant la période officielle. La collecte québécoise a été prise en charge par l'Université de Montréal, en particulier le Département de linguistique et de traduction et le Département d'informatique et de recherche opérationnelle. Le projet a été soutenu par la société de portails pour opérateurs mobiles Adenyo et le Centre de recherche interdisciplinaire sur les technologies émergentes (CITÉ). La collecte a été étendue aux populations francophones du Canada (Québec et environs).

Le corpus de messages québécois est monolingue. Ceci découle d'une politique de traitement des données qui a écarté les messages qui n'étaient pas écrits en français (principalement en anglais). Ce qui signifie que des sms écrits en d'autres langues ont été collectés, mais qu'ils n'ont pas été maintenus dans le corpus final. Seuls les messages totalement rédigés dans une ou plusieurs autre(s) langue(s)

30. Chiffre basé sur le corpus original (non anonymisé) d'après l'application Wordcount de Microsoft Word.

que le français ont été éliminés. Les messages contenant du code-mixing, du code-swiching ou d'autres phénomènes de contact des langues où le français est présent (partiellement), ont été maintenus.

La collecte québécoise a permis d'assembler 4.731 messages qui ont été transcrits en graphie standard. Ce corpus de 4.731 messages bruts comprend **225.256 caractères** espaces non comprises[31]. Le corpus de 4.731 messages transcrits présente, selon AntConc, **60.809 mots occurrences** pour 7.056 mots types. En termes de densité lexicale, on obtient donc un rapport type/token de 11,6 %.

Suisse La collecte suisse a officiellement eu lieu du 1er novembre 2009 au 15 décembre 2009 (officieusement jusqu'au 14 janvier 2010, http://sms4science.ch/). Une seconde collecte italienne et romanche s'est déroulée en Suisse en 2011, mais ne fait pas l'objet de cette étude. La collecte suisse a été prise en charge par l'Université de Neuchâtel, en particulier l'Institut des sciences du langage et de la communication et le Centre de linguistique appliquée, par l'Université de Zurich, en particulier les Deutsches et Romanisches Seminars, et par l'Université de Berne, en particulier l'Institut für italienische Sprache und Literatur. Le projet a été encouragé financièrement par la société de téléphonie Swisscom.

Le corpus de messages suisse est plurilingue ; on compte principalement l'allemand, le français, le dialecte alémanique et l'italien. Ceci découle d'une politique de traitement des données qui a accordé à toute langue le même intérêt scientifique. Pour les besoins de cette dissertation, nous écartons les messages intégralement non francophones.

Une partie majeure du déroulement et les premiers résultats du projet suisse sont décrits dans Duerscheid et Stark (2010).

TABLE 6.1 – *Répartition des langues en Suisse (en nombre de messages)*

dialecte alémanique	allemand	français	italien	rhéto-roman
10.664	7.174	3.924	487	214

La répartition présentée dans la table 6.1 est le résultat d'une sélection des langues réalisée via un un script de reconnaissance des langues en Suisse. Ce processus est donc imparfait et nécessite un deuxième passage manuel, qui aura effectivement lieu d'ici 2013, en Suisse. Pour la partie francophone, ce travail a été réalisé par une étudiante à Louvain-la-Neuve en 2011[32].

31. Chiffre basé sur le corpus original non anonymisé d'après l'application Wordcount de Microsoft Word.

32. Un remerciement à Arianne Torné, étudiante d'origine suisse romande, qui a assuré le nettoyage et la transcription du corpus suisse francophone. Nous la remercions également pour l'intérêt

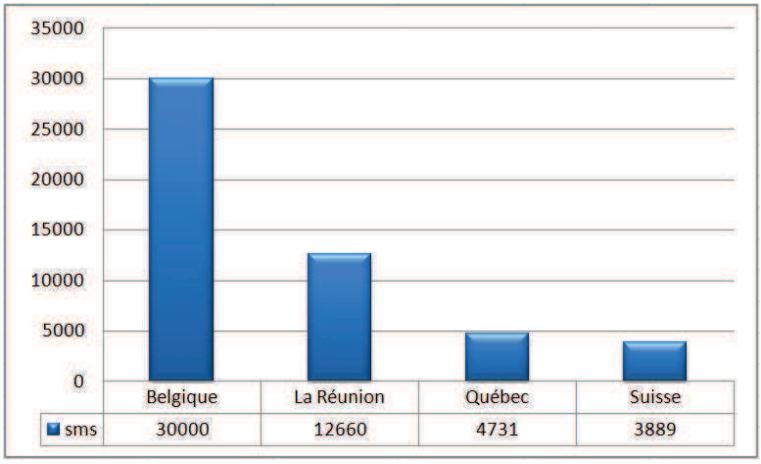
La collecte suisse a permis d'assembler 3.889 messages francophones qui ont été transcrits en graphie standard. Ce corpus de 3.889 messages bruts comprend **374.644 caractères** [33] espaces non comprises [34]. Le corpus de 3.889 messages transcrits présente **94.398 mots occurrences** pour 7.395 mots types. En termes de densité lexicale, on aura donc un rapport type/token de 7,83 %.

6.1.3.3 L'échantillon de données sms

Un tel projet de collecte de données sms et profils pour la recherche scientifique a fondamentalement pour volonté d'obtenir un échantillon *aléatoire* et *représentatif*. Bien que l'ensemble de la réflexion méthodologique sur la collecte et le traitement des données ait toujours tendu vers l'obtention de ces deux caractéristiques de l'échantillonnage, nous montrons en quoi celles-ci restent utopiques. Nous présentons plutôt notre corpus comme un **corpus spécialisé**, c'est-à-dire par définition, **non représentatif** et nous décrivons l'**échantillon** comme **contingent**, puisque l'échantillonnage se fait par volontaire.

L'hétérogénéité des quantités de données sms par région mérite d'être détaillée. La Figure 6.3 reprend la distribution des corpus de sms.

FIGURE 6.3 – *Répartition des sms francophones (sms4science)*



Le corpus belge est composé de 30.000 messages pour une population francophone de Belgique ayant envoyé 363,5 millions de sms durant 3 mois, en 2004. Ce chiffre est calculé à partir de données provenant de plusieurs sources : la première (une

qu'elle a accordé aux phénomènes linguistiques, ainsi que pour les échanges informels, très instructifs pour nous, qui en ont découlé.

33. On remarque, dès ce niveau que la Suisse, avec un corpus de messages plus restreint que le Québec, dépasse pourtant ce dernier en nombre de caractères.

34. Chiffre basé sur le corpus original anonymisé (d'où une légère réduction du nombre de caractères par rapport au nombre de caractères réel) d'après l'application Wordcount de Microsoft Word.

enquête de la Libre Belgique de 2005 [35]) indique que le nombre de sms envoyés en 2004 en Belgique s'élève à 3 milliards. Nous arrivons à ce nombre en combinant les chiffres donnés par Proximus et Mobistar par la Libre (respectivement 1,9 milliard et 1,095 milliard - 3 millions par jour -). Chez Base, aucun interlocuteur n'a accepté de nous donner les chiffres (qui n'apparaissent pas non plus sur le réseau Internet). Cependant, nous pouvons engendrer un chiffre par simulation en fonction des parts de marché (communiquées dans le même article de la Libre) : 33% pour Mobistar et 18% pour Base (on en déduit 49% chez Proximus). Selon cette méthode approximative, le nombre de sms envoyés en Belgique en 2004 chez les trois opérateurs (α) est donc de 3,57 milliards. Ce chiffre doit ensuite être ramené à 40% pour correspondre à la population francophone de Belgique (ϕ) en 2004, selon notre seconde source (le CIA World Fact Book [36]). Nous obtenons alors 1,454 milliards de sms envoyés par des usagers du sms francophones en 2004. Enfin, on restreint ce nombre aux trois mois de collecte (δ) afin d'obtenir le corpus cible de messages à analyser en fonction de notre projet en Belgique [37].

Nombre de sms échangés en Belgique sept-déc 2004 :

$$\alpha \times \phi \div \delta = 363,5 \text{ (millions de sms)}$$

Proportion de l'échantillon sms4science belge :

$$30.000 \text{ (sms)} \div 363,5 \text{ (millions de sms)} \times 100 = 0,008\% \text{ (sms représentés)}$$

Le corpus réunionnais est composé de 12.660 messages pour une population ayant envoyé 220.900.000 sms (α) en 2008 [38] On va restreindre ce chiffre pour coïncider avec les 2,5 mois de collecte de sms à la Réunion ($\delta = 4,8$).

Nombre de sms échangés à la Réunion entre le 15 avril et le 30 juin 2008 :

$$\alpha \div \delta = 46,02 \text{ millions de sms}$$

Proportion de l'échantillon sms4science suisse francophone :

$$12.659 \text{ (sms)} \div 46,02 \text{ (millions de sms)} \times 100 = 0,027\% \text{ (sms représentés)}$$

Le corpus réunionnais représente donc une part plus conséquente des messages réellement échangés durant la période concernée.

35. http://www.lalibre.be/economie/actualite/article/221038/en-croissance.html
36. http://en.wikisource.org/wiki/CIA_World_Fact_Book,_2004/Belgium
37. Ce calcul, ainsi que les suivants, est à considérer de manière prudente. En effet, les pourcentages permettant d'identifier la population francophone et la durée de collecte sont fictifs au regard des messages envoyés : rien ne nous assure que la distribution des envois est équilibrée entre les différentes communautés linguistiques, ni que le taux d'envoi est régulier tout le long de l'année. De fait, on connait les pics d'envoi au moment du nouvel an ou d'autres grands événements.
38. Chiffre provenant des Observations de ARCEP : http://www.arcep.fr/index.php?id=9939#c15018.

Le corpus québécois est composé de 4.731 messages pour une population francophone du Canada ayant envoyé 5,07 milliards de sms durant 5 mois, en 2010. Ce chiffre est calculé à partir de données provenant de plusieurs sources. La première (une enquête des Rapports textuels pour le Canada de 2010[39]) indique que le nombre de sms envoyés en 2010 au Canada s'élève à 56,4 milliards. Ce chiffre doit ensuite être ramené à 21,6% pour correspondre à la population francophone du Canada (ϕ), selon notre seconde source (le CIA World Fact Book[40]). Nous obtenons alors 12,18 milliards de sms envoyés par des usagers du sms francophones en 2010. Enfin, on restreint ce nombre aux cinq mois de collecte (δ) afin d'obtenir le corpus cible de messages à analyser en fonction de notre projet au Canada.

Nombre de sms échangés au Canada nov 2009-avril 2010 :

$$\alpha \times \phi \div \delta = 5,07 \text{ (milliards de sms)}$$

Proportion de l'échantillon sms4science canadien francophone :

$$4.731 \text{ (sms)} \div 5,07 \text{ (milliards de sms)} \times 100 = 0,0001\% \text{(sms représentés)}$$

Enfin, le corpus suisse francophone est composé de 3.889 messages pour une population francophone de Suisse ayant envoyé 241,7 millions de sms durant 2,5 mois, en 2009. Ce chiffre est calculé à partir de données provenant de plusieurs sources : la première (une enquête de l'Office fédéral de la communication suisse de 2009[41]) indique que le nombre de sms envoyés en 2009 en Suisse s'élève à 5,67 milliards (α). Ce chiffre doit ensuite être ramené à 20,4% pour correspondre à la population francophone de Suisse (ϕ) en 2009, selon notre seconde source (U.S. Department of State[42]). Nous obtenons alors 1,16 milliard de sms envoyés par des usagers du sms francophones en 2009. Enfin, on restreint ce nombre aux deux mois et demi de collecte (δ) afin d'obtenir le corpus cible de messages à analyser en fonction de notre projet en Suisse.

Nombre de sms échangés en Suisse nov-dec 2009 (janvier 2010) :

$$\alpha \times \phi \div \delta = 241,7 \text{ (millions de sms)}$$

Proportion de l'échantillon sms4science suisse francophone :

$$3.889 \text{ (sms)} \div 241,7 \text{ (millions de sms)} \times 100 = 0,0016\% \text{(sms représentés)}$$

Au vu de ces pourcentages, à la fois faibles et inconstants d'une région à l'autre, les sous-corpus obtenus ne sont pas représentatifs de la totalité des messages échangés durant cette période dans les pays concernés.

39. Pour en savoir plus : http://www.txt.ca/english/consumer/press.html.

40. Chiffre provenant du recensement 2006 : https://www.cia.gov/library/publications/the-world-factbook/geos/ca.html.

41. Pour en savoir plus : http://www.bakom.admin.ch/dokumentation/zahlen/00744/00746/index.html?lang=fr.

42. http://www.state.gov/r/pa/ei/bgn/3431.htm.

6.1.4 Comparabilité des données

Nous nous efforçons de détailler la question de la comparabilité des données afin de cerner au mieux notre corpus et les analyses et conclusions possibles à partir des qualités intrinsèques de ce corpus. Selon L'Homme (2004, p. 133) « La notion de corpus comparables a fait son apparition assez récemment dans différentes communautés, notamment en lexicographie, en linguistique de corpus et en traitement automatique de la langue. D'une manière générale, les corpus comparables sont composés de deux ensembles de textes (ou davantage) qui possèdent des caractéristiques communes ». Ces caractéristiques appartiennent tant à la méthode d'obtention qu'au traitement des données. En effet, on estime qu'un but de collecte semblable, une méthodologie d'obtention et un traitement des données identiques peuvent mener à des corpus comparables. D'autres auteurs (Fung et Yee, 1998) introduisent de nouvelles définitions pour les corpus comparables ; ainsi, les textes constituant les corpus comparables peuvent :

1. Montrer la même fonction dans une situation de communication ;
2. Montrer le même thème, le même domaine ou le même type (ou sous-type) de discours [43] ;
3. Etre multilingues ou monolingues.

Nous avons expliqué précédemment que le but de la collecte sms4science était unique : obtenir des copies de messages envoyés dans des situations de communication véritables en vue de les étudier scientifiquement et « contribuer à l'étude de la communication par sms et du langage qu'elle véhicule » [44]. Nous avons également montré que la méthodologie de collecte des messages, même si elle présentait des spécificités par zone, suivait le même fil rouge (participation par volontariat, appel à participation par les médias, etc.). Après un travail de nettoyage, nous pouvons également assurer que nos corpus sont monolingues (francophones) [45] ; il s'agit d'un ensemble de textes appartenant au même genre (la communication par sms), mais non à la même situation de communication. Il existe une très grande diversité des situations de communication par sms : le nombre de participants, la vitesse (engendrée par le stress d'une situation), le lieu, etc. varient considérablement. Cette diversité des situations de communication constitue justement aussi un point commun entre tous les corpus.

6.1.4.1 Format, annotation et nettoyage

Nous présentons ici le seul point dont nous n'avons pas encore discuté : le traitement des données [46].

43. Au sens exprimé dans (Dubois *et al.*, 1994) : scientifique, pédagogique, etc.
44. Passages provenant de la présentation détaillée du projet.
45. Tout en assumant qu'ils incluent d'autres langues, comme nous l'avons signalé, par mixing ou switching.
46. Notre totale gratitude pour le travail de nettoyage et d'anonymisation des données va à Louis de Viron et au Dr Thomas François.

Les modalités de collectes (dispersion géographique, équipes de recherche différentes, etc.) ont rendu indispensable l'établissement d'un protocole précis pour chacune des étapes menant du stockage des messages et questionnaires reçus à l'étiquetage précis. Malgré des précautions dans la démarche initiée dès 2007 (nous présentions aux partenaires des lignes directrices pour le traitement, basées sur Fairon *et al.* 2006b), le traitement des données a largement différé, notamment en ce qui concerne l'anonymisation [47] et la transcription en graphie standard (cf. la partie 6.1.4.2).

Généralités La méthodologie de traitement à la Réunion, au Québec et en Suisse a pris des distances par rapport à la proposition initiale. Nous reprenons dans la Figure 6.4 les grandes étapes de traitement de corpus et les problèmes rencontrés lors de la réception des 4 corpus sms4science constitués par les équipes locales.

FIGURE 6.4 – *Résumé des problèmes généraux liés au traitement des données (sms4science)*

Etape	Point spécifique	Problème	Notre solution
Nettoyage	Doublons (messages envoyés à plusieurs reprises)	Relevés partout sauf au Québec	Repérage manuel et suppression
	Soudure multi-messages (messages dépassant 160 car.)	Quelques coquilles à la Réunion	Rectifiées manuellement
	Retours à la ligne intra-messages	Doivent être maintenus	Le maintien des retours engendre des erreurs lors du traitement des données. Remplacement par une balise.
	Encodage des caractères	Les corpus ont été fournis en des formats réellement différents avec encodage varié (parfois dans le même corpus)	Transcodage de tous les corpus en utf8
Anonymisation	Technique d'anonymisation	Suisse et Belgique : semi-manuelle (logiciels de reconnaissance automatique limitée à certains numéros, prénoms, etc.) Québec et La Réunion : manuelle	Suisse et Belgique : repassage manuel pour affinage. Québec, La Réunion : application de dictionnaires de prénoms et recherche de patterns d'adresses, etc. pour affiner
	Système d'étiquetage	Tous différents	Fusion et harmonisation (d'où une certaine perte d'information)

Trois précautions supplémentaires sont liées à l'anonymisation des corpus de sms. Premièrement, la suppression des numéros de téléphone associés aux messages. Nous

47. (Nous abordons uniquement ici les traitements sur données textuelles.

remplaçons ces numéros par des identifiants neutres. Plusieurs numéros pouvant être liés à un même auteur, on prendra garde de les relier au même identifiant. Deuxièmement, un certain pourcentage du corpus de messages a été envoyé par des usagers qui n'ont pas rempli de profils (certaines régions n'obligeant pas à remplir un profil pour participer au projet). La solution choisie est d'associer le message à un identifiant (au profil vide) afin d'associer les messages appartenant à un même auteur. Enfin, l'anonymisation suppose de neutraliser les dates et les heures d'envoi des messages. Or, ces informations sont utiles pour différents types de recherches :

– des recherches du type de celles réalisées dans Weiser *et al.* (2011) en vue de l'extraction d'informations temporelles pour leur insertion dans un agenda. Pour les expressions relatives telles que « demain », « mardi prochain », « dans une heure », etc. la date d'envoi est primordiale. La solution est de créer une date factice de manière aléatoire ;
– le repérage de « conversations ». Comme nous l'avons explicité précédemment, nous recevons uniquement des messages unilatéraux. Pourtant les recherches sur conversations intéressent un très grand nombre de scientifiques. On peut s'intéresser aux unilatéraux dans le cadre d'une conversation : on aura ainsi besoin, pour reconstituer l'ordre des événements sur une ligne temporelle chronologique, des dates et heures d'envoi de tous les messages. Cependant, les auteurs nous envoient des copies des messages réellement envoyés à d'autres utilisateurs, ce qui signifie que les informations temporelles associées à ces messages ne correspondent pas (forcément) à la temporalité de l'envoi réel du message (certains participants allaient rechercher d'anciens messages stockés dans leur dossier d'envoi). Nous n'avons pas encore trouvé de solution à ce problème [48].

Etiquetage pour l'anonymisation

Belgique L'étiquetage proposé et adopté par l'équipe belge est le suivant (Fairon *et al.*, 2006b) :

1. {???,.NOM} pour les noms, prénoms et (quelques) surnoms ;
2. {???,.EMAIL} pour les adresses électroniques personnelles ;
3. {???,.ADR} pour les adresses postales, ceci correspondant uniquement aux numéros de rue (ex. *2, rue haute 1348* devient *???,.ADR, rue haute 1348*) ;
4. {???,.TEL} pour les numéros de téléphone personnels (fixes ou portables) ;
5. {???,.COMPTE} pour les numéros de comptes bancaires ;
6. {???,.COM} pour les communications associées à des virements de comptes bancaires ;

48. Notons toutefois que l'équipe du projet sms4science à Montpellier (`www.sud4science.org`) a prévu une collecte de dialogues par sms pour 2012. Leur méthodologie permettrait, via une application pour téléphones intelligents, de recueillir des conversations enregistrées comme telles (chronologiquement) dans ces téléphones. Notons que cette méthode exclut les participants ne possédant pas de téléphones intelligents.

7. {???,.RECHARGE} pour les numéros permettant de recharger un téléphone portable en crédit d'appel ;
8. {???,.BLOG} pour les adresses de blogs personnels ;
9. {???,.SITEWEB} pour les adresses de site Internet ;
10. {???,.NUM} pour tout autre type de numéro

Limites de l'annotation belge :

1. Elle est non exhaustive : énormément de prénoms et de noms de famille ont été oubliés, malgré la vérification manuelle en raison de la transcription manuelle ; ce problème sera résolu par un deuxième passage semi-manuel ;
2. L'anonymisation des numéros de rue laisse transparaitre un trop grand nombre d'informations qui permettent l'identification de certains individus. Ainsi *2, rue haute 1348, la porte rouge en face du bar à kebab*, remplacé par *{???,.ADR}, rue haute 1348, la porte rouge en face du bar à kebab*, ne nous semble pas assurer une anonymisation suffisante. Ce problème sera résolu par un deuxième passage semi-manuel ;
3. Ce type d'étiquettes engendre une perte d'information du nombre de caractères, et d'information ethnique ou encore du genre de l'individu mentionné (dans le cas des prénoms et surnoms). Ce problème n'a pas été résolu par nous-même, puisque les informations ont été perdues dès la collecte belge ; les étiquettes des autres corpus doivent donc être alignées aux belges.

La Réunion L'étiquetage proposé et adopté par l'équipe réunionnaise est le suivant :

1. {NOM} pour les noms de famille ;
2. {PRENOM} pour les prénoms et (quelques) surnoms ;
3. {MAIL} ou {mail} (non uniforme) pour les adresses électroniques personnelles ;
4. {adresse} pour les adresses postales ;
5. {TEL} ou {tél} (non uniforme) pour les numéros de téléphone personnels (fixes ou portables) ;
6. {site} pour les adresses de site Internet ;
7. {num} pour toute information numérotée ;
8. Les étiquettes {Guignard}, {Jordan}, {x}, {NOM.com} et {NOM1} ont également été trouvées dans le corpus. {NOM.com} a été compris comme site Internet, {NOM1} comme un nom, {Guignard} et {Jordan} comme des prénoms. L'étiquette {x} a été supprimée.

Limites de l'annotation réunionnaise :

1. Les étiquettes manquent d'uniformité (mail, tel, site) ;

2. Ce système d'étiquettes perd l'information du nombre de caractères, et l'information ethnique et du genre de l'individu mentionné (dans le cas des prénoms et surnoms). Ce problème ne sera pas résolu, les informations étant perdues dès ce niveau.

Québec L'équipe québécoise n'a adopté aucune balise d'anonymisation. Nous avons donc effectué une anonymisation semi-automatique en suivant l'harmonisation des balises, telle que nous la présentons plus tard dans cette partie. Il y avait un certain nombre de doublons que nous avons demandé à l'équipe québécoise de supprimer *a posteriori*.

Suisse L'équipe suisse est celle qui a pris le plus de liberté (Stähli *et al.*, 2011) par rapport aux recommandations belges (Fairon *et al.*, 2006b). Ceci est notamment dû à la spécialisation et à la compétence des personnes enrôlées. L'étiquetage proposé et adopté par l'équipe suisse est le suivant :

1. Les prénoms : une liste de prénoms préétablie dans chaque langue a été appliquée à l'ensemble du corpus. Quelques tests prouvent que 95% des prénoms ont été anonymisés. L'originalité du traitement suisse tient au choix de la méthode d'anonymisation. Contrairement aux autres équipes réunionnaise et belge, la Suisse n'a pas remplacé les prénoms de son corpus par des étiquettes abstraites de type [FirstName]. Ils ont opté pour une rotation des prénoms ; ceci signifie que les véritables prénoms n'ont pas été réellement éliminés du corpus, mais qu'ils ont été déplacés. Ainsi, chaque occurrence de *Marie* a été remplacée par un autre prénom féminin tel que *Anne*. Ceci a plusieurs avantages : on conserve les informations de genre, on obtient un texte plus facile à la lecture qu'avec des étiquettes, et des conversations peuvent être plus aisément identifiées puisque le destinataire est reconnu, même s'il l'est sous une identité différente ;
2. Les noms de famille : peu de noms de famille sont présents dans le corpus. Puisque les noms de famille sont plus facilement identifiables que les prénoms, l'équipe suisse a décidé de remplacer les noms de famille de leur corpus par l'étiquette [LastName]. Ici aussi, les tests ont prouvé que la méthode semi-automatique de reconnaissance des noms de famille a permis d'anonymiser 95% de ceux-ci ;
3. Les nombres : afin d'anonymiser massivement toutes les informations téléphonique, bancaire, etc. sous forme de nombres, l'équipe locale a décidé de remplacer chaque chiffre du corpus appartenant à un ensemble de 3 chiffres ou plus, par l'étiquette N. Ainsi 0477 serait remplacé par NNNN. Les tests prouvent une reconnaissance des patterns de 100%. Cette méthode permet de ne pas perdre l'information du nombre de caractères ;
4. Les adresses postales : celles-ci ont été anonymisées et remplacées par l'étiquette [StreetAddress]. Ceci ne couvre que le nom de la rue lui-même, étant donné que les numéros de maison (s'ils font trois chiffres ou plus) et de codes

postaux sont anonymisés, selon la méthode présentée plus haut. Les noms de ville ont été maintenus dans le corpus anonymisé puisqu'ils ne sont pas considérés comme des informations privées [49] et qu'ils sont d'une importance sémantique non négligeable dans les messages. Ainsi, 3 Place des Carmes à 76390 Rouen devient 3 [StreetAddress] à NNNNN Rouen ;

5. Les adresses électroniques : toutes les adresses ont été supprimées et remplacées par une étiquette suivant le pattern xxx@yyy.ch, où chaque lettre précédant l'arobase a été remplacée par x et chaque lettre la suivant, par y ;
6. Les sites Internet : ils ont été maintenus puisqu'ils présentent des contenus publiquement accessibles.

Limites de l'annotation suisse :

1. Le système de rotation des prénoms est intéressant, mais ne permet de maintenir ni l'information ethnique, ni l'information en termes de nombre de caractères ;
2. Les expressions temporelles sont souvent abrégées dans les sms et, bien qu'il soit plus courant de noter 16h15 ou 16.15, il arrive aussi que les heures soient notées 1615 dans un contexte de type *J'arrive à 1615*. Dans le cas de la Suisse, cela engendre *J'arrive à NNNN*. Il s'avère que le nombre d'occurrences d'un pattern aussi abrégé est non négligeable dans le corpus suisse ; aussi, il y est impossible de réaliser des recherches sur l'extraction d'informations temporelles en vue de les insérer dans un agenda, comme expliqué au point 6.1.4.1, sans engendrer des biais puisqu'une part non négligeable des données serait omise. On notera encore d'autres contextes où cette erreur se produit, tel que dans l'exemple suivant, où l'anonymisation a été appliquée sur une expression du type *dans 10000 ans* :

   (6) Je peux peut-être vite avoir la voiture pour passer chez toi pendant que mon papa attend mon oncle...Par contre j'ai juste aucune idée de par où passer, donc je risque d'arriver dans NNNNN ans :-)

3. Les adresses de sites Internet maintenues peuvent faire référence à des sites personnels qui permettent l'identification du destinateur, du destinataire ou d'un tiers, comme dans le message *vas voir mon site perso www.siteperso.com.*

Harmonisation de l'étiquetage Etant donné l'ampleur de la variation des annotations et du nettoyage des corpus telle qu'elle vient d'être présentée, une harmonisation était nécessaire avant l'exploitation des données. Cette harmonisation est le résultat d'un travail de groupe tant au niveau de la réflexion qu'au niveau du traitement des données [50]. Le travail a abouti à un protocole de nettoyage et d'anonymisation, destiné aux équipes locales, que nous détaillons ici et qui reprend,

49. Elles peuvent permettre l'identification des personnes dans quelques rares contextes contenant d'autres informations. Ces quelques cas doivent être anonymisés manuellement.

50. Ce travail n'aurait jamais pu se faire sans la précieuse collaboration des Drs Patrick Watrin et Thomas François, de Hubert Naets, et surtout, de Louis de Viron.

pour la partie *Anonymisation*, les choix d'harmonisation que nous avons réalisés par rapport à nos 4 corpus francophones [51].

Nettoyage

1. Reconstitution des messages en plusieurs parties : un certain nombre d'utilisateurs rédigent des sms dépassant la limite de 160 caractères. Dans quelques cas, le message est découpé en autant de sms que nécessaire par l'opérateur téléphonique. Chaque morceau est ainsi annoté à l'aide d'un code permettant de l'assembler avec les autres morceaux. Avant la collecte, au moment des tests, il faut vérifier que les messages entiers sont récupérés ; ou bien, *a posteriori*, au moment de la récupération du corpus final, il faut penser à demander la soudure à l'opérateur. Un autre moyen, plus laborieux, est de recoller manuellement les messages une fois collectés ;
2. Retours à la ligne : seule l'équipe suisse n'a pas respecté le maintien des retours à la ligne au sein des messages. Cette information est donc perdue. Dans les autres régions, nous les avons remplacés par une étiquette {return} afin d'éviter des problèmes de traitement automatique (y compris statistique) des corpus ;
3. Élimination des doublons : une procédure automatique permet d'éliminer les doublons dans le cas de messages strictement identiques envoyés depuis le même numéro de téléphone. Cette étape peut fausser légèrement les résultats si quelqu'un a envoyé volontairement deux fois un sms comme : *Oui !*. Nous pensons toutefois, et de façon subjective, que cette perte est minimale puisqu'elle ne concerne que des messages triviaux et qu'elle n'occulte aucun phénomène linguistique (excepté dans le cadre d'une étude statistique sur les thématiques, par exemple). Notons qu'il peut malgré cela rester des doublons dans le corpus. Cela est dû au fait que la normalisation rend identiques des messages qui étaient différents. Ainsi, les deux messages : *Mon e-mail est machin@blabla.org* et *Mon e-mail est bidule@blalbla.be* auront tous les deux été transformés en *Mon e-mail est {MAIL}*. Nous prenons le parti de garder ces informations redondantes. Notons que le travail de suppression des doublons devra donc précéder la procédure d'anonymisation ;
4. Suppression des messages envoyés aux participants du projet : à la suite de l'inscription d'un participant, ce dernier reçoit un sms de confirmation de type *Merci de participer à notre enquête. Veuillez visiter notre site et envoyer une copie de vos messages au 3232*. Ce message d'information est parfois lui-même renvoyé aux organisateurs. Nous prenons évidemment le parti de les supprimer ;
5. Élimination de sms adressés à l'équipe du projet : dans certains messages, les participants s'adressent aux organisateurs du projet pour leur poser des questions, donner leur avis sur le projet, etc. Ces messages sont systématiquement

51. Notons, pour la partie Nettoyage, qu'ils sont particulièrement inspirés de Fairon *et al.* (2006b)

éliminés. En revanche, il faut se garder de supprimer les messages qui ne sont pas clairement adressés aux organisateurs (c'est-à-dire les messages ambigus comme *Je n'ai pas d'adresse email ; comment je fais ?*) ;

6. Élimination des « Chaines de l'amitié » : à l'instar du courrier électronique, le sms est devenu la cible des messages en chaine à renvoyer à plusieurs personnes. Tous ces messages sont supprimés puisqu'il ne s'agit pas de communication spontanée ;
7. Suppression des sms promotionnels et de sonnerie : une possibilité d'identification des sms promotionnels est de rechercher automatiquement tous les sms contenant certains mots-clés, tels que *envoyez*, *pay*, *service*, *recharge* ou *sonnerie*, et de les trier manuellement ;
8. Problèmes d'encodage : toutes les équipes n'utilisaient pas le même encodage pour leurs sous-corpus. Nous avons donc transcodé tous les corpus en UTF-8 afin d'obtenir un ensemble cohérent, ce qui a impliqué un certain nombre de vérifications manuelles ;
9. Mise à l'écart des sms non francophones : cela dépend de la région et de l'objectif des organisateurs. En Belgique, le corpus contenait 714 sms intégralement rédigés dans une langue autre que le français (33 % wallon, 30 % anglais, 13 % néerlandais, 12 % espagnol, 8 % langue non identifiée, 4 % divers). Ces sms ont été écartés du corpus. En Suisse, au contraire, plusieurs langues étaient visées et ont été classées en sous-corpus. Une politique de tri doit néanmoins être déterminée avant le nettoyage des données ;
10. Sms graphiques : les messages contenant des jeux graphiques peuvent être écartés du corpus. Dans la plupart des cas, il s'agit de l'insertion d'un dessin réalisé en combinant des caractères (phénomènes ASCII), mais on trouve également des jeux dans lesquels le destinataire du message doit faire défiler celui-ci pour le lire jusqu'au bout. Il est peu pertinent de les laisser dans le corpus général, étant donné que la plupart des recherches visées par le projet portent sur la langue. Mais il est important de les garder dans un corpus séparé au cas où des recherches portant sur l'aspect graphique, la communication en général, etc. surviendraient.

Anonymisation

1. Suppression des numéros de téléphone des expéditeurs : il s'agit de remplacer les numéros de téléphone associés à chaque message par un identifiant neutre. Il faut conserver la table de correspondance numéro de téléphone - identifiant (en cas de problème futur). La syntaxe de l'identifiant est la suivante : RE_XXX où RE est la région ou le pays où la collecte a été effectuée et XXX est le numéro de profil sociodémographique. Exemple : CH_230 (Suisse), BE_1785 (Belgique), etc. Chaque identifiant unique liera donc tous les messages d'un même utilisateur et, dans le cas d'un auteur ayant répondu au questionnaire, il reliera les messages au profil ;

2. Anonymisation des données personnelles dans les messages : toutes les informations permettant d'identifier les auteurs des messages doivent être supprimées et remplacées par des balises neutres. Ce choix engendre une perte d'information, mais l'hétérogénéité des étiquetages des 4 premiers corpus ne permet pas de garder plus d'information. Les données concernées sont les suivantes :
   (a) prénoms, noms propres ou surnoms trop particuliers : {NPERS} ;
   (b) adresses e-mail : {MAIL} ;
   (c) numéros d'habitation, rue, code postal et/ou ville : {ADRESSE} ;
   (d) numéros de téléphone : {TEL} ;
   (e) numéros de compte bancaire, numéros de communication bancaire, numéros de recharge de téléphone mobile et divers numéros : {NUM} ;
   (f) photographies personnelles : {IMAGE} ;
   (g) adresses de blogs et sites web personnels[52] : {SITEWEB} ;

   Dans certains cas, nous devons effacer quelques données supplémentaires. Dans le doute, notamment en cas d'ambigüité entre un nom propre et un mot inconnu, nous décidons d'anonymiser.

   Comme nous l'avons expliqué, cet étiquetage anonymise mais ne permet pas de conserver les informations concernant le nombre de caractères (utiles notamment lorsque l'on fait des études statistiques) ni les informations concernant le référent (par exemple la connotation socio-ethnique d'un prénom). Or, cette problématique est importante. Nous conseillons aux futures équipes locales de posséder deux versions d'anonymisation : la première emprunterait une des solutions suivantes :
   (a) Le prénom est remplacé par un nombre de symboles égal au nombre de caractères. *Abdel* devient *XXXXX* (pour garder des informations de type : longueur des messages) ;
   (b) Le prénom est remplacé par un nom ayant la même connotation ethno-sociologique[53]. *Abdel* devient *Khaled* ;
   (c) Le prénom est remplacé par un nom commençant par la même lettre (au cas où il y aurait des appels de cette personne avec abréviation à un autre endroit du message). *Abdel* devient *Antoine* ;
   (d) Une combinaison des 3 précédentes : respect du nombre de caractères, de la connotation et de la première lettre : *Abdel* deviendrait *Ahmed*. Cette solution risque d'être assez couteuse en termes de travail et n'est peut-être réalisable que pour les corpus de taille restreinte.

52. Les adresses de sites non personnels seront laissées telles quelles.

53. Cette connotation étant assez subjective et liée aux représentations de l'annotateur, nous conseillons à celui-ci de définir, avant de commencer, des règles d'anonymisation auxquelles il pourra se référer lui-même et qu'il pourra communiquer ultérieurement pour préciser son système d'anonymisation.

La deuxième version du corpus anonymisé suivrait les règles générales admises pour tous les corpus sms4science (telles que présentées précédemment) et serait la version homogénéisée pour le corpus commun.

6.1.4.2 Alignement : transcription en graphie standard

Corpus parallèles et alignés La littérature relative à la linguistique de corpus a tendance à confondre corpus alignés et parallèles. Ainsi, selon Do (2011, p. 1), « Un corpus parallèle se compose de textes bilingues alignés au niveau des phrases ». Les corpus parallèles sont pourtant uniquement des corpus composés d'un même texte traduits (ou obtenus) en plusieurs langues et mis en vis-à-vis (Guidère, 2008, p. 94). Il existe ensuite la possibilité de les aligner, c'est-à-dire de mettre en relation les unités correspondantes de chaque version de langue. Nous montrons dans cette partie en quoi notre corpus de sms peut-etre considéré comme un ensemble de sous-corpus parallèles et en quoi la question de l'alignement est présente dans ce travail.

Même s'il n'existe pas à proprement parler de « traduction » des messages d'une langue A à une langue B, il y a, dans le corpus de sms, une transcription [54] d'un système graphique A à un système graphique standardisé du français (B). En ce sens, il semble déjà que nous soyons en présence de corpus dits « parallèles ».

Cette considération nous amène à préciser la méthode de transcription en graphie standard. A ce jour, il existe deux possibilités de transcription des corpus de sms : manuelle ou semi-automatique. La méthode manuelle consiste à transcrire manuellement (de préférence par une seule personne, pour assurer l'homogénéité de la transcription) chaque sms en son correspondant en graphie standard. Cette méthode présente deux inconvénients : elle est couteuse en termes de ressources humaines et de temps et dépend de la constance du transcripteur. De plus, elle entraine les coquilles inévitables, associées à tout travail manuel. La méthode semi-automatique permet d'accélérer considérablement le travail de transcription et réduit également le nombre de coquilles possibles. L'automatisation de la transcription repose sur l'apprentissage à partir d'une transcription manuelle précédente (Beaufort *et al.*, 2010). Ainsi, comme l'expliquent Habert *et al.* (1997, p. 137), « le point de départ n'est pas une formalisation de deux langues et de leur mise en correspondance, mais la réutilisation des traductions existantes produites par des traducteurs humains ». L'inconvénient d'un tel système pour la constitution d'un corpus linguistiquement valide est qu'il fonctionne parfaitement pour le corpus sur lequel il a été entrainé, mais moins sur les autres. Plusieurs facteurs peuvent expliquer cette limitation :

1. Le lexique diffère d'un corpus à l'autre et donc du corpus d'entrainement au corpus suivant (le système fonctionnant notamment sur la fréquence) ; on a par exemple *vendr*, qui est entrainé à être transcrit en *vendredi*, et qui n'autorise

54. La notion de transcription est à prendre dans le même sens que celle utilisée pour l'oral, par exemple dans Ochs (1979).

pas la transcription en *vendre*. C'est également le cas de *km* entrainé à une transcription en *comme*, là où on attendrait simplement l'unité de distance ;

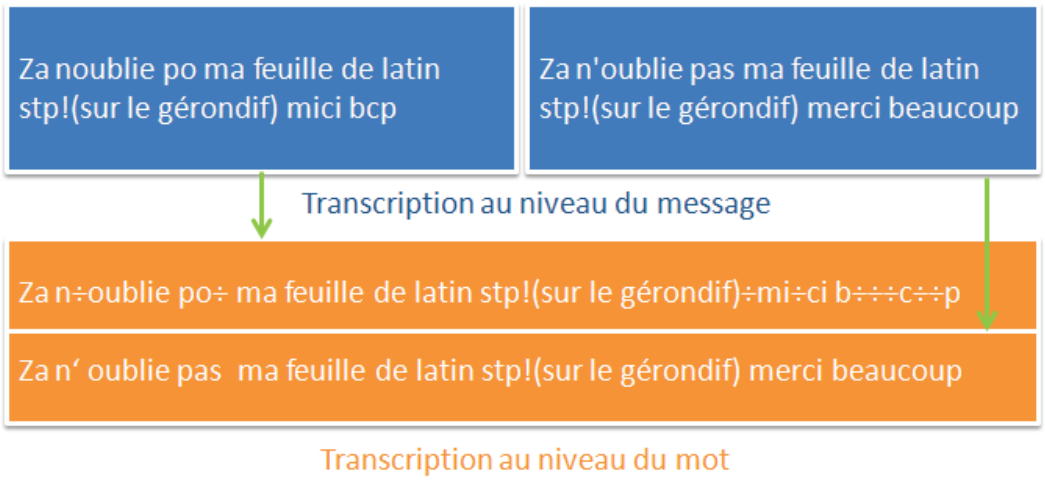
2. La proportion de lexique étranger mélangé au français (notamment dans des régions plurilingues) peut rendre hermétique la compréhension du texte transcrit de façon inadéquate par la machine.

Ainsi, un travail manuel de relecture et de correction est indispensable et a effectivement lieu dans le cas d'une transcription semi-automatique [55]. Notons que dans la situation d'un mélange de langues très récurrent, comme c'est le cas à la Réunion, la correction du texte transcrit sera plus conséquente en termes de temps que sa transcription directe à partir de la version brute. Une transcription semi-automatique devient donc inutile.

Les deux types de transcription offrent des corpus parallèles et alignés. Il y a une mise en correspondance des unités du texte ou « segments de grain équivalent ». La mise en correspondance des unités (ou composantes formelles) de texte peut se faire à partir d'un message, d'un paragraphe, d'une phrase, d'un mot ou encore d'un caractère. Au sein du projet sms4science, la mise en relation peut se situer à deux niveaux de segments différents : le message ou le mot. Les transcriptions manuelles aboutissent toutes, pour l'instant, à une transcription au message près.

La transcription semi-automatique permet, pour sa part, une correspondance – *a posteriori* [56] - au caractère près. La Figure 6.5 présente les deux méthodes.

FIGURE 6.5 – *Les deux niveaux de transcription des sms en graphie standard*



De nombreux auteurs ont relevé « les difficultés de la mise en correspondance automatique » (Habert *et al.*, 1997, p. 136) des segments dans les corpus parallèles ;

55. Notons qu'il est également utile, afin de perfectionner la transcription, de réaliser un travail de relecture sur un corpus transcrit manuellement. Cette étape est toutefois moins indispensable que dans le cas de la transcription automatique.

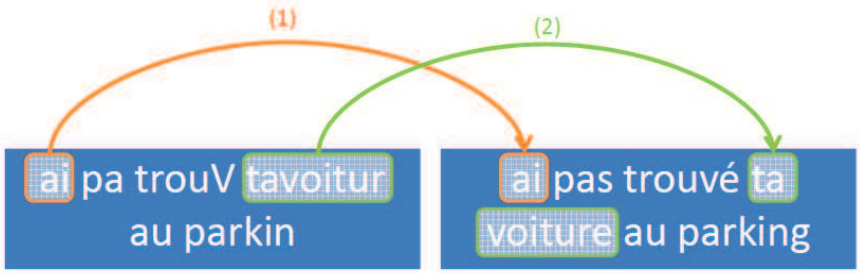
56. Grâce à un système élaboré par Beaufort *et al.* (2010)

l'alignement d'un texte (ou même d'un enregistrement sonore) et de sa transcription en graphie standard pose des problèmes distincts de l'alignement entre plusieurs textes de langues différentes, qui sont d'ordinaire :

- une phrase en L1 peut donner deux phrases en L2 et inversement ;
- des constructions plus complexes dans une phrase d'une langue donnée que dans l'autre ;
- les métaphores propres à une langue.

Puisqu'il s'agit chez nous de la même langue dans les deux corpus, il n'existe pas de problèmes de métaphores ou de nombre de phrases. Par contre, on rencontre dans les sms un problème différent : une unité lexicale peut donner deux unités (*allerviens* pour *aller viens*) et inversement (*t'es éner V* pour *t'es énervé*). La Figure 6.6 illustre la politique de transcription (cf. 6.1.4.2) : elle a pour but de ne pas ajouter d'information, ni au niveau typographique (un manque de ponctuation n'est pas comblé) [57], ni au niveau syntaxique (1). La nouvelle difficulté engendrée au niveau de la frontière de l'unité lexicale (unités agglutinées) est également illustrée dans la Figure (2).

FIGURE 6.6 – *Politique et limites de transcription des sms en graphie standard*



La transcription manuelle engendre de la variation parce qu'elle est liée à la subjectivité de l'interprétation des formes brutes. En effet, comme l'expliquent Dister *et al.* (2009), « transcrire est une activité sélective et interprétative, qui engage une théorie et oblige à poser des choix » (ex. *biz* pour *bisou* ou pour *bise*).

Afin de limiter cette variation de la transcription, un protocole de transcription a été élaboré dans chaque région. Nous détaillons ces protocoles de transcription par région et présentons ensuite l'harmonisation de cette transcription, que nous avons entreprise afin d'obtenir des corpus comparables.

Belgique La transcription belge des sms bruts en graphie standard est décrite dans Fairon *et al.* (2006b). Elle a largement - mais pas tout à fait - inspiré l'harmonisation des transcriptions qui sera décrite ci-après. La transcription manuelle s'est accompagnée d'un alignement au niveau du message. Une relecture de cette transcription a ensuite eu lieu. Une partie originale de la transcription belge concerne

57. Et donc de ne pas donner deux phrases à partir d'une phrase.

l'insertion d'étiquettes pour compléter l'information syntaxique. En effet, la politique de transcription du projet sms4science, comme expliqué précédemment, veut que le transcripteur n'intervienne qu'au niveau morphologique afin de compléter les informations manquantes pour former des unités lexicales. La Belgique a elle-même recommandé cette règle, mais elle a également proposé une version différente de la transcription incluant des étiquettes syntaxiques là où il y avait des insuffisances. Ces étiquettes suivent la syntaxe proposée par le Delaf[58]. On aura donc un corpus belge comprenant deux versions de transcription, tel que l'illustre la Figure 6.7.

FIGURE 6.7 – *Double transcription adoptée par l'équipe belge du projet sms4science*

Salut, on à fai ou pas le chap 31? J'croi po ms sui po sure... Rép stp	Salut, on a fait ou pas le chapitre 31? Je {ne,.ADV+MISS} crois pas mais {je,.PRO+MISS:1s} {ne,.ADV+MISS} suis pas sûre... Réponds stp	Salut, on a fait ou pas le chapitre 31? Je crois pas mais suis pas sûre... Réponds stp

Une deuxième particularité caractérise le corpus belge ; il s'agit des règles suivies en matière de ponctuation. Comme nous le verrons, nous avons entrepris de normaliser la ponctuation en respectant la règle des espaces associée à l'usage de la ponctuation en langue française. La Belgique ne réintroduit pas ces espaces, là où elles sont absentes.

Enfin, puisque nous abordons la notion d'alignement, notons que le corpus belge est, pour l'instant, le seul qui soit aligné au caractère près suivant la méthode de Beaufort *et al.* (2010). Ce second alignement permet ainsi des recherches plus variées que celles se basant uniquement sur les corpus alignés au message près (par ex. analyser la confusion de l'infinitif et du participe passé en extrayant toutes les formes verbales en *er* qui ont été remplacées par des *é* après transcription).

La Réunion Comme nous l'avons vu précédemment, l'équipe réunionnaise du projet a choisi de respecter autant les messages en créole que ceux en français. Ainsi, la transcription en graphie standard[59] a-t-elle dû être appliquée aux deux langues. Cette transcription est réalisée au message près. La Réunion a également choisi d'insérer dans les messages des étiquettes pour compléter l'information syntaxique. Contrairement à la Belgique, ils n'ont cependant pas offert de version sans étiquettes. Nous avons entrepris, en Belgique, de les repérer et les éliminer automatiquement. Enfin, l'équipe réunionnaise a été confrontée à la transcription de plusieurs langues,

58. Dictionnaire électronique du français du LADL, initié par Maurice Gross (Courtois, 1990).

59. Cette transcription a été réalisée en intégralité par Gauthier Caron et revue par la maitre de conférence Gudrun Ledegen, de l'université de la Réunion (aujourd'hui maitre de conférence à l'Université de Rennes).

dont le créole réunionnais, qui a quatre graphies standards distinctes [60]. Il s'agissait donc de choisir parmi ces dernières, en sachant que le choix influencerait tout le processus de transcription. C'est la graphie 83 (au sein de laquelle les lettres *k*, *w* et *z* sont très présentes) qui a été élue, parce que, selon l'équipe scientifique locale, elle se calque sur la phonétique du mot et semble donc plus simple à la lecture. On peut se demander également si ce choix n'est pas basé sur la ressemblance de la graphie 83 (l'usage inhabituel des lettres *k*, *w* et *z*) avec l'écrit sms (phonétisation par *k* et *z*). L'équipe locale ne pense pas avoir été influencée par la graphie en usage dans les sms, mais avoue qu'il était intéressant de constater, au moment de la transcription, que pour certains items, ils hésitaient quant à leur appartenance : phénomène de l'écrit sms ou créole ? En effet, il est légitime de se demander où se situe la frontière entre les deux pratiques. Pourrait-on penser que les deux pratiques participent en réalité du même phénomène : une volonté de marquer des traits de l'oral à l'écrit ? C'est en tout cas ce que semble penser l'équipe réunionnaise [61].

Québec La transcription québécoise a été réalisée par deux étudiantes de licence en traduction [62]. Elle s'est écartée, en quelques points, du protocole général, par exemple pour ce qui est de la ponctuation (insertion d'espaces blancs là où la norme n'en demande pas, non insertion là où elle en demande, etc.). Pour ce qui est de ces points de ponctuation, elle a donc hérité de quelques corrections ultérieures par l'équipe belge.

La seconde limitation de cette transcription concerne les trous syntaxiques. En effet, l'équipe québécoise a pris le parti de combler quelques absences telles que les pronoms sujets, le *ne* de négation, etc. Ces quelques rectifications sur corpus transcrit limitent certains pans de notre analyse (cf. partie 10.2) pour lesquels le corpus québécois ne peut être utilisé.

Suisse Le corpus suisse est le seul qui n'a pas été transcrit par l'équipe locale à l'heure où nous rédigeons cette dissertation. Ainsi, nous avons entrepris de le transcrire nous-même. Pour éviter les biais engendrés par une méconnaissance des particularités linguistiques suisses, nous avons décidé d'engager une étudiante suisse francophone pour traiter le corpus [63]. La méthode est toutefois semi-automatique puisque le travail s'est basé sur une transcription automatique du corpus (à partir de l'outil entrainé sur le corpus belge) corrigée manuellement.

60. Les trois premières tiennent leur nom de leur année de création : la graphie 77 (lékritir), la graphie 83 (kwz) et la graphie 2001 (tangol). On nomme la dernière la graphie *étymologique* ; celle-ci se calque sur la graphie du français. Notons qu'aucune des variantes ne s'est imposée dans la réalité des pratiques réunionnaises.

61. Ces conclusions proviennent d'une discussion menée avec Gauthier Caron en 2011.

62. Nous remercions Amélie Paulus et Eugénie Rompré Brodeur pour ce travail.

63. Nous remercions Arianne Torné pour la qualité du travail réalisé.

Harmonisation des transcriptions La notion de français « correct » ou « standard » semble maitrisée par les locuteurs et ancrée dans leurs représentations linguistiques [64]. Pourtant, cette notion est, scientifiquement parlant, très vague pour la simple raison qu'il n'y a pas *un* mais *des* français qui varient, selon de nombreux critères (la région où est pratiquée la langue, le médium oral ou écrit, etc.). Dans le contexte de la transcription des sms, il est encore plus difficile de fixer une norme, car certains phénomènes, comme nous l'avons vu précédemment, sont propres à la Cémo et posent des questions nouvelles. La meilleure façon de procéder est donc d'établir un protocole qui traite les problèmes au cas par cas, en proposant, pour chacun, une solution la plus proche possible du texte original.

Quelques grands principes guident la transcription en graphie standard que nous proposons en vue de l'obtention de corpus tels que ceux exploités dans le cadre de cette dissertation. L'idée générale de la transcription est de **respecter au mieux les messages d'origine pour ce qui est de leur lexique et de leur syntaxe**. **En ce qui concerne l'orthographe, on essayera au contraire de tendre vers une norme graphique**. La transcription dans le cadre du projet sms4science concerne donc bien (quasi-)uniquement la graphie. Afin d'obtenir une transcription homogène et de qualité, il est bon de trouver un ou deux membres de l'équipe qui assureront l'intégralité du travail, plutôt que de répartir la tâche entre tous les membres de l'équipe. Ensuite, et dans la mesure du possible, les transcriptions sont relues par une tierce personne. Bien que de très nombreuses règles soient fixées au départ, certaines apparaissent au fur et à mesure que les transcripteurs rencontrent de nouveaux problèmes ; celles-ci doivent donc être consciencieusement notées et enregistrées.

La question de l'harmonisation du lexique doit être précisée. Un certain nombre de points mérite un détail de la méthodologie suivie :

1. Contact des langues dans un message : les unités lexicales qui ne sont pas en français ne doivent pas être traduites. En revanche, on s'efforcera de les orthographier correctement dans la mesure du possible. Ex. *Helo mi amor commo stas ? J'arrive dans 1h en tr1, bisou* sera transcrit en *Hello mi amor como estas ? J'arrive dans 1h en train, bisou* ;
2. Régionalismes : on laissera la forme telle quelle (qu'on corrigera le cas échéant et s'il existe une orthographe de référence) ;
3. Néologismes, verlan, argot : on conservera la forme originale, qu'on corrigera. Ex. *g rien catche* deviendra *J'ai rien catché* ;
4. Formes ambigües : une forme telle que *bis* peut se transcrire en *bisou*, *bise*, *bises*, etc. Dans la même logique de non-interprétation personnelle du transcripteur, on laissera telle quelle l'expression d'origine *bis*.

64. Compris comme « les savoirs sociaux partagés dont les locuteurs disposent quant à leur langue » (Petitjean, 2009), ou encore quant à leurs pratiques.

Le niveau typographique concerne :

1. La ponctuation : on rétablira l'usage correct du trait d'union, de l'apostrophe et des espaces avant et/ou après les ponctuations[65]. Ex. *Où vas tu ?* deviendra *Où vas-tu ?* Par contre, on n'introduira pas de nouvelle ponctuation, pour éviter toute interprétation subjective du transcripteur. Ex. *Où vas tu moi je vais à la mer* ne doit pas devenir *où vas-tu ? Moi je vais à la mer* ;
2. Les émoticônes : les émoticônes seront préservés dans les messages, sauf s'ils sont clairement utilisés comme équivalents de mots (lorsqu'ils occupent une fonction dans la phrase). Dans ce cas, on remplacera l'émoticône par le mot qu'il représente. Ex. *si c'est comme ça, je te :-p* deviendra *si c'est comme ça je te tire la langue* ;
3. Les symboles mathématiques : on laisse tels quels les symboles mathématiques quand ils sont justement utilisés pour leur valeur mathématique (y compris au sens large, comme dans *Rodolphe+Emma=Amour*) ou dans les émoticônes. Dans les autres cas, on les remplacera par leur équivalent en lettres. Ex. *on se voit + tard* devient *on se voit plus tard* ;
4. Les nombres : à l'inverse des nombres apparaissant dans des expressions (par ex. dans *à 2main*), qui seront retranscrits en toutes lettres, les nombres qui représentent bien une quantité numérique seront laissés tels quels. Si l'on doute, on essayera de remplacer le nombre par un autre nombre. Ainsi, *vi1 1 peu* sera remplacé par *viens un peu*, car on ne pourrait pas avoir *vi2 2 peu* ;
5. Le format des heures : on harmonisera les heures au format de type *12h45*. Par ex. *rdv a 1245* deviendra rendez-vous à *12h45*. Cependant, si l'heure est au format *12h*, on n'ajoutera pas de précision en mettant *12h00* ; on maintiendra la forme *12h* ;
6. Les ponctuations, symboles ou suites de lettres incompréhensibles : on ne modifiera rien. Par ex. *C'est pr ça**** devient *C'est pour ça**** et *on va à la gffr* reste tel quel ;
7. le signe & : quand le signe & est utilisé pour dire *et*, on le remplacera par *et* dans la transcription. Sinon, on le laissera tel quel. S'il est collé à un mot à gauche et/ou à droite, on mettra un espace entre le signe et les mots qui l'entourent. Ex. *il y a mon père&ma mère* devient *il y a mon père et ma mère* ;
8. La casse : on respecte la casse traditionnelle. Les messages intégralement en majuscules seront transformés en minuscules, avec majuscule en début de phrase et à l'initiale des noms propres. On considérera comme début de phrase soit le début du message, soit ce qui suit un point, un point d'exclamation ou d'interrogation. On considérera également comme début de phrase les cas où l'auteur du message a lui-même marqué le début de phrase par une majuscule. Attention : les mots écrits en majuscules pour des raisons emphatiques seront

65. Notons que le corpus belge ne respecte pas cette règle ni la règle inverse. L'insertion des espaces liés à la ponctuation a été réalisée de manière aléatoire.

laissés en majuscules. Ainsi, *chez léon* deviendra *chez Léon* (restaurant bruxellois) et *COMT TU T'APL* deviendra *Comment tu t'appelles*. Mais *Te voilà oncle ET parrain* se transformera par contre en *Te voilà oncle ET parrain* ;

9. La mise en page : on respectera la mise en page du message d'origine, tant pour les espacements et les tabulations que pour les retours à la ligne. Les retours à la ligne seront toutefois remplacés par un tag {return} au cas où ils engendrent des problèmes techniques.

La morphologie est probablement le point le plus important de cette transcription.

1. Abrégements et abréviations : si l'abrégement (ou la troncation) est lexicographié (c.-à-d. s'il occupe une entrée dans un dictionnaire de référence[66]), comme c'est le cas avec *ciné*, on laissera le mot tel quel. En revanche, s'il s'agit d'une abréviation (non ambigüe), qui n'est pas usuelle, on remplacera par l'expression française complète. Ainsi *Alors, pour vend. c'est ok ?* devient *Alors, pour vendredi c'est ok ?* S'il est difficile de décider si l'abrègement est usuel ou non, on laissera l'expression telle quelle. Une attention particulière sera également portée aux abrègements qui n'ont pas l'air ambigus, mais qui le sont pourtant. C'est le cas de *bio*, par exemple, qui peut signifier notamment *biologie*, *biologique* ou *biochimie*. On maintiendra alors *bio* dans le texte ;
2. Acronymes, sigles et abréviations : les acronymes, sigles propres aux sms (*mdr*, *lol*, *atpr*, etc.) seront maintenus, de même que les abréviations familières *stp* et *svp*. On conservera la casse de ces mots. On mettra par contre en majuscules les acronymes et sigles usuels (*SMS*, *GSM*, *ONU*, etc.) ;
3. Répétitions de lettres : on simplifiera les répétitions, sauf dans le cas où elles sont essentielles à une onomatopée. Dans ce cas, on limitera la répétition à 3 lettres. Ainsi *aaaaaammouuurr* deviendra *amour* mais *pfffffffff* sera transformé en *pfff* ;
4. Transformations phonétiques (de type stylistique) : s'il n'y a pas d'ambigüité, on rétablira le mot d'origine ; sinon, on maintiendra la forme brute. Ainsi, *mon ti chéri* devient *mon petit chéri* ;
5. Interjections et onomatopées : on conservera l'orthographe des interjections et onomatopées, à l'exception des répétitions de lettres qui seront supprimées (cf. point 3) ;
6. Cédilles et accents : on rétablira la cédille et les accents, en veillant à systématiquement accentuer les majuscules (et mettre un *Ç* si nécessaire) ;
7. Accords : on accordera, quand c'est évident, les éléments du discours avec le sexe de l'expéditeur du message (exemple : *jsui payé* devient *je suis payé*, si c'est un homme, et *je suis payée*, si c'est une femme). En cas de doute, on laissera tel quel.

Ce protocole de transcription permet de tendre vers l'harmonisation des corpus en formant un ensemble cohérent de messages. C'est avec celui-ci que se clôt la partie

66. Le dictionnaire de référence que nous emploierons est le Rey *et al.* (2011).

consacrée à la présentation des données textuelles. Cette partie dédiée à la présentation du projet général, des chiffres relatifs au corpus de sms, des techniques de collecte et des méthodes de transcription et d'annotation permet de dresser plusieurs conclusions :

1. L'obtention d'un corpus de données privées relève d'enjeux distincts de ceux que l'on peut observer pour l'obtention d'autres types de corpus : l'accès aux données et l'anonymisation de ces données constituent les points centraux de cette spécificité ;
2. Les corpus de sms, comme de la Cémo en général, se distinguent d'autres types de corpus parce que les unités textuelles sont non standard, une phase de transcription, couteuse en temps et en technique, et génératrice de nouvelles erreurs, s'avère nécessaire si l'on veut étudier leur lexique et syntaxe ;
3. Le corpus francophone collecté dans le cadre du projet sms4science comporte **51.280 messages et 2.563 profils d'auteurs** de sms ;
4. Il s'agit d'un corpus **ouvert**, **monolingue** et **diachronique** ;
5. Le corpus de sms bruts présente 3.862.783 caractères, celui de sms transcrits **1.077.042 mots** (tokens) ;
6. Puisqu'il concerne un genre dans différentes variétés, notre corpus est **spécialisé** et donc, par définition, non représentatif.
7. Les sous-corpus régionaux, parce qu'ils ont été collectés selon la même méthodologie sont dits « **comparables** » ;
8. Chaque sous-corpus brut et son équivalent transcrit forment des corpus **alignés au message près**. Le corpus belge est également aligné au caractère près.

6.2 Recueil de données sociodémographiques

[...] les puristes vont contre le français.
Les langues ont toujours changé,
elles appartiennent à ceux qui les parlent.
Calvet (2004)

Sommaire

6.1 Constitution de corpus de sms 64
6.1.1 Linguistique de corpus et sms 66
6.1.2 Des sms comme corpus de données textuelles 70
6.1.3 Collecte des données 73
6.1.4 Comparabilité des données 83
6.2 Recueil de données sociodémographiques 101
6.2.1 Constitution et traitement de formulaires 102
6.2.2 Distribution des variables 110
6.2.3 L'échantillon de locuteurs 118
6.3 Conclusions 124

Introduction

La collecte de sms dans chaque région s'est déroulée parallèlement à une enquête sociolinguistique qui a permis de rassembler les profils des participants. Cette enquête s'est concrétisée par la constitution d'un questionnaire qui a permis une économie en termes de temps et de ressources matérielles, par rapport à la technique de l'entretien, par exemple. Ce questionnaire a été « distribué » (en version papier et sur Internet) et n'a donc pas nécessité la présence de l'enquêteur ; conséquemment, il a permis d'atteindre un échantillon plus large de la population. En outre, le recours au formulaire a permis une collecte objective de données, puisque l'enquête par formulaire ne dépend pas de l'humeur et de la personnalité de l'enquêteur et que l'enquêté y répond – souvent – dans un environnement qu'il choisit, et à son rythme.

L'enquête sociolinguistique auprès d'une population donnée pose la question du paradoxe de l'observateur[67] : « la relation d'enquête [...] reste une relation sociale qui exerce des effets sur les résultats obtenus » (Bourdieu, 1993), mais dont on peut se rendre compte, tenir compte, voire maitriser. Le questionnaire que nous avons distribué via le réseau Internet permet de diminuer les effets de ce paradoxe en n'imposant pas la présence de l'observateur. Pourtant, le locuteur se sent observé parce que (1) l'enquête est présentée clairement en tant que telle (et non dissimulée sous d'autres prétextes) et que (2) l'envoi des sms, comme nous l'avons vu au

67. « the fact that the presence of the observer may destroy the phenomenon that s/he is observing » (Pooley, 1996, p. 80).

sous-chapitre 6.1.3.1, requiert, à chaque envoi, une action volontaire du participant, ce qui exclut de faire oublier l'existence de l'enquête, voire de l'enquêteur, au participant [68]. Une sélection, par l'auteur, des messages à transmettre aux organisateurs a donc probablement eu lieu et constitue un biais dont nous devons tenir compte.

Nous n'avons élaboré nous-même aucun des questionnaires qui sont présentés dans cette partie. Le questionnaire belge préexistait à notre travail et les équipes locales ont elles-mêmes construit leur propre questionnaire. Il est toutefois important de présenter les questions qui ont mené au recueil de données parce que, d'une part, **un tronc commun** a été soumis aux partenaires et que, d'autre part, la formulation des questions élue par chaque équipe influe sur les données que nous manipulons lors de nos analyses : « toute erreur, omission ou ambigüité dans les questions a nécessairement une incidence sur les résultats de la recherche »(Calvet et Dumont, 1999).

Dans la première partie, nous détaillons la constitution des formulaires et la méthodologie de collecte qui a permis l'obtention des informations portant sur les locuteurs. Selon la région, le remplissage du formulaire sociodémographique était plus ou moins obligatoire, ce qui a notamment altéré la quantité d'informations recueillies. Dans la deuxième partie, nous présentons les données sociodémographiques recueillies par région : nous abordons la question des variables possibles pour notre propre enquête (quelles sont les informations communes aux quatre corpus et harmonisables en vue d'être exploitées lors de nos analyses ?). Enfin, nous traitons de la distribution des variables, notamment par rapport à la population réelle concernée.

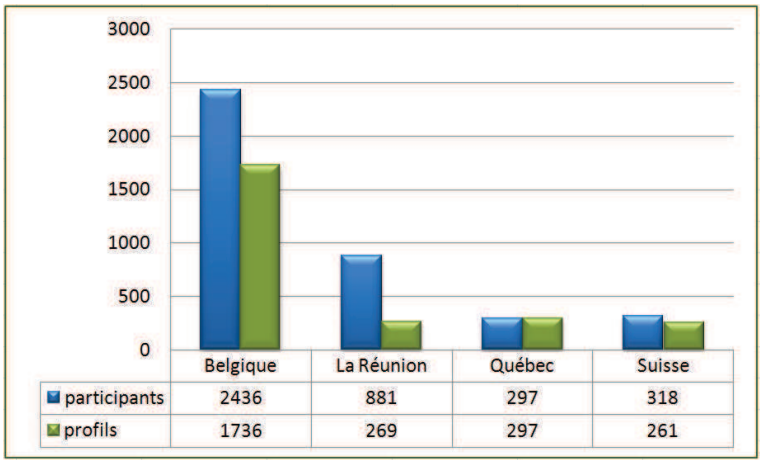
6.2.1 Constitution et traitement de formulaires

Selon la région, la participation des locuteurs à l'enquête sociolinguistique dépendait de facteurs divers tels que les conditions éthiques, qui influaient par exemple sur les conditions de participation au concours pour recevoir des lots, ou encore les préférences et objectifs de chaque équipe locale : recevoir un très grand nombre de sms et rendre le formulaire optionnel ou recevoir un moins grand nombre de sms mais rendre le formulaire obligatoire et donc obtenir plus de profils. En Belgique, en Suisse et à La Réunion, le remplissage du questionnaire n'était pas obligatoire mais nécessaire pour participer au concours. Au Québec, l'envoi des sms nécessitait, au préalable, le remplissage du formulaire. Au total, 82% des participants ont rempli le questionnaire en Suisse francophone, 71% en Belgique (pour les 30.000 sms qui font

68. Nous avons par exemple vu arriver une série de messages adressés aux enquêteurs, notamment en Belgique, par une personne qui explicitait délibérément ses activités quotidiennes, comme elle l'aurait fait dans un journal intime. En Suisse également, l'équipe locale a collecté tous les messages de deux amants envoyés par une des deux parties, une sorte de dialogue amoureux qui ne répond pas aux exigences du projet (puisque nous demandions uniquement des copies de messages envoyés et non reçus) et qui s'inscrit dans un exhibitionnisme conscient. Ces deux exemples illustrent le paradoxe de l'observateur qui a probablement influencé une partie des pratiques de communication/de rédaction lors de la collecte.

l'objet de notre étude) et seulement 30,5% à la Réunion. La Figure 6.8 présente les chiffres concernant les participants dans chaque région.

FIGURE 6.8 – *Répartition des participants et des profils de participants francophones (sms4science)*



Deux remarques importantes émanent de la collection des profils sociodémographiques. La première a trait à l'anonymisation. En effet, même à ce niveau de collecte, une précaution (que nous avons déjà mentionnée) doit être prise pour protéger la vie privée des participants : la suppression des numéros de téléphone qui servent de clés entre les messages envoyés et les profils. Il s'agit donc de remplacer ces numéros par un identifiant neutre dans les deux types de corpus (sms et profils). Lorsque les messages ne sont pas associés à un profil (participant n'ayant pas rempli le questionnaire), un identifiant neutre lui est tout de même attribué afin de faire le lien entre tous les messages d'un même participant. Ceci nous mène à la deuxième remarque : étonnamment, un certain nombre de participants ont rempli un questionnaire sans envoyer de messages par la suite [69]. Ces profils sont enregistrés, car ils sont précieux pour des études portant uniquement sur les profils. Dans le cadre de notre recherche sociolinguistique portant sur l'écrit sms et les profils des auteurs de ses sms, nous les avons bien entendu écartés du corpus d'analyse.

6.2.1.1 Tronc commun

Suite à la sélection des questions réalisée par l'équipe belge en 2004, nous avons proposé aux partenaires suivants un tronc commun de questions en vue de l'exploitation transversale des données [70]. Nous présentons ici le questionnaire commun en le

69. Ce n'est évidemment pas le cas au Québec, où les deux activités étaient liées.

70. Nous n'avons pas obligé les partenaires à utiliser exactement le même questionnaire car nous avons vite ressenti la nécessité d'une souplesse et d'une liberté de chaque membre dans ce partenariat. L'imposition d'un tronc commun a constitué une solution diplomatiquement plus acceptable.

segmentant en fonction de deux ensembles d'information (les profils et les pratiques) et de deux grands types de questions (les questions de fait et les questions d'opinion).

Les questions relatives au profil sont les suivantes :

1. Questions de fait : âge, sexe, profession, niveau d'étude, langue 1, autres langues pratiquées et code postal du domicile ;
2. Questions d'opinion : aucune.

Les questions relatives aux pratiques sont les suivantes :

1. Questions de fait : utilisation d'un dictionnaire d'aide à la rédaction, nombre d'années de pratique du sms, nombre de sms par semaine, types de destinataires des sms (amis, famille, collègue, jeux/concours, etc.), pratique d'autres types de Cémo et compréhension des sms avec codes/abréviations ;
2. Questions d'opinion : « Avez-vous quelque chose à dire sur la façon dont vous rédigez vos sms ? ».

Au moment où nous avons débuté notre recherche, nous avons pensé qu'un certain nombre de questions non abordées dans le questionnaire belge pourraient être pertinentes pour la recherche scientifique en général (en linguistique, en didactique ou en sociologie). Nous les avons également proposées aux nouveaux partenaires, bien que sachant qu'elles ne feraient pas partie de notre recherche personnelle (étant donné que ces questions n'étaient pas posées en Belgique, nous ne pouvions pas les exploiter dans le corpus harmonisé pour toutes les régions).

1. Nombre d'heures de lecture par jour ;
2. Type de lecture ;
3. Nombre d'heures d'écriture sur papier par jour ;
4. Type d'écriture sur papier.

Rappelons qu'il existe également deux grands types de réponse : les réponses fermées (liste finie de réponses possibles) et les réponses ouvertes (texte libre avec un nombre de caractères maximal – variable –). L'erreur que nous avons commise à ce niveau de constitution de corpus, et que nous approfondissons plus tard dans cette partie, concerne l'absence de tronc commun de réponses à ces questions. Ainsi, la disparité de formulation et le choix des réponses dans chaque région a drastiquement réduit la comparabilité des données concernant les profils.

6.2.1.2 Limites des questionnaires élaborés

En annexe A.2 se trouve le détail des questions et la structure des réponses pour chaque région de collecte. Si l'on suit strictement les propriétés d'un « bon questionnaire », telles qu'elles ont été énumérées par exemple par Calvet et Dumont (1999, pages 18-22), les questionnaires construits au sein du projet sms4science montrent un certain nombre de limites. Nous reprenons donc ces propriétés et les confrontons aux questionnaires du projet.

1. Systématicité du questionnaire : « le même questionnaire doit être administré à tous les groupes et sous-groupes de l'échantillon, c'est la condition même de la mesurabilité de l'objet de recherche ». Comme nous l'avons vu, ce n'est pas le cas au sein du projet sms4science. Si nous avons pris garde de respecter un « tronc commun » au niveau du type de question, une certaine liberté a été prise par rapport à leur formulation et aux types de réponses possibles. Certaines de ces variantes proviennent de la volonté personnelle des chercheurs, mais la plupart sont le résultat de demandes spécifiques d'instances juridiques, en vue de respecter des règles d'éthique, qui diffèrent d'une région à l'autre. Conséquemment, le nombre de données croisées utilisables dans le cadre de cette recherche est relativement restreint (cf. le point 6.2.2).
2. Durée du questionnaire : « élaborer un questionnaire aussi bref que possible permet d'économiser la durée de la passation et celle du traitement [...] Les sujets sont plus enclins à répondre quand le questionnaire est court et précis ». Le nombre de questions diffère d'une région à l'autre : 18 questions en Belgique, 30 à la Réunion, 33 au Québec et 25 en Suisse. Le taux plus élevé de questions à la Réunion, au Québec et en Suisse est notamment dû à l'ajout des questions relatives à l'écriture et à la lecture. Pourtant nous estimons qu'elles font sens et ne sont pas redondantes. On se posera par contre la question de l'intérêt de certaines autres questions tel que l'envoi de sms via un service web : qu'apporte cette question comme information ? Comment les réponses à cette question pourront-elles être exploitées ? En vue de quels genres d'étude ?
3. Maitrise de la langue du questionnaire : « le questionnaire doit être rédigé dans une langue parfaitement maitrisée par les sujets ». Il existe dans les questionnaires des mots et expressions dont le référent n'est pas clair, y compris pour nous même. Par exemple, au Québec, la distinction entre la ou les langue(s) parlée(s) « couramment » et la ou les langue(s) parlée(s) « fréquemment ». Les participants peuvent-ils faire la différence entre ces deux questions ? Ne risque-t-on pas d'obtenir les mêmes réponses ? Nous pensons que cette ambigüité sémantique provient notamment de la polysémie relative à l'adverbe « couramment » : il peut signifier autant « qui est fréquent » que « avec aisance ».
4. Exhaustivité des réponses suggérées : toutes les langues sont-elles incluses ? Tous les avis possibles sont-ils représentés ? Dans le cas de La Réunion, la question de la profession est traitée en proposant une réponse fermée comprenant *cadre*, *profession libérale*, *enseignant en secondaire*, *enseignant en primaire ou maternel*, *infirmier*, *commerçant/artisan*, *employé administratif*, *employé de commerce*, *ouvrier*, *sans activité professionnelle* et *retraité*. On imagine très bien des participants ne sachant pas s'identifier à une proposition de la liste (par exemple, professeur des universités). Au Québec, on trouve une erreur de rédaction qui peut être déconcertante pour le participant : à la question du nombre de sms envoyés par semaine, les réponses sont fermées et incluent *moins de 5*, *6 à 10*, etc. Le chiffre 5 n'est donc pas proposé. Une réponse peut

encore être déconcertante à La Réunion : lors des questions sur le mélange de langues dans les sms, il est demandé de préciser les langues en question et les réponses sont fermées : *langue des signes*, *créole*, *français* et *autre* (réponse ouverte avec zone de commentaire limitée en nombre de caractères). Or, l'option *langue des signes* doit faire douter le participant quant à sa compréhension de la question, qui ne devrait évidemment pas proposer une réponse *langue des signes* mais plutôt *SignEcriture* (SignWriting [71]).

5. Une idée par question : « Pour éviter d'obtenir des réponses ambigües, il convient d'exprimer une idée et une seule par question ». Cette propriété d'un bon questionnaire semble en grande partie respectée au travers de nos questionnaires. Nous n'avons trouvé qu'un seul cas de pluralité d'idées dans une question, à la Réunion, où, pour la pratique de l'écriture sur papier, il s'agit de *préciser le but et la fréquence* (réponse ouverte avec zone de commentaire limitée en nombre de caractères). Le participant risque de ne répondre qu'à une seule des deux questions ou bien encore de ne pas savoir comment répondre aux deux questions à la fois. Au Québec, par exemple, cette question a été segmentée en deux questions distinctes.

6. Des questions qui font sens : « La formulation des questions ne doit pas dérouter le sujet, le cadre de référence des questions doit être clair et cohérent pour tous les sujets ». Nous pensons qu'une question commune à tous les questionnaires n'a pas un référent clair. Il s'agit du nombre de sms envoyés par semaine. Nous pensons qu'il aurait été beaucoup plus simple pour le participant d'estimer son nombre de messages envoyés par jour. Conséquemment, nous pensons que les réponses à cette question sont probablement teintées de subjectivité et nous doutons de leur pertinence. Une autre question pose véritablement problème en Suisse : le niveau d'étude. Aucune question n'est véritablement posée dans le questionnaire, on y lit uniquement l'indication « Niveau d'étude ». Or, est-ce le dernier diplôme obtenu ou le diplôme en cours ? Ceci a dû créer de l'ambigüité pour les participants et nous permet donc difficilement d'harmoniser ces réponses avec les autres régions.

7. Simplicité et clarté de l'énoncé des questions : « le lexique employé doit [...] éviter le jargon en étant à la portée des sujets ». Les questions relatives, par exemple, aux types de destinataire des messages, proposent un ordonnancement « en utilisant (1) pour le plus fréquent (allant jusque maximum (5) pour le moins fréquent). Notez (0) pour les éventuelles catégories qui ne vous concernent pas ». Nous pensons que cette consigne n'est pas des plus claires et complexifie en tout cas la participation, notamment par rapport à des réponses par le biais de listes déroulantes ou cases à cocher. Enfin, à La Réunion, à la question « Écrivez-vous en créole », les réponses possibles sont « oui » ou

71. Le SignEcriture est un système d'écriture visuel composé à partir de symboles simples et iconiques né d'un système de notation initié par la chorégraphe Valerie Sutton. Le but de ce système est de transcrire les signes à l'écrit afin de mieux les enseigner, d'en faire des dictionnaires, etc. (Sutton, 2009)

« non » ; si oui, il est demandé de « préciser » et la réponse sera rédigée sous forme de zone de commentaire libre. La question manque de clarté : doit-on répondre en termes de fréquence, de but ou de caractéristique de l'écrit ?

8. Neutralité des questions : « Les termes et les expressions chargés de subjectivité feront [...] place à une langue neutre ». Nous pensons que ce critère est particulièrement pris en considération dans l'élaboration des divers questionnaires sms4science. Nous n'avons pas trouvé de trace de subjectivité ou d'engagement personnel dans la rédaction des questions et des réponses proposées. Le seul point de formulation que nous regrettons se trouve dans le formulaire québécois : lorsque la question de la rédaction des sms est abordée, l'expression « français normal » est utilisée (« En général, quand vous rédigez vos textos, utilisez-vous le français "normal" ou le "langage texto" ? »). Comme nous en avons déjà discuté dans la première partie du chapitre 1, nous ne pensons pas qu'il y ait de « français normal » ni de « langage sms », mais plutôt des pratiques orthographiques aux fréquences variées. Nous pensons donc qu'il aurait été plus neutre de parler d'« orthographe standard » : « En général, quand vous rédigez vos textos, utilisez-vous une orthographe standard ou une orthographe plus abrégée ? » (par exemple).
9. Ordonnancement des questions : « le rangement des questions selon un ordre logique et psychologique approprié permet de mettre à l'aise le sujet, par exemple en faisant en sorte que les questions générales précèdent les questions spécifiques et que les questions d'opinion suivent les questions de fait ». Dans notre cas, l'unique question d'opinion suit bien les questions de fait. On regrettera toutefois l'ordonnancement dans certains questionnaires (par exemple au Québec) où les questions relatives à l'usage des langues s'entremêlent aux questions relatives aux pratiques sms et Cémo. Il aurait été plus clair de regrouper les questions par thématique.

Nous avons volontairement laissé une marge de liberté aux différents partenaires concernant la formulation des questions et les propositions de réponse, notamment parce qu'il n'était pas possible de mener une discussion préalable au lancement général du projet, avec l'ensemble des partenaires, puisque les partenaires ont rejoint progressivement le projet. Les principales limites que nous avons rencontrées concernent donc l'harmonisation des données rassemblées. Les types de réponse proposés, comme nous l'avons dit précédemment, ont significativement posé problème puisqu'ils ont largement varié en fonction des régions. Par exemple, proposer une réponse ouverte (Belgique, La Réunion, Suisse et Québec : autre) aux questions sur les langues a posé un problème d'harmonisation des données que nous avons dû résoudre en les rectifiant manuellement. En effet, l'orthographe variait en termes de casse, de caractères spéciaux (cédille, accents), etc. Ce problème ne fut toutefois pas le plus important, étant donné que l'harmonisation était rendue possible après traitement. D'autres différences ont, elles, engendré une réduction des données significative. Nous choisissons donc de reprendre dans les figures 6.9 et 6.10 les questions du tronc commun et de montrer comment elles ont varié, soit dans leur formula-

tion soit dans les réponses proposées, réduisant ainsi les possibilités d'harmonisation.

FIGURE 6.9 – *Harmonisation des questions de profil du tronc commun (sms4science)*

Variable	Belgique	La Réunion	Québec	Suisse	Harmonisation
âge	numérique	numérique	numérique	numérique	possible
sexe	M / F	M / F	M / F	M / F	possible
profession	Réponse ouverte	11 propositions	19 propositions	4 propositions	réduction aux catégories suisses
niveau d'étude	9 propositions	10 propositions	6 propositions	5 propositions	réduction à 4 catégories. Problème en Suisse et au Québec
langue 1	réponse ouverte. Un seul choix	réponse semi-ouverte. Plusieurs choix	réponse semi-ouverte. Plusieurs choix	réponse semi-ouverte. Plusieurs choix	possible après traitement. Problème en Belgique
autres langues	2e, 3e et 4e langues classées	autres langues + contexte : *couramment*. Non classées	autres langues : *couramment* et *fréquemment*. Non classées	autres langues *parlées* + contexte. Non classées	possible après création d'une catégorie globale « autres langues ».
code postal domicile	numérique	numérique	numérique	numérique	possible
dico aide à la rédaction	possession + utilisation	possession	possession + utilisation	utilisation	non

On remarque par exemple que les niveaux d'étude sont tous différents. Des tables d'équivalence sont donc nécessaires, permettant une harmonisation *a posteriori*. Toutefois, deux problèmes majeurs de subjectivité se posent pour les niveaux d'étude : au Québec, la formulation de la question (« Niveau d'étude actuel (ou le dernier complété si vous avez complété vos études) ») ne permet pas d'exploiter les données correctement. Comment distingue-t-on alors une personne âgée de 20 ans ayant obtenu un dernier diplôme (secondaire) ou une personne âgée de 20 ans ayant doublé et se trouvant toujours en secondaire ? En Suisse, la situation est plus difficile encore en raison du référent vague induit par le label globalisant « Niveau d'étude ». Comment harmoniser sans trancher quant à l'interprétation de la question par les participants ? Nous avons donc décidé de rester prudente quant aux résultats relatifs au niveau d'étude. Pour ce qui est de la profession, nous avons rencontré un problème différent : l'équipe suisse a restreint les catégories à quatre choix. Nous pouvons harmoniser via des tables de correspondance, mais la perte d'information est très importante dans les trois autres régions. Il nous est toutefois possible de maintenir ces informations pour l'exploitation des données. Concernant la langue première, l'équipe belge est la seule à avoir proposé une réponse unique

FIGURE 6.10 – *Harmonisation des questions de pratiques du tronc commun (sms4science)*

Variable	Belgique	La Réunion	Québec	Suisse	Harmonisation
expérience rédaction sms	4 propositions	4 propositions	7 propositions	7 propositions	possible avec tables d'équivalence
nb de sms par semaine	6 propositions	6 propositions	6 propositions	6 propositions	possible
types de destinataires	5 propositions graduées	5 propositions graduées	6 propositions graduées	5 propositions Non graduées	possible avec réduction au non gradué
autres Cémo	3 propositions graduées	3 propositions graduées	4 propositions non graduées	3 propositions graduées	possible avec réduction à 3 non graduées
compréhension des sms abrégés	4 propositions	4 propositions	3 propositions	X	non
commentaire libre sur rédaction des sms ?	champ libre	champ libre	champ libre	champ libre	non, mais exploitation manuelle possible
nb d'heures de lecture	X	numérique	liste déroulante	numérique	non
types de lecture	X	6 propositions	6 propositions	5 propositions	non
nb d'heures d'écriture sur papier	X	champ libre	liste déroulante	liste déroulante	non
types d'écriture sur papier	X	4 propositions	4 propositions	4 propositions	non

(les autres équipes permettant des réponses multiples). Ceci ne pose pas de réel problème pour l'exploitation des données, mais l'on se posera toutefois la question de la pertinence des réponses multiples à la langue première. Le bilinguisme, reconnu par un locuteur, à qui l'on pose la question de la langue première, semble assez subjectif. Gadet (2006) explique : « comment s'y prend le récepteur pour situer ce à quoi il a affaire [...] ? Il faut qu'interviennent des processus de l'ordre de la catégorisation, de l'évaluation [...] tous les faits linguistiques ne sont pas de l'ordre du 1/0 (présence/absence), ni du A/B (concurrence, donc choix) ; mais il existe des faits qui se présentent en continuum ». De plus, on peut penser que la liberté offerte par un deuxième choix de langue dans le questionnaire influe sur l'auto-évaluation du locuteur : là où il n'y aurait que maitrise d'une langue seconde, le locuteur ne serait-il pas enclin, sous l'effet de cette liberté, à y voir du bilinguisme ?

L'harmonisation des langues « autres » (c'est-à-dire différentes de la ou les langue(s) première(s)) a également posé problème puisque la formulation des questions et le format des réponses étaient très variés. Nous avons pris le parti de créer une grande

catégorie « autres langues » que la langue première, sans distinguer le niveau de connaissance ni le contexte d'utilisation de cette ou ces langue(s).

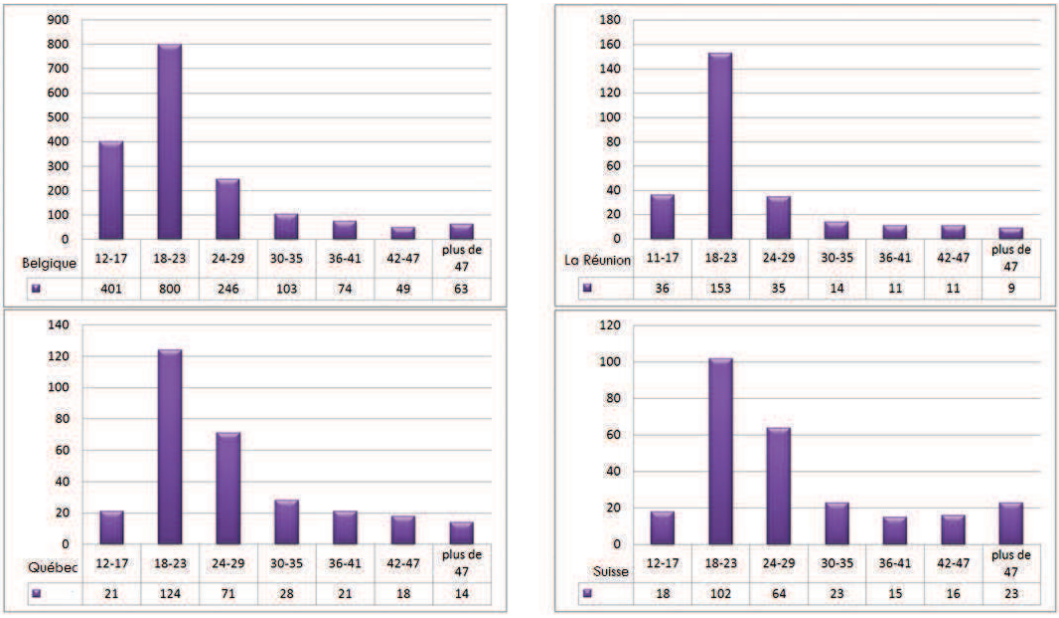
Un autre point concerne le dictionnaire d'aide à la rédaction, qui est un outil influençant de façon importante l'eSMS. Cette variable nous aurait donc été utile. Toutefois, les quatre régions ne se sont pas mises d'accord sur la formulation de cette question et nous perdons donc toute information transversale. La question portant sur le nombre d'années d'expérience dans la pratique du sms est tout à fait harmonisable mais nous pensons qu'à l'heure actuelle elle n'apporte qu'une information minime, la plupart des utilisateurs du téléphone portable envoyant des messages depuis plusieurs années. Les populations les plus jeunes pourraient se démarquer sur ce point, mais dans le cadre de notre recherche, cette information semble triviale. C'est également le cas de l'information sur le code postal qui n'est d'aucun intérêt au sein même de notre recherche puisque l'information régionale (Belgique francophone, Suisse francophone, La Réunion et Canada francophone) suffit. Dans la même lignée de non-adéquation avec notre sujet de recherche, nous écartons le nombre de sms envoyés par semaine (par ailleurs très utile pour des études sur les abréviations, par exemple) et le type de destinataire (par ailleurs très utile pour des études dialogiques). L'information sur la compréhension des sms s'écartant de la norme orthographique aurait été intéressante ; malheureusement, elle manque dans les données suisses et ne permet donc pas l'harmonisation. Nous écartons évidemment les quatre derniers types de questions concernant les pratiques de lecture et d'écriture sur papier, qui sont d'un très grand intérêt sociolinguistique, mais qui n'ont pas été envisagés en Belgique, et ne permettent donc aucune exploitation transversale. Enfin, la question d'opinion sur la pratique du sms (zone de commentaire libre) n'est pas harmonisable, mais est exploitable manuellement pour les études sur les représentations, notamment.

6.2.2 Distribution des variables

L'âge Les distributions de l'âge des auteurs, présentées dans la Figure 6.11, permettent de mettre en avant plusieurs tendances. Tout d'abord, nous voyons qu'il existe partout une nette prédominance des participants de la tranche d'âge 18-23 ans. Ceci est confirmé par la Figure 6.13 qui montre une concentration de la catégorie sans emploi/étudiant/élève. La deuxième classe d'âge est celle des 12-17 ans, en Belgique et à La Réunion. Cette tendance n'est pas suivie par le Québec et la Suisse, où la classe d'âge 24-29 ans se place en deuxièmement position. Les moins de 29 ans sont les locuteurs ayant le plus contribué à l'enquête. On note encore une tendance globale de faible participation des plus de 47 ans, excepté en Suisse où elle excède même la participation des 12-17 ans (une classe d'âge d'ordinaire particulièrement friande de sms). Ces tendances peuvent s'expliquer de diverses manières : elles peuvent refléter la réalité des pratiques ; en effet, il existe peut-être une propension réelle chez les jeunes de moins de 29 ans à envoyer plus de messages que leurs ainés. Si ce n'est pas le cas, cela signifie que c'est notre méthodologie elle-même

(appel à participation avec récompense en lots) ou le tenant même du projet (projet médiatisé, sujet de la recherche, etc.) qui ont engendré une plus grande attraction des plus jeunes.

FIGURE 6.11 – *Répartition de l'âge dans les corpus francophones (sms4science)*

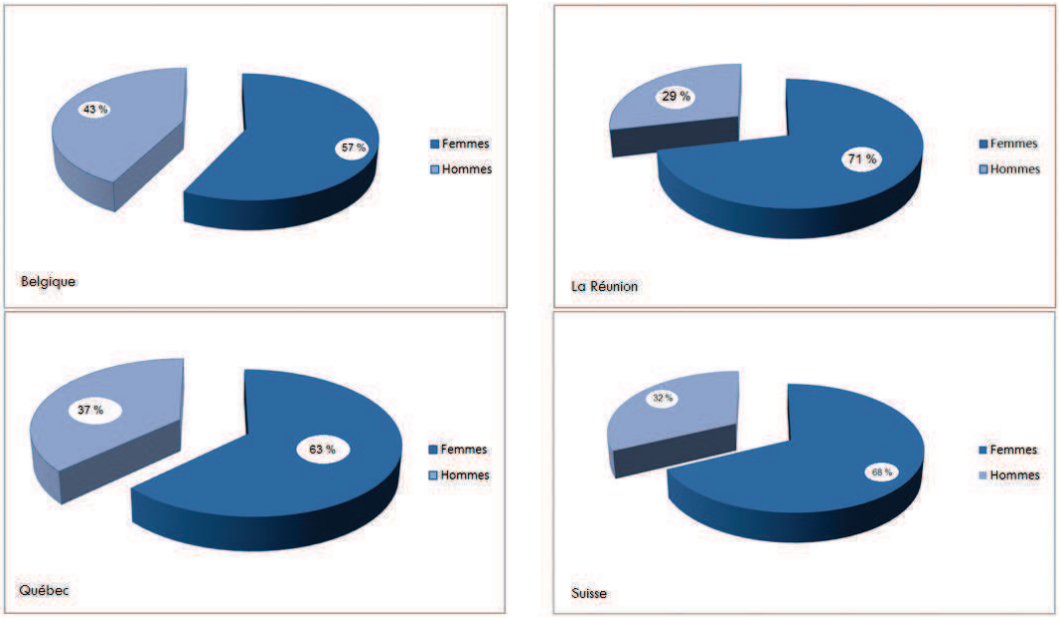


Le sexe La Figure 6.12 illustre la proportion supérieure de femmes ayant participé au projet, une tendance qui se retrouve dans les quatre régions. Là encore, on pourrait penser que la tendance reflète la réalité des faits. Il est pourtant plus difficile de le croire que dans le cas de l'âge. Pourquoi les femmes enverraient-elles à ce point (71% à La Réunion et 66% en Suisse) davantage de sms que les hommes ? Le sujet de recherche lui-même et la méthodologie du projet ne semblent pas pouvoir non plus expliquer ce phénomène. La réponse se situe peut-être dans une tendance générale des femmes à répondre davantage aux enquêtes. Il s'agit d'une tendance mise en avant par différents auteurs tels que Riandey et Widmer (2009). Les auteurs relient ce phénomène au fait que les femmes ont plus de temps libre que les hommes puisqu'elles sont plus nombreuses à rester au foyer [72]. On pourrait également expliquer ce phénomène via le taux plus important de femmes dans les Facultés de Lettres d'où sont généralement parties les enquêtes sms4science. Nous ne prendrons pas position sur ces sujets mais retiendrons les pistes, notamment en ce qu'elles pourraient nous aider dans l'interprétation de la Figure 6.13 (par exemple pour interpréter le haut taux de participation des étudiants et sans-emplois).

La profession Un bref aperçu de la Figure 6.13 pourrait nous indiquer que les tendances sont comparables dans les quatre régions, sauf au Québec où le secteur

72. Une hypothèse qui est, à l'heure actuelle, de moins en moins plausible.

FIGURE 6.12 – *Répartition du sexe dans les corpus francophones (sms4science)*

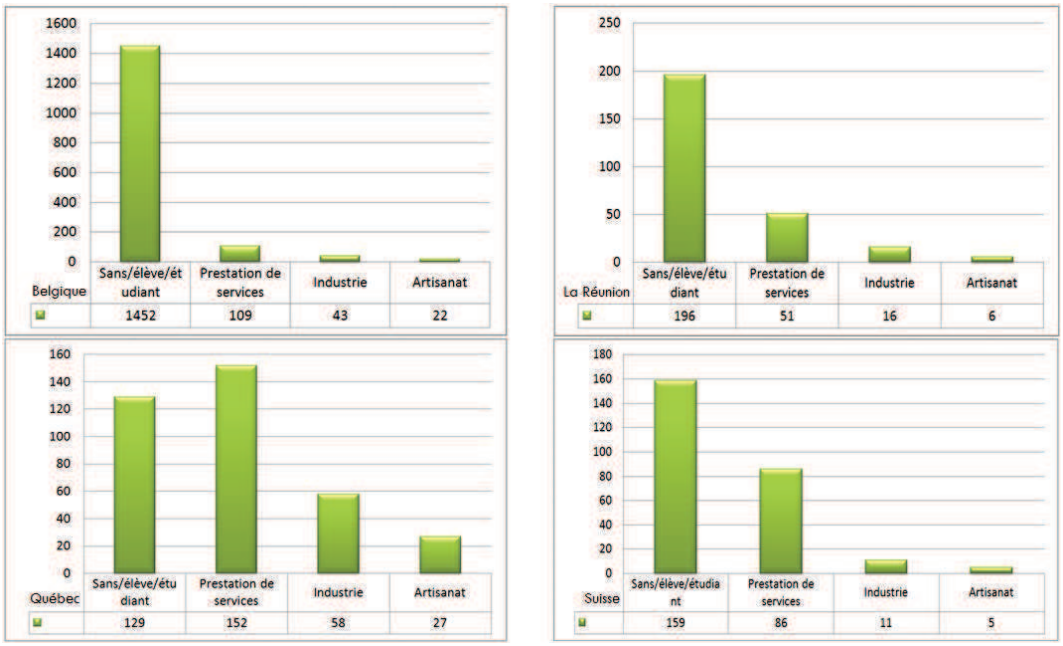


des services, surreprésenté par rapport aux autres régions, offre une courbe distincte. Comme nous l'avons expliqué précédemment, nous pensons que le rassemblement des professions en quatre catégories (via le plus petit dénominateur : le classement suisse) engendre une perte drastique d'information. On pourrait confondre la catégorie « sans emploi/élève/étudiant » en Belgique et en Suisse (deux régions qui offrent les mêmes proportions pour cette catégorie). Or, lorsqu'on réalise une analyse plus fine de données, on s'aperçoit qu'en Belgique, la catégorie surreprésentée est non, comme ailleurs, la catégorie des « élèves/étudiants », mais bien celles des « sans-emploi » (1295 sur 1452 items de la catégorie). Ce chiffre anormal peut s'expliquer, selon nous, de diverses manières. Premièrement, on peut prendre en compte la théorie selon laquelle les personnes n'exerçant pas de profession comme les mères au foyer seraient plus disposées à répondre à une enquête. Nous pensons cependant que cela n'explique pas un si haut taux de participation. Deuxièmement, le type de réponses proposées pour cette question pourrait expliquer ce chiffre anormal. En effet, alors que dans les autres régions, la question profession est suivie d'une réponse fermée obligatoire avec différentes propositions (menu déroulant ou case à cocher), en Belgique, elle est suivie d'une réponse ouverte non obligatoire. Nous pensons alors à deux explications possibles : un oubli de réponse pour les uns, une volonté d'indiquer qu'ils étaient étudiants/élèves (ce qui n'est pas forcément considéré comme une profession) pour les autres.

Un autre problème lié à la mise en commun des professions concerne la difficulté de dresser des tables d'équivalence entre les professions et les catégories socio-professionnelles proposées en Suisse : le choix de mettre « acteur » dans « Arti-

sanat », « Prestation de services » ou « Industrie » est par exemple partiellement arbitraire. Cela est dû à la différence fondamentale entre profession et secteur d'activité [73] datant d'avant la Seconde Guerre mondiale. En Belgique, à La Réunion et au Québec, il s'agit bien de professions, alors qu'en Suisse, les catégories concernent des secteurs d'activités (une profession pouvant appartenir à plusieurs secteurs). Cela rend peut-être encore plus la réduction en 4 catégories non pertinente. Cette variable sera donc peu exploitable, ses conclusions étant facilement discutables.

FIGURE 6.13 – *Répartition de la profession dans les corpus francophones (sms4science)*



Les études Confectionner des tables de correspondance pour les niveaux d'études des différents projets sms4science a été une tâche particulièrement périlleuse en raison de la très grande disparité des catégories proposées par le partenaire. Chaque zone a donc adopté les classifications propres à sa région ou qui leur paraissaient pertinentes dans le cadre de leur objectif personnel de recherche. Nous sommes parvenue à établir 4 catégories principales (primaire, secondaire inférieur, secondaire supérieur et post-secondaire) en réalisant des choix que certains pourraient trouver contestables. En Suisse, par exemple, les formations par canton sont particulièrement différentes, ce qui pose problème pour le projet local également. L'équipe suisse a harmonisé ses données en proposant les catégories suivantes :

1. Éducation primaire ;

73. La classification en trois secteurs d'activité telle qu'elle a été proposée en Suisse (jouxtant une question portant l'intitulé « profession ») s'inspire de la théorie de Clark (1940). Depuis, la distinction profession-secteur d'activité est tout à fait de rigueur.

2. L'International Standard Classification of Education (ISCED) : Niveau 2, pour les secondaires niveau 1 ;
3. L'International Standard Classification of Education (ISCED) : Niveau 3, pour les secondaires niveau 2 ;
4. Le General Certificate of Education (GCE) : Niveau A, pour les personnes ayant répondu Gymnase/Lycée ;
5. Bachelier et diplôme universitaire supérieur.

On constate que les items 2, 3 et 4 renvoient à des niveaux qui se chevauchent partiellement. Pour harmoniser dans le corpus global, nous avons donc pris le parti d'associer l'item 2 au niveau secondaire inférieur et les items 3 et 4 au niveau secondaire supérieur, même s'il s'agit d'une simplification de la réalité. Au Québec, la situation est semblable :

1. 6 années de primaire (6 à 12 ans environ) ;
2. 5 années de secondaire (13 à 17 ans environ) ;
3. 2-3 années de CÉGEP (18 à 20/21 ans environ) ;
4. 3 années de baccalauréat ;
5. 2 années de maitrise ;
6. 3 années de doctorat.

Les items 4, 5 et 6 peuvent être rassemblés dans la catégorie études post-secondaires. Le problème se pose au niveau de l'item 3 : il ne s'agit pas, à proprement parler, de l'école secondaire, mais de 2-3 années prolongeant le secondaire avant d'entamer des études supérieures. Il ne s'agit donc pas non plus d'études post-secondaires. Toutefois, nous avons entrepris, vu l'âge des locuteurs concernés, de les associer à la catégorie post-secondaire.
La Figure 6.14 résume la distribution de ces catégories dans les 4 zones. On constate que la participation des élèves de primaires est quasi inexistante. La progression est constante et semblable dans toutes les zones, avec une participation majeure des étudiants post-secondaires.

La fréquence d'envoi de messages La Figure 6.15 présente une nette différence dans la distribution du Québec. En effet, les trois autres régions montrent une fréquence d'utilisation qui se situe majoritairement entre 10 et 20 sms par semaine. Au Québec, nous avions précédemment conclu que les messages envoyés étaient d'ordinaire plus courts (Cougnon et François, 2011), ce qui nous avait amenée à penser que les Québécois étaient peut-être moins loquaces dans leurs sms, préférant peut-être communiquer des informations plus longues (en nombre de caractères) via des conversations téléphoniques. Ici, nous observons également que la fréquence d'envoi par semaine est largement inférieure à celle des autres régions (taux le plus important à moins de 5 sms par semaine). Notons encore que les données suisses comprenaient des valeurs absentes, raison pour laquelle son graphique inclut une catégorie supplémentaire (NC).

FIGURE 6.14 – *Répartition du niveau d'étude dans les corpus francophones (sms4science)*

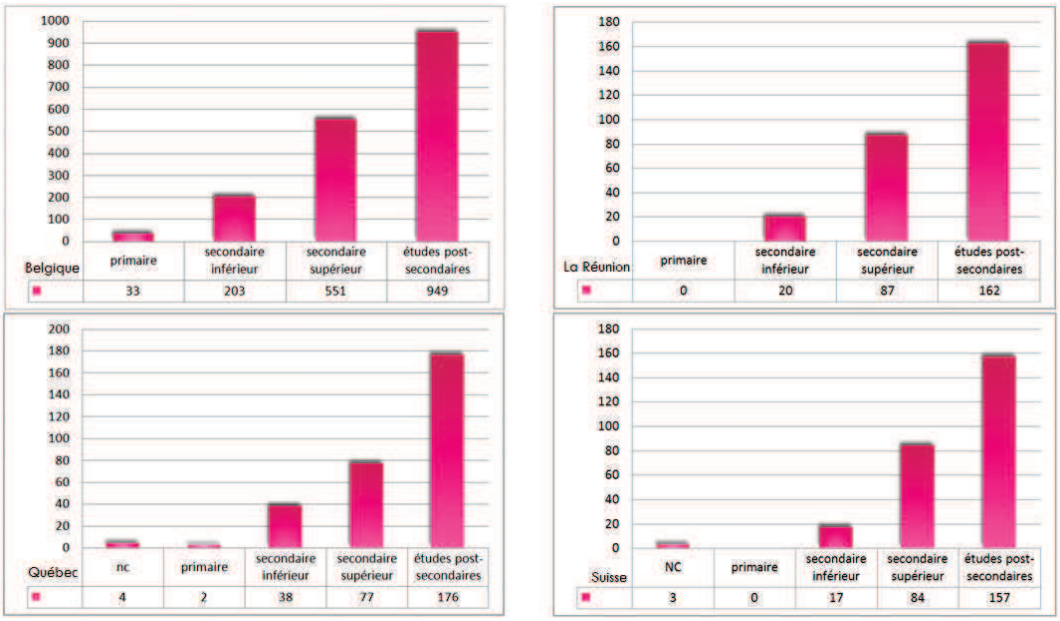
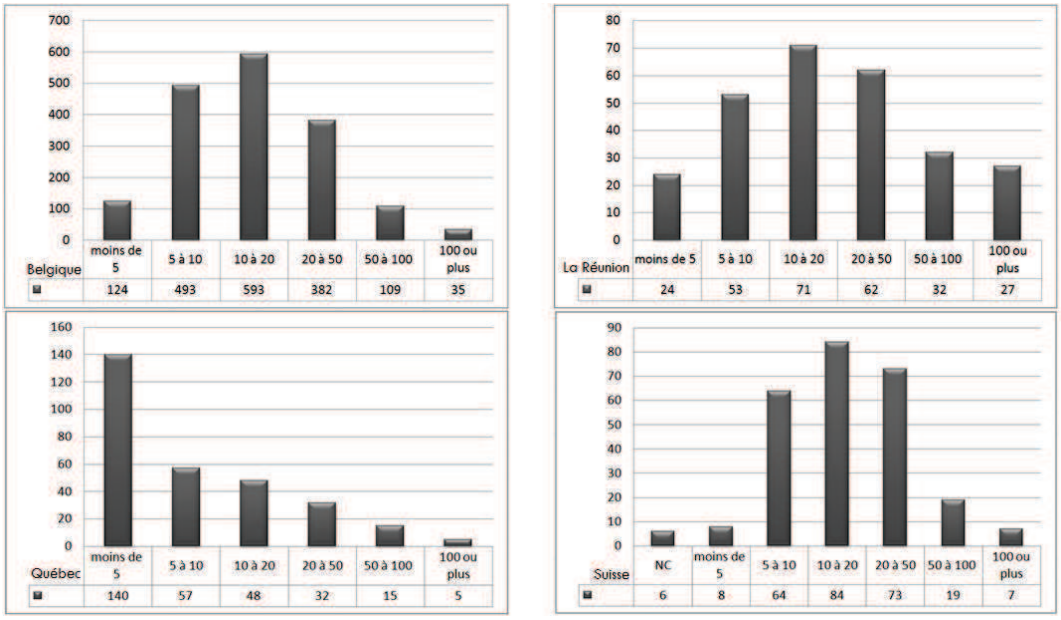


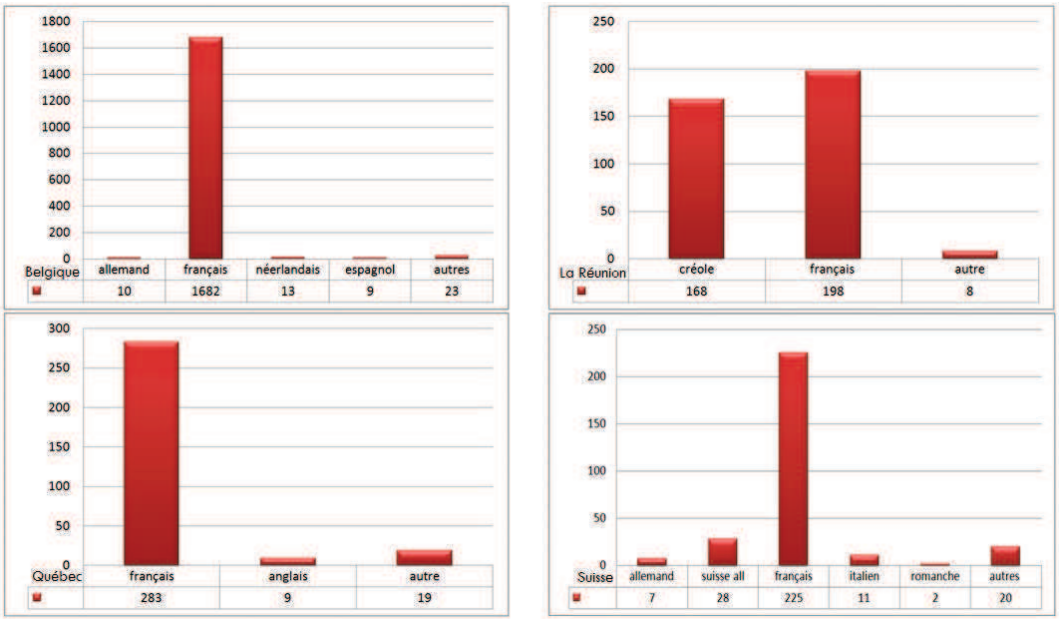
FIGURE 6.15 – *Répartition de la fréquence d'envoi des sms dans les corpus francophones (sms4science)*



La langue première La gestion de la question relative à la langue première dans les différents corpus nous amène à quelques précisions préalables. En effet, là où la Belgique a rendu une réponse unique obligatoire, les autres régions ont rendu

possibles les réponses multiples [74].
Les différents graphiques de la Figure 6.16 représentant les langues premières mettent en valeur la prédominance - logique - de locuteurs francophones dans les quatre corpus francophones. Dans le cas de La Réunion toutefois, le français est nettement concurrencé par le créole, en tant que deuxième langue maternelle. En Suisse par contre, le français arrive en première position, bien devant le suisse alémanique, l'italien ou le romanche. On observe une distribution identique dans les corpus québécois et belge, où l'on aurait pu s'attendre à une concurrence avec, respectivement, l'anglais et le néerlandais ou le wallon.

FIGURE 6.16 – *Répartition de la langue première dans les corpus francophones (sms4science)*

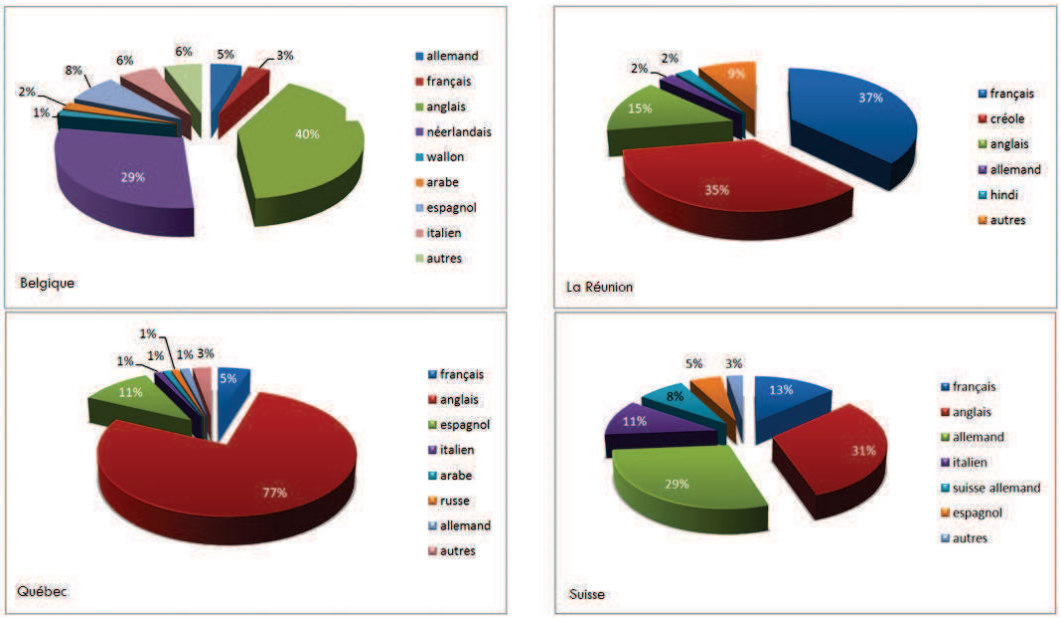


Les autres langues pratiquées Le traitement des données relatives aux langues pratiquées (autres que langues premières) s'est révélé également délicat étant donné la nature ou l'ambigüité des questions posées. En effet, la Suisse et La Réunion ont clairement interrogé les locuteurs sur les « autres » langues pratiquées à la maison ou en dehors de la maison. Un bref aperçu des données montre que les locuteurs ont tendance à associer ces questions à leur langue première (la réponse aux deux questions est donc la même). En outre, les locuteurs pouvant pratiquer la même « autre » langue à la maison et en dehors de celle-ci, des répétitions apparaissaient à ce niveau également. Au Québec, un problème identique se pose avec les questions « Quelle(s) autre(s) langue(s) parlez-vous couramment ? » et

74. Plusieurs langues peuvent être considérées comme des langues premières (voir par exemple les cas de bilinguisme précoce, comme dans Titone 1972), bien que l'auto-évaluation de cette qualité par le locuteur se révèle difficile.

« Quelle(s) langue(s) parlez-vous le plus fréquemment ? » [75]. Afin de neutraliser ces différences régionales, nous avons entrepris de supprimer les doublons apparaissant dans les questionnaires remplis par les locuteurs afin de regrouper des réponses générales concernant les « autres langues pratiquées » (que nous appellerons L2 même s'il ne nous est plus permis de distinguer, dans l'ordre, les langues 2, 3, etc.), quelque situation ou maitrise de la langue que ce soit.
La Figure 6.17 met en avant plusieurs phénomènes. Tout d'abord, l'importance accordée à l'anglais, langue « vernaculaire », apparait dans toutes les régions. Comme on pouvait s'y attendre, d'autres L2 sont également pratiquées, en nombre important de locuteurs, en Belgique, à La Réunion et en Suisse. C'est le cas du néerlandais (29%) en Belgique, du français (37%) et du créole (35%) à La Réunion ainsi que de l'allemand (31%) et du français (18%) en Suisse. Notons que la présence du français dans la catégorie L2 révèle l'apport aux corpus francophones de sms de locuteurs non francophones.

FIGURE 6.17 – *Répartition des autres langues pratiquées par les usagers dans les corpus francophones (sms4science)*



Cémo La dernière variable commune concerne les pratiques écrites dans d'autres média de communication : le chat, les forums et le courrier électronique. Quelques réajustements ont eu lieu sur les corpus afin qu'ils soient comparables. C'est le cas, notamment, des degrés de pratique, qui ont dû être neutralisés pour s'aligner sur le Québec. Nous avons entrepris d'assembler les réponses à la question « Utilisez-vous d'autres systèmes de communication sur Internet ? » comme suit : « Tous les

75. On notera dans cette dernière question que l'adjectif « autre » a disparu, ce qui rend la question d'autant plus ambigüe.

jours » et « Régulièrement » sont devenues des réponses affirmatives, tandis que « Rarement » et « Jamais » ont été transformées en réponses négatives. Cela peut entrainer un biais que nous avons pensé négligeable dans l'objectif d'une enquête sur les pratiques générales.
Nous avons également entrepris d'interpréter les réponses absentes (par exemple en Belgique) comme des réponses négatives : étant donné l'intitulé de la question, il nous semblait normal que les participants ne s'intéressent pas aux réponses proposées pour une question qui ne les concerne pas. En effet, en raison des réponses fournies aux questions « temps sms » et « sms » en moyenne (des questions qui permettaient de ne pas répondre et auxquelles tout le monde a répondu) et au vu de la formulation de la question (« Utilisez-vous d'autres systèmes de communication sur Internet ? » et non pas « Comment ou à quelle fréquence utilisez-vous d'autres systèmes de communication sur Internet ? »), nous pouvons raisonnablement émettre l'hypothèse que les participants qui n'ont pas répondu voulaient signifier qu'ils n'utilisaient pas ce type de Cémo. Il pourrait également s'agir d'oublis, mais nous pensons que ces cas sont minoritaires. Nous fusionnons donc la catégorie « non communiqué » avec la catégorie « non ».
La Figure 6.18 laisse entrevoir une évolution des pratiques dans le temps. En effet, nous constatons que lors du projet en 2004, le courrier électronique était majoritairement non employé, alors qu'il devient au fil du temps la pratique de Cémo la plus courante [76]. Le chat apparait comme le deuxième mode de Cémo à La Réunion, en Suisse et au Québec. Les discussions sur forum sont minoritaires dans toutes les régions interrogées.

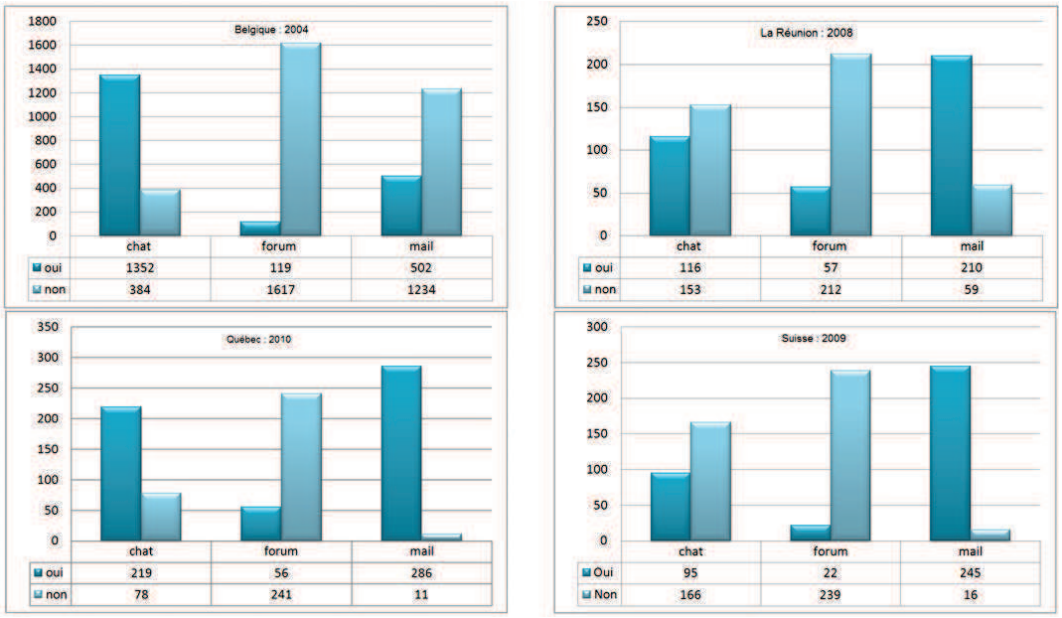
6.2.3 L'échantillon de locuteurs

La question de la représentativité des corpus de profils sms4science est une question que nous avons déjà entrepris d'étudier dans Cougnon et François (2010), notamment, pour le corpus belge. Dans la partie 6.1.3.3, nous avons montré que les données sms ne constituaient pas un corpus représentatif. Mais qu'en est-il de l'échantillon des locuteurs interrogés par rapport à l'ensemble des locuteurs pratiquant la communication par sms ? La distribution de ces locuteurs par région, par âge, etc. de notre corpus est-elle pertinente ?

La méthodologie du projet sms4science consiste, comme nous l'avons vu, en un appel à participation diversifié : distribution de lots, rédaction d'articles dans la presse, affichage ciblé dans des endroits-clés (maisons de repos, centres de jeunes, centres sportifs, etc.), etc. Cet appel vise donc l'ensemble de la population, à travers les âges et les milieux sociaux. Comme nous l'avons précisé précédemment, la participation au projet se veut gratuite [77] ; elle permet d'effacer les barrières économiques qui pourraient dissuader de participer et donc entraver l'équilibre de notre échantillon.

76. Afin de vérifier l'hypothèse selon laquelle il ne s'agit pas d'une différence due à la région mais bien à l'année, on pourrait imaginer l'organisation d'une deuxième collecte en Belgique francophone.
77. Bien que cette gratuité n'ait pas été possible partout.

FIGURE 6.18 – *Répartition des pratiques de Cémo dans les corpus francophones(sms4science)*



Enfin, d'un point de vue linguistique, nous cherchons évidemment à colliger des messages de tous styles et registres, quelle que soit la maitrise orthographique et les thématiques abordées ; nous précisons ceci à chaque intervention dans les médias, en expliquant le pourquoi même de cet intérêt diversifié. Si chacune des conditions est respectée, le corpus se veut donc aléatoire, dans le sens que le projet n'encourage pas plus qu'il ne refrène certaines parties de la population à la participation.

Cependant, même si cette méthode comporte des aspects très positifs et pallie effectivement les problèmes d'authenticité ou de taille du corpus, elle reste insatisfaisante d'un point de vue purement statistique, puisqu'elle opère un échantillonnage par volontaires, contingent [78], lequel ne produit généralement pas un échantillon représentatif de l'ensemble des caractéristiques de la population [79]. En outre, dans certaines régions participant au projet, les conditions énumérées plus haut n'étaient pas rencontrées (non-gratuité des envois, lots moins attractifs, etc.),

78. Notons que nous ne pouvions pas non plus user du critère interne de saturation (« le moment lors duquel le chercheur réalise que l'ajout de données nouvelles dans sa recherche n'occasionne pas une meilleure compréhension du phénomène étudié », Mucchielli 1996, p. 204) puisque les phénomènes à étudier étaient encore inconnus (défrichage du domaine).

79. L'échantillonnage par volontaires appartient à l'ensemble des échantillonnages non probabilistes. « Le fait d'échantillonner des participants volontaires plutôt que la population en général peut introduire des biais marqués. Souvent, à l'occasion des sondages d'opinion, seuls les gens qui se soucient assez fortement d'une façon ou d'une autre de la question étudiée ont tendance à y répondre. La majorité silencieuse n'y répond généralement pas, ce qui entraine un important biais sur le plan de la sélection » (Documentation de l'organe statistique du gouvernement canadien : `http://www.statcan.gc.ca/edu/power-pouvoir/ch13/nonprob/5214898-fra.htm`).

ce qui peut entrainer un biais encore plus important dans la répartition des données.

Nous présentons la distribution de l'échantillon de la population par rapport à la population concernée, telle que précédemment détaillée dans Cougnon et François (2010). Afin de vérifier la distribution d'un jeu de données, il est possible, en statistiques, d'effectuer une série de tests d'ajustement (test khi-carré pour une distribution non paramétrique ou test de normalité pour des données normales [80]) pour chacune des dimensions prises en compte (région, âge, sexe, etc.). Toutefois, dans le cas d'un corpus portant sur des auteurs 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), en Suisse (francophone), à La Réunion et au Québec – 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 téléphones portables ou qu'il ait un téléphone portable sans utiliser la fonction sms [81]. Confrontée à ces difficultés méthodologiques, nous avons choisi de nous focaliser sur les proportions entre les variables de la population et d'effectuer une série de tests khi-carré d'ajustement (Agresti, 2002, p. 22) sur ces variables modélisées aussi précisément que possible [82].

Nous avons expliqué précédemment (sous-chapitre 6.2.2) pourquoi les variables construites à partir des profils sms4science étaient à utiliser avec prudence. C'est pourquoi, en vue d'évaluer la distribution de notre jeu de données, nous avons décidé d'utiliser des variables claires et stables, y compris après harmonisation : le sexe et l'âge des auteurs des messages [83].

En ce qui concerne le sexe des utilisateurs, on observe une plus grande proportion de femmes ayant participé au projet. Dans le corpus belge, on compte 57% de femmes et 43% d'hommes ; en comparant avec le nombre de résidents en Belgique francophone, où la proportion de femmes est de 51,6% contre 48,4% pour les hommes, il n'est pas difficile de prévoir une non-représentativité de la population par genre. Dans le corpus réunionnais, on dénombre une participation de 71,4% de femmes et de 28,6% d'hommes pour une population qui montre seulement 48,4%

80. Pour plus de précisions à propos des tests statistiques, voir le sous-chapitre 7.3.

81. C'est également en raison de cette impossibilité de définir la population cible de l'échantillon que nous n'avons pas eu recours à un rééchantillonnage du corpus *a posteriori*.

82. Pour ce faire, nous nous sommes basée sur une liste limitée de sources statistiques fiables : pour la Belgique, le site de Statbel, le principal acteur de la statistique officielle en Belgique : http://statbel.fgov.be/fr/statistiques/chiffres/population/structure/agesexe/, pour La Réunion, l'Institut national de la statistique et des études économiques : http://www.insee.fr/fr/themes/document.asp?reg_id=24&ref_id=8710, au Québec, l'Institut de la statistique du Québec : http://www.stat.gouv.qc.ca/donstat/societe/demographie/struc_poplt/201_2010.htm et pour la Suisse, la Section statistique de l'Administration fédérale suisse : http://www.bfs.admin.ch/bfs/portal/fr/index/themen/01/02/blank/data/01.html.

83. Notons que la distribution de la variable région a déjà été succinctement abordée dans la *Problématique* de cette dissertation doctorale.

TABLE 6.2 – *Répartition des utilisateurs belges de sms, selon leur classe d'âge, dans le corpus, et dans la population résidante (globale)*

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%

TABLE 6.3 – *Répartition des utilisateurs réunionnais de sms, selon leur classe d'âge, dans le corpus, et dans la population résidante (globale)*

Classes	Proportion : corpus	Proportion : population	Classes	Proportion : corpus	Proportion : population
1 : - de 15 ans	3,35%	26,3%	4 : 25 – 34 ans	13,01%	13,71%
2 : 15 – 19 ans	34,2%	8,95%	5 : 35 – 44 ans	6,69%	16,37%
3 : 20 – 24 ans	36,8%	6,99%	6 : + de 45 ans	5,95%	27,69%

de femmes pour 51,6% d'hommes. Au Québec, le corpus présente une répartition de 63% de femmes et 37% d'hommes pour une population comprenant en réalité 50,4% de femmes et 49,6% d'hommes. Enfin, en Suisse, le corpus de profils est constitué de 66% de femmes et de 34% d'hommes, pour une population Suisse francophone répartie en 51% de femmes et 49% d'hommes. Le résultat du test d'ajustement belge ($\chi^2(1) = 21,2\,;\ p < .0001$) indique que sa population ne peut être représentative de celle d'où provient l'échantillon. On note la même tendance à La Réunion, au Québec et en Suisse[84]. Ces résultats indiquent donc un biais considérable dans l'échantillonnage, à moins que, pour une raison qui reste à éclaircir, les femmes aient réellement davantage tendance à communiquer via sms que les hommes[85].

Au niveau de l'âge, nous avons regroupé les utilisateurs en 6 classes d'âge. Les tableaux 6.2 à 6.5 précisent leurs frontières, leurs proportions dans le corpus et dans la population.

84. Où les résultats sont de $\chi^2(1) = 56,6\,;\ p < .0001$ pour la Réunion, $\chi^2(1) = 18,7\,;\ p < .0001$ pour le Québec et $\chi^2(1) = 25,8\,;\ p < .0001$ pour la Suisse.

85. Ce qui renvoie à ce que nous avons annoncé précédemment : la distribution de la population utilisant la fonction sms d'un téléphone portable peut être très différente de la population totale d'une région donnée.

TABLE 6.4 – *Répartition des utilisateurs québécois de sms, selon leur classe d'âge, dans le corpus, et dans la population résidante (globale)*

Classes	Proportion : corpus	Proportion : population	Classes	Proportion : corpus	Proportion : population
1 : - de 15 ans	1,35%	15,64%	4 : 25 – 34 ans	28,62%	13,69%
2 : 15 – 19 ans	17,85%	6,32%	5 : 35 – 44 ans	11,78%	13,28%
3 : 20 – 24 ans	33,33%	6,28%	6 : + de 45 ans	7,07%	44,79%

TABLE 6.5 – *Répartition des utilisateurs suisses de sms, selon leur classe d'âge, dans le corpus, et dans la population résidante (globale)*

Classes	Proportion : corpus	Proportion : population	Classes	Proportion : corpus	Proportion : population
1 : - de 15 ans	1,08%	15,17%	4 : 25 – 34 ans	25,99%	13,26%
2 : 15 – 19 ans	17,33%	5,84%	5 : 35 – 44 ans	9,75%	15,59%
3 : 20 – 24 ans	34,66%	6,05%	6 : + de 45 ans	11,19%	44,09%

Les tableaux montrent 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 [86], 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 jouent ensemble : le sms constitue certainement une technologie séduisant davantage un public jeune, bien qu'il faille nécessairement admettre que les lots de stimulation utilisés pour la collecte du corpus ainsi que la thématique même du projet de recherche ont très probablement influencé le type de population ayant participé à l'étude.

Nous venons donc de montrer la présence de plusieurs biais dans notre corpus. La présence de ces biais doit-elle conduire à rejeter toute approche statistique ? Nous avons montré que notre corpus se différencie nettement de la population résidant dans chaque région francophone, sans pouvoir déterminer si cet écart s'explique uniquement par des biais liés à la méthode d'échantillonnage par volontaires ou s'il met en avant des caractéristiques réelles de la population d'utilisateurs des sms. Pour trancher cette question, il conviendrait d'effectuer une nouvelle collecte de sms qui respecte les principes d'un échantillonnage aléatoire simple. Or, cette option reste actuellement non envisageable, pour les raisons légales d'impossibilité de colliger tous les messages arrivant chez l'ensemble des opérateurs d'une région. 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 et ne se calcule que par approximation. Par conséquent, ces diverses difficultés méthodologiques limitent fortement toute tentative d'inférence statistique, mais n'empêchent pas d'étudier le corpus en tant que tel, sans étendre automatiquement les conclusions à la population totale. Nous interprétons les résultats de notre recherche dans la lignée de Van Campenhoudt (2002) [87] comme valables pour eux même et non généralisables aux pratiques de l'ensemble de la population étudiée. Ainsi, nous garderons à l'esprit que les distributions non représentatives de la population peuvent expliquer certaines tendances et certains phénomènes que nous observerons. Suivant Dister *et al.* (2009, p. 121), nous affirmons donc que notre projet « ne prétend pas représenter [...], mais vise plutôt [...] à documenter la diversité des pratiques linguistiques observables ».

86. $\chi^2(5) = 3941,5$; $p < .0001$ pour la Belgique, $\chi^2(5) = 648,9$; $p < .0001$ pour La Réunion, $\chi^2(5) = 590,5$; $p < .0001$ pour le Québec et $\chi^2(5) = 581,6$; $p < .0001$ pour la Suisse.

87. « Loin de l'empirisme dont font preuve certains, l'observation d'énoncés attestés dans un environnement textuel participe en réalité d'une démarche scientifique qui s'avère rigoureuse dès lors qu'elle prend en considération les paramètres de l'expérimentation et en relativise les résultats » (Van Campenhoudt, 2002, p. 3).

6.3 Conclusions

Selon le parcours méthodologique que nous venons de présenter ainsi que la description des corpus obtenus, nous pouvons affiner la définition de notre double banque de données : d'un côté, nous avons en notre possession une banque de 51.280 sms comprenant plus d'un million de mots, propres à « l'écrit spontané ». Il s'agit d'un corpus *monolingue* (francophone), *ouvert* et *dynamique*.

De l'autre côté, nous nous basons sur une banque de 2.563 profils sociodémographiques, harmonisables, *a posteriori*, pour ce qui est des variables suivantes : âge, sexe, langue 1, langue 2 (et autres), niveau d'étude, profession, nombre de sms envoyés par semaine et pratique d'autres types de Cémo.

Bien que la distribution des données des deux corpus soit inégale en termes de région (la Belgique est surreprésentée, le Québec et la Suisse sont sous-représentés), ou selon les variables étudiées (âge, sexe), en raison de la méthode d'échantillonnage non probabiliste employée, ces corpus sont toutefois précieux et valables particulièrement au regard de la rareté de tels corpus à l'heure actuelle et de la difficile accessibilité des données confidentielles. Nous pouvons encore nous réjouir d'avoir collecté des données *véritables*, dans le sens qu'elles n'ont pas été créées pour l'occasion ou encore qu'elles ne proviennent pas d'un réseau proche de l'enquêteur. Pour ce qui est des profils, même si l'harmonisation des 4 régions a engendré une perte d'information conséquente, il n'en reste pas moins que les variables harmonisées permettent d'approcher les corpus de sms dans une optique sociolinguistique. En outre, nous prendrons garde de ne pas interpréter nos résultats trop hâtivement, en considérant notamment le risque de circularité relatif à la définition des variables : « la tentation de faire des phénomènes linguistiques des indices de localisation spatiale, sociale ou situationnelle. Par exemple, une catégorie "jeunes" existe parce que les jeunes ont un comportement langagier spécifique ; et comme leur comportement langagier est spécifique, il y avait bien lieu de poser une catégorie "jeunes" » (Gadet, 2004, p. 70).

Ainsi, armée de ces deux corpus conséquents, et des précautions théoriques qui leur sont inhérentes, nous sommes prête à entamer l'exploitation sociolinguistique de la variation des données.

Troisième partie

De la variation des données

Chapitre 7

Préliminaires

In conducting a microscopic analysis of data, it is important not to become overwhelmed by the vast amount of statistical information that such a study will be able to generate, but to focus instead on using statistical analysis to confirm or disconfirm the particular hypotheses one has set out to test.
Meyer (2002, p. 125)

Les analyses lexicale et syntaxique des données sms nécessitent quelques mises au point préalables. Il s'agit de répondre à un certain nombre de questions théoriques et méthodologiques :

1. En quoi consiste la notion de variation en sociolinguistique et en quoi est-elle applicable à notre cas ?
2. Pourquoi approchons-nous le corpus à l'aide d'une dichotomie lexique/syntaxe ?
3. Quelle double méthodologie cette dichotomie induit-elle ? En quoi cette dichotomie est-elle également liée à la différence entre analyses fondées sur corpus et analyses guidées par corpus ?
4. À quoi servent les statistiques dans l'analyse de la variation en sociolinguistique ? Quels tests allons-nous utiliser et que faut-il savoir à leur propos ?

7.1 La notion de variation pour un corpus de sms

L'analyse variationnelle utilise les techniques et les théories de la linguistique, de l'anthropologie et des statistiques pour étudier l'usage et la structure du langage (Poplack, 1993, p. 251). Nous définirons la variation sociolinguistique, plus précisément, à la suite de Tagliamonte (2006, p. 5) : il s'agit de « la branche de la linguistique qui étudie les propriétés majeures du langage, en considérant avec autant d'importance – structure linguistique et structure sociale ; sens grammatical et sens social – les propriétés du langage qui nécessitent de prendre en compte, dans les explications, autant les facteurs externes (sociaux) qu'internes (systémiques) ».

La littérature sur la variation montre que cette notion est apparue très tôt dans l'histoire de la linguistique française, notamment avec l'œuvre dialectologique de Jules Gilliéron ainsi qu'avec l'approche de la variation sociale dans une communauté linguistique donnée par Louis Gauchat au début du XX^{e} siècle. La variation géographique a fait l'objet de très nombreux travaux en analyse variationniste, et

la variation sociale (Gadet, 2004), en particulier, n'a pas été mise à l'écart : on recense notamment des travaux sur la variation stylistique (Biber, 1988) et plus généralement sur la variation phonologique (Bauvois, 2002 ; Laks, 1977), en fonction de la classe sociale, la région ou l'âge et l'influence de l'identité sociale (Beeching, 2007) sur la parole. La variation sociolinguistique n'a toutefois pas marqué de la même façon la communauté scientifique européenne et la communauté québécoise, par exemple, ou la recherche anglo-saxonne (Auger, 2006, p. 355).

L'œuvre de William Labov, d'abord qualifiée de dialectologie sociale ou de linguistique corrélationniste, s'inscrit en réalité dans la linguistique variationniste, que l'auteur a profondément marquée (Labov, 1966, 1990). La linguistique de Labov se distingue de la linguistique chomskyenne puisqu'elle envisage la langue comme un système au sein d'une communauté linguistique, contrairement à celle de Chomsky qui renvoie aux structures cognitives du cerveau humain.

> « Je désigne comme "sociaux" les traits de langage qui, dans une communauté hétérogène, caractérisent divers sous-groupes ; et comme "stylistiques", les modifications au moyen desquelles un locuteur adapte son langage au contexte immédiat de son acte de parole. Ces deux éléments appartiennent au comportement "expressif", à la façon dont le locuteur, non content d'informer l'auditeur en lui représentant le monde extérieur, lui révèle en même temps un peu de lui-même et de ses dispositions. La variation sociale et stylistique présuppose que l'on peut choisir de "dire la même chose" de plusieurs façons différentes, soit par des variantes dont la valeur dénotative ou assertorique est constante, mais qui s'opposent par leur significativité sociale et/ou stylistique ». (Labov, 1976, p. 366)

Définir la variation dans l'eSMS s'avère une tâche plus compliquée : nous avons vu (au sous-chapitre 5.2.1) que l'eSMS pouvait être considéré comme un nouveau code, partiellement distinct de l'écrit traditionnel. Lorsque nous énonçons, comme nous l'avons fait dans ce sous-chapitre, des règles, des principes graphiques qui régissent l'eSMS, nous projetons en réalité la norme de l'écrit traditionnel sur l'eSMS, en prenant comme point de départ pour l'analyse les écarts entre eSMS et écrit traditionnel. Mais comment étudier la variation lexicale et syntaxique d'un corpus « transcrit » en graphie standard ? C'est la question que nous allons poser dans cette partie afin d'introduire la notion de variation qui est à la base de l'analyse. La démarche analytique qui suivra portera donc sur quatre types de variation :

1. Diatopique : étudier la variation de phénomènes linguistiques d'une région à l'autre de la francophonie ;
2. Diastratique : étudier la variation d'une sous-catégorie de la population à l'autre, en utilisant les données de profil collectées avec les messages ; le regroupement en classes de personnes se fait selon les catégories suivantes : âge, sexe, niveau d'éducation, nombre de sms envoyés par semaine et pratique ou non d'un autre type de Cémo ;

3. Par rapport à la syntaxe du français dit « standard » : étudier la variation de phénomènes syntaxiques par rapport à la norme définie par les grammaires du français ;
4. Par rapport à une « norme d'usage », au sens de Girard et Lyche (2004), c'est-à-dire par rapport à la norme inhérente au genre d'écrit étudié (une moyenne des pratiques observées pour ce genre) : on pourra, à partir de l'ensemble des corpus de sms, découvrir les unités lexicales n'apparaissant que dans une aire géographique donnée (c'est par exemple le cas pour l'étude des mots-clés).

La variation diatopique du français a fait l'objet de nombreuses recherches et notre but consiste à examiner en quoi cette variation se concrétise dans un écrit ordinaire. La variation diastratique de l'écrit sms a été amorcée dans quelques études (Bieswanger, 2010 ; Goumi *et al.*, 2011 ; Ursini, 2005) qui nous ont offert un petit nombre de conclusions que nous voulons infirmer ou confirmer. La variation de l'eSMS par rapport à une syntaxe de référence en français fait l'objet de peu d'études mais de nombreuses accusations du grand public et des médias, qui voient dans les nouveaux moyens de communication une « déstructuration » du langage. Nous aurons pour but d'analyser une série de phénomènes syntaxiques majeurs afin de vérifier objectivement ces présupposés.

Nous envisagerons enfin une approche plurifonctionnelle de la variation. Il s'agira de voir si les pratiques linguistiques de l'eSMS répondent à un besoin unique de communication économique ou si elles sont le reflet de volontés et de contextes individuels et sociaux plus complexes. En effet, nous pensons, à la suite de Gadet (2004, p. 54) que la variation remplirait « des fonctions de désambigüisation, de résolution de conflits, et d'économie, avec un compromis entre couts et gains de la variabilité, entre prestige de la norme et facilité de production », entre volonté de véhiculer des émotions, marquer une identité, tout en répondant à des tendances et pressions sociales.

7.2 Approches guidées par corpus et fondées sur corpus

L'approche des données de sms que nous envisageons est double : il s'agit d'une approche mêlant des analyses « guidées par corpus » (ou *corpus-driven*) aux analyses « fondées sur corpus » (ou *corpus-based*). En effet, il existe deux grands types distincts d'approches utilisées en linguistique de corpus. McEnery et Hardie (2012) définissent l'approche fondée sur corpus comme étant une méthode qui consiste à confirmer ou infirmer des hypothèses, souvent présentes dans la littérature ou ailleurs, en les confrontant à des données issues des corpus (McEnery et Hardie, 2012, p. 6). On cherchera par exemple à vérifier que le vocabulaire des 12-18 ans dans la communication par sms est assez « pauvre ». Les approches guidées par corpus, quant à elles, ne considèrent pas la linguistique de corpus en tant que méthode pour vérifier des hypothèses émises par ailleurs (Tognini Bonelli, 2001). Elles utilisent le corpus comme un point de départ pour faire émerger les hypothèses

sur la langue[1]. C'est en regardant avec attention les corpus de sms que l'on y trouve des traces de régionalismes lexicaux dont nous n'aurions pu nous douter, *a priori*.

Certains linguistes posent une complémentarité entre les deux approches (Tagg, 2009, p. 47) et incitent les chercheurs à effectuer un va-et-vient constant entre leurs hypothèses et les données afin de ne pas limiter les points de vue. Nous aborderons les deux conceptions de manière distincte (nous avons scindé les chapitres en analyse lexicale et analyse syntaxique), sans jamais les dissocier réellement.

Pour la partie lexicale, nous avons appliqué un dictionnaire au corpus de sms. Le dictionnaire employé pour cette thèse est le dictionnaire Delaf. Le Delaf, ou dictionnaire des formes fléchies du français, est un dictionnaire électronique constitué par l'Université de Marne-la-Vallée et composé de 683 824 entrées simples pour 102 073 lemmes différents et 108 436 entrées composées pour 83 604 lemmes différents (Courtois, 2004). Les processus d'identification et d'application du corpus d'exclusion constitué par ce dictionnaire ont été menés à bien à l'aide de l'outil Unitex (Paumier, 2010). Cette procédure informatique permet de récupérer un fichier contenant les mots connus du dictionnaire et un fichier contenant les mots inconnus. Les fichiers de mots inconnus comptent 6.609 lexies pour la Belgique, 2.311 pour La Réunion, 2.036 pour le Québec et 1.177 pour la Suisse. Chaque liste de lexies inconnues a été triée manuellement : les prénoms et noms (par exemple, de personnalités), les unités issues de fautes de frappe de la part du transcripteur (pour les corpus belge et suisse), les erreurs du transcripteur automatique (pour les deux autres corpus) et les marques. Cette liste présente des unités appartenant à des systèmes linguistiques autres que le français (éléments allogènes), des créations lexicales (néologismes) et des unités régionales (régionalismes). Ce sont les listes de mots connus et inconnus du Delaf qui font émerger les hypothèses sur la langue.

Pour les parties consacrées à l'orthographe et à la syntaxe, nous avons au contraire choisi de partir des assertions faites au sujet de la Cémo et qui portaient sur la « destruction de la langue » (un nombre important de variations orthographiques et syntaxiques), de la confusion oral/écrit. Pour la syntaxe, nous avons donc répertorié les phénomènes majeurs que la littérature exploite lorsqu'elle tente de départager la syntaxe de l'oral et de l'écrit, notamment. Notre interprétation se gardera bien d'être dépendante de cette dichotomie et suivra plutôt le continuum allant de la proximité (*Nähesprache*, que d'autres nomment syntaxe de l'oral) à la

1. Tognini Bonelli (2001) va jusqu'à nier l'existence d'une théorie en dehors des données (en cela elle s'oppose à Chomsky pour qui les corpus ne peuvent jamais constituer une source d'information sur la nature du langage). Des auteurs tels que McEnery et Hardie (2012) réfutent cette assertion qu'ils trouvent beaucoup trop extrême.

distance (*Distanzsprache*, que d'autres nomment syntaxe traditionnelle, normée ou « de l'écrit »), suivant la terminologie de Koch et Oesterreicher (1985).

Il existe toutefois une limite théorique à cette scission. En effet, lors de l'analyse lexicale, nous abordons des phénomènes qui touchent éminemment à la syntaxe (par ex. les types d'alternance de codes, les structures de collocations, les expressions figées). De la même façon, la variation des structures syntaxiques est également étudiée en fonction du contexte lexical. Suivant la conception initiée par Harris (1965), puis conceptualisée par la théorie du lexique-grammaire (Gross, 1975), nous n'envisageons donc cette scission qu'arbitrairement en raison de nos deux méthodologies générales distinctes et complémentaires.

7.3 Outils d'analyse

L'importance de compléter les analyses qualitatives à l'aide d'outils, tels que les statistiques, a déjà été soulignée par maints auteurs : « A good corpus study [...] combines qualitative and quantitative research methods » (Meyer, 2002, p. 123). Bien entendu, l'étude statistique ne peut constituer, en linguistique, la seule base en vue de poser des conclusions scientifiques. C'est pourquoi nous prendrons soin, tout au long de notre investigation des données, d'allier les outils statistiques [2] à l'analyse manuelle, qualitative. Ainsi procèdent également Degand et Bestgen (2004, p. 49) lorsqu'ils énoncent clairement : « la complémentarité entre analyses de corpus qualitatives (traditionnellement manuelles) et analyses de corpus quantitatives (traditionnellement semi-automatisées) [...] en opérationnalisant de manière adéquate les facteurs linguistiques étudiés, l'on peut en quelque sorte aboutir à une analyse automatisée alliant quantité et qualité ». La construction des hypothèses linguistiques pour chaque partie se réalisera à partir de conclusions qualitatives empiriques manuelles qui seront ensuite infirmées, validées ou complétées par l'étude quantitative.

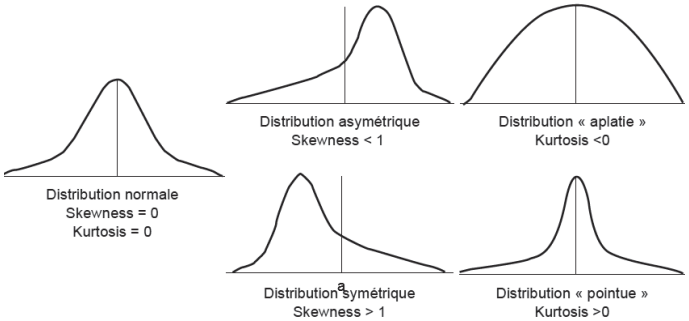
L'analyse statistique constitue un outil précieux pour l'exploration et l'étude de corpus de sms. On peut décliner les statistiques selon deux axes majeurs, suivant que l'approche se limite à décrire les données (statistiques descriptives) ou qu'elle vise à inférer des tendances globales à partir d'un corpus (considéré alors comme un échantillon représentatif d'une population plus large : statistiques inférentielles). Chacune de ces deux approches présente des avantages, mais également des limites méthodologiques. La première démarche, descriptive, permet d'explorer un corpus de manière systématique et fournit des indices résumant diverses caractéristiques du corpus. Ces statistiques, qu'elles soient de localisation ou de dispersion, restent sommaires : elles doivent être conçues comme un simple moyen d'épauler l'analyse qualitative du linguiste, que ce soit en l'informant d'une problématique qui lui

2. La quasi intégralité des calculs statistiques de la partie analytique de notre recherche a été effectuée à l'aide du logiciel SPSS (2010).

aurait échappé ou en lui fournissant une mesure précise d'un phénomène qu'il aurait repéré. Ainsi, on remarque dans le corpus belge de nombreux éléments allogènes. Après une exploration systématique des données, on observe que 83% des éléments proviennent de l'anglais. Reste au linguiste à détailler, voire à expliquer ce phénomène.

La seconde démarche, inférentielle, est plus ambitieuse, mais également plus risquée. Il s'agit de chercher à expliquer des phénomènes linguistiques en fonction d'une série de variables explicatives supposées entretenir une relation de corrélation avec le phénomène d'intérêt. En d'autres termes, il s'agit de procéder à des généralisations sur une population donnée à partir d'observations sur un échantillon de cette population. Étant donné que les variables à disposition pour ce type d'analyse sont par exemple de nature socio-démographique (sexe, âge, région d'origine, langues pratiquées...), on adopte alors une perspective sociolinguistique. Une telle démarche a déjà été mise en pratique dans Cougnon et François (2010) et Cougnon et François (2011). Lors de ces premières recherches, nous avons montré que la démarche inférentielle était limitée par la nature non représentative de notre corpus (cf. sous-chapitre 6.2.3) ; l'analyse de l'ensemble des données (belges, réunionnaises, suisses et québécoises) permet d'ajouter deux limites supplémentaires : la distribution non équilibrée des données sociodémographiques (les profils) entre les régions et la distribution non équilibrée des données pour chaque variable dans ces régions. Cette dernière limite ne permet pas toujours la mise en œuvre des tests statistiques, notamment quand un certain nombre d'observations (khi-carré) ou qu'une distribution dite « normale » (t-test) est requise pour effectuer le test (cf. 7.1). Dans ces cas précis, nous devrons choisir de ne pas effectuer le test par région et d'étudier l'ensemble du corpus international comme un tout non sécable. Ce choix n'entrave pas la nature des conclusions, mais force à mettre de côté certaines d'entre elles.

FIGURE 7.1 – *Représentations des distributions statistiques (Carricano* et al.*, 2010)*



Un des principes fondamentaux sur lesquels repose l'analyse principale de cette thèse est de montrer l'association entre deux variables (par exemple : absence ou présence du *ne* de négation en fonction de l'âge du locuteur). La simple statistique descriptive

permet de montrer les fréquences, les variances, les moyennes, etc. des observations et même de croiser les observations de deux variables entre elles. Cependant, ces premiers calculs ne permettent pas d'estimer la force d'association de deux variables, et surtout d'éliminer une association apparente qui, en réalité, n'est que le fruit du hasard. Nous décrivons ici succinctement les outils de la statistique inférentielle que nous emploierons, l'objectif de leur emploi et leurs limites. On distinguera les tests paramétriques, qui se basent sur la comparaison de données métriques, des tests non paramétriques qui, puisqu'ils ne se servent pas de données métriques, sont moins sensibles aux échantillons plus restreints.

1. Le test t pour échantillons indépendants (test paramétrique) a pour but de comparer des différences de moyennes entre deux types de population distincts (par exemple, les hommes et les femmes : nombre de sms envoyés par semaine en fonction du sexe du locuteur). Il s'applique lorsque l'on cherche à **comparer une variable qualitative et une variable quantitative**, lorsque la qualitative ne montre que **deux catégories** (ou « modalités ») différentes de données (par exemple : hommes et femmes). Pour appliquer un test t, il est nécessaire de s'assurer de la normalité de distribution des données quantitatives (test de normalité). Cette condition, dans notre cas, rend souvent le test inexploitable, étant donné l'hétérogénéité des données pour certaines variables (la distribution de l'âge, par exemple). On aura alors recours, lorsque cela se montre pertinent, à deux transformations permettant l'exécution du test *a posteriori* : le logarithme naturel et la mise en catégorie. Le log(n) est applicable lorsque la distribution est dite « asymétrique » (cf. 7.1). La mise en catégorie, applicable dans les autres cas, permet de regrouper différemment les données afin de rééquilibrer leur distribution (par exemple : réunir les observations des individus de plus de 45 ans, ayant participé plus faiblement à notre enquête, dans une même classe d'âge). Le test t s'effectue en deux temps : un test de Levene permet de juger de l'égalité des variances. En fonction de cette hypothèse (égalité ou non des variances), nous effectuons un test t pour l'égalité des moyennes, spécifique à cette hypothèse.
2. L'Anova (test paramétrique) repose sur l'analyse de la variance. Il s'applique lorsque l'on cherche à comparer une variable qualitative et une variable quantitative pour une variable qualitative montrant **plus de deux modalités de données**.
3. Le test U de Mann-Whitney (test non paramétrique) s'applique à deux échantillons indépendants. On l'emploie pour comparer une variable qualitative et une variable quantitative lorsque **la normalité de distribution des données quantitatives n'est pas assurée (même après transformations)**.
4. Le khi-carré, encore écrit chi-carré, khi-deux ou χ^2, bien que non paramétrique, permet de vérifier une hypothèse de recherche. Il s'applique à des variables nominales et ordinales comportant un nombre restreint de catégories ou de niveaux (par exemple, langue de l'emprunt en fonction du niveau d'étude). On l'utilise donc lorsque l'on désire analyser **deux variables qualitatives** (nous

y recourrons très souvent). Il est conçu pour déterminer si la différence entre deux distributions de fréquences est attribuable à l'erreur d'échantillonnage (le hasard) ou est suffisamment grande pour être statistiquement « significative ». On pose deux hypothèses : H0 (l'hypothèse nulle) et H1. Si la différence entre les deux distributions est réduite, l'hypothèse nulle (H0) est acceptée (il y a indépendance entre les deux variables). Si elle est rejetée, on parle de différence statistiquement significative entre les deux groupes. La plupart des statisticiens recommandent d'éviter d'utiliser le khi-carré lorsque le nombre de fréquences est peu élevé. Hughes et Grawoig (1971) insistent par exemple pour qu'aucune cellule du tableau ne contienne moins de 5 fréquences et que le total des fréquences ne soit pas inférieur à 50. Ce test limite donc certaines de nos analyses, comme explicité précédemment, lorsque les régions montrent un nombre insuffisant d'observations de la variable 1 en fonction d'une donnée socio-démographique minoritaire (par exemple : l'emploi d'emprunts grecs chez les enquêtés de plus de 40 ans). L'effet du khi-carré se calcule par l'analyse des résidus ajustés standardisés, qui permettent de détailler le résultat de dépendance en montrant dans quelles cellules se trouvent des écarts significatifs par rapport à la situation d'indépendance des variables.

5. Le coefficient de corrélation de Pearson mesure le degré d'association entre deux variables et permet d'obtenir un indice dont la valeur fluctue entre -1,0 et 1,0. Il s'applique lorsque l'on désire comparer **deux variables quantitatives** dont la distribution est normale.

Chapitre 8

Variation lexicale

Souvent langue varie,
Bien fol est qui s'y fie.
Klein (2012)

Sommaire

8.1 Questions générales de lexicologie 135
8.1.1 Lexicométrie 136
8.1.2 Cooccurrences 141
8.1.3 Néologismes 143
8.1.4 Conclusions 151
8.2 Éléments allogènes 154
8.2.1 Approche qualitative 154
8.2.2 Variation diastratique 166
8.2.3 Conclusions 182
8.3 Régionalismes lexicaux 183
8.3.1 Régionalismes de Belgique 185
8.3.2 Régionalismes de la Réunion 193
8.3.3 Régionalismes du Québec 197
8.3.4 Régionalismes de Suisse 202
8.3.5 Conclusions 209
8.4 Conclusions 212

8.1 Questions générales de lexicologie

L'objectif de cette partie consacrée à la lexicologie générale [1] est de donner un aperçu des caractéristiques linguistiques globales du corpus et des phénomènes qui seront traités plus en profondeur par la suite.

1. Le titre exact de ce sous-chapitre constitue un clin d'oeil au cours de linguistique que nous avons suivi durant nos études universitaires et pour lequel nous avons, par la suite, assumé les fonctions d'assistante.

8.1.1 Lexicométrie

La lexicométrie[2] est née dans les années 60 du croisement de l'influence des courants structuraliste et quantitativiste et de plusieurs disciplines : l'étude de texte, la statistique, l'analyse du discours et l'informatique, notamment. Le principal reproche fait à la lexicométrie concerne la définition du mot comme unité de base, comme échelle unique ignorant la phrase, la grammaire, la sémantique du texte. Les travaux des Français André Salem et Jean-Philippe Genet, notamment, ont permis de résoudre en partie cette tension, grâce, par exemple, à des mises en contexte de l'unité dans la phrase (concordances) ou à des rassemblements d'unités sémantiquement semblables (groupes de forme). Aujourd'hui, les traitements lexicométriques s'attachent donc à éclairer la répartition et le fonctionnement des unités lexicales dans un corpus, via le recensement par exemple des types, des tokens, du rapport type-token, des mots-clés et des hapax, sans négliger pour autant les unités supérieures au mot, l'agencement des unités lexicales et leur sens. Cette double méthode appliquée au corpus de sms révèle des tendances générales sur le lexique propres au corpus étudié, qui semblent tout à fait pertinentes pour constituer un point de départ à l'analyse des données. Les chiffres contenus dans cette partie ont été calculés par le biais du logiciel WordSmith Tools qui, sur la base du corpus de sms transcrit, découpe les unités lexicales au niveau des espaces et de la ponctuation (caractères *délimiteurs*), excepté le tiret.

Nous avons vu précédemment (partie 6.1.1) le nombre de types et de tokens pour chaque région, ainsi que leur rapport[3](Johnson, 1944). Le nombre d'hapax[4] d'un texte permet, en complément du rapport type-token, d'évaluer sa richesse. L'ensemble de ces chiffres et le rapport hapax-tokens[5] sont présentés dans la Figure 8.1.

FIGURE 8.1 – *Principales caractéristiques lexicométriques des 4 corpus de sms francophones*

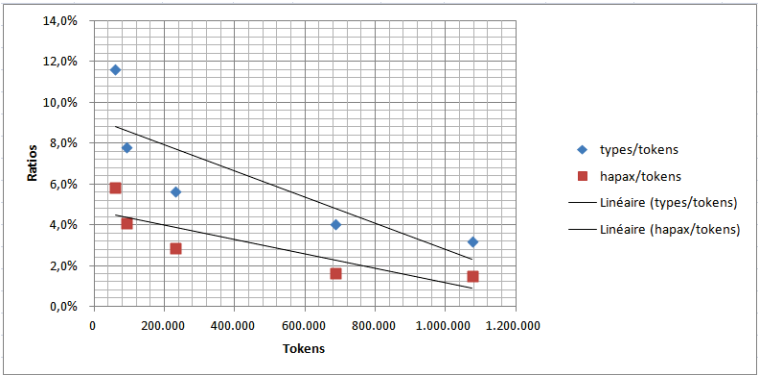
	tokens	types	TTR	Std TTR	hapax	hapax/tokens
BE	688.550	24.265	4,0%	40,7%	11.096	1,6%
LR	233.285	13.080	5,6%	39,1%	6.610	2,8%
QU	60.809	7.056	11,6%	43,1%	3.530	5,8%
CH	94.398	7.395	7,8%	40,3%	3.841	4,1%
Global	1.077.042	34.380	3,2%	40,5%	15.687	1,5%

Le rapport type/token standardisé (*std TTR*[6]) montre une similarité des chiffres dans les corpus belge, suisse et réunionnais et un pourcentage plus élevé dans le

2. Encore appelée statistique linguistique, statistique lexicale ou linguistique quantitative.
3. TTR = (nombre de types/nombre de tokens) * 100.
4. Du grec $\alpha\pi\alpha\xi\ \lambda\varepsilon\gamma\mu\varepsilon\nu\omicron\nu$, *hapax legomenom*, « dit une seule fois ».
5. Notons donc que nous évaluons le rapport entre formes fléchies, ce qui peut différer largement du rapport sur les formes canoniques (des chiffres dont nous ne disposions pas).
6. La moyenne des TTR de chaque séquence de n mots (n étant souvent égal à 1.000)

corpus québécois qui disposerait donc d'une richesse lexicale plus élevée. Une comparaison du rapport hapax/tokens basé sur un million de mots (corpus de sms francophones total) avec celui d'un autre corpus écrit de même taille met en avant la faiblesse du rapport pour le corpus de sms. En effet, une analyse identique réalisée sur un corpus de type journalistique (échantillon aléatoire du quotidien belge *Le Soir* de 2004 à 2008 [7]) livre un rapport type/token de 5,88% et un rapport hapax/tokens de 2,62%. Nous obtenons des chiffres beaucoup plus faibles (respectivement 3,2% et 1,5%), ce qui nous encouragerait à penser que le lexique du corpus de sms est lexicalement moins diversifié que celui du corpus journalistique de comparaison.

FIGURE 8.2 – *Rapports types/tokens (TTR) et hapax/tokens (HTR) dans les corpus de sms*



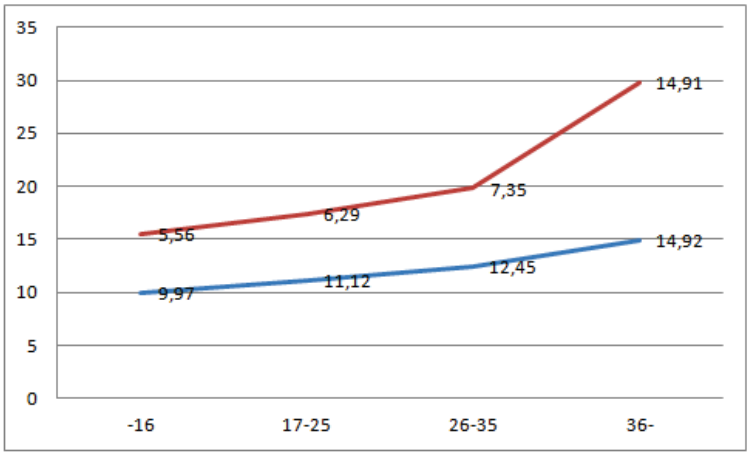
La Figure 8.2 qui présente les rapports types/tokens et hapax/tokens permet trois types de commentaires. Premièrement, on constate distinctement que les deux types de rapport (types/tokens et hapax/tokens) suivent la même courbe. Deuxièmement, la courbe montre une décroissance des ratios lorsque la taille du corpus croit et inversement. Il s'agit d'une tendance s'approchant de la normale. Enfin, on remarque que, contrairement à une tendance tout à fait normale, celle étudiée, même si elle tend à se stabiliser, ne se stabilise pas tout à fait.

La Figure 8.3 met en avant les ratios lexicaux en fonction de l'âge. Afin d'évaluer les différences objectives dans les pratiques lexicales entre générations, nous avons opté pour un ré-échantillonnage du corpus en fonction de l'âge. Nous avons aléatoirement extrait 2.000 sms pour chaque tranche d'âge. Nous constatons tout d'abord l'écart dans le nombre de tokens pour 2.000 messages entre les moins de 25 ans et les plus de 25 ans. Ce constat souligne que les plus jeunes exploitent de façon maximale l'espace que leur offre le médium de communication sms. Les plus âgés, quant à eux, sont moins loquaces. Ensuite, on observe une différence dans les unités linguistiques uniques, les hapax : les plus de 35 ans montrent un taux d'hapax beaucoup plus élevé. Il semblerait donc que le lexique utilisé dans les sms de cette classe d'âge soit plus diversifié que ne l'est celui des autres classes.

7. Merci à H. Naets pour la transmission de ces informations.

FIGURE 8.3 – *Rapports types/tokens et hapax/tokens en fonction de classes d'âge*

âge	sms	tokens	TTR	hapax	HTR
-16	2000	43383	9,97	2413	5,56
17-25	2000	43240	11,12	2723	6,29
26-35	2000	36761	12,45	2703	7,35
36-	2000	30126	14,92	4494	14,91



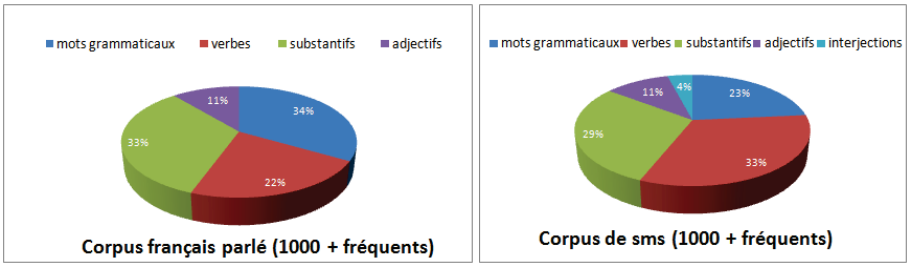
Un panorama des unités lexicales simples aux fréquences les plus élevées permet d'aborder une autre facette des corpus de sms. La Figure disponible en Annexe A.5 présente les 190 unités simples de **plus de 1.000 occurrences**[8]. Le taux de mots grammaticaux (déterminants, prépositions et pronoms) et de mots-outils (conjonctions) est évidemment très élevé (40%). Par ailleurs, on retrouve un pourcentage important d'adverbes (22%) : il s'agit, entre autres, d'adverbes de négation (*pas*, *rien*), de temps (*quand*, *alors*, *demain*, *bientôt*, *hier*), de lieu (*là*), de manière (*comment*) et d'intensité (*très*, *trop*). Les verbes représentent pas moins de 24% des unités les plus fréquentes : on recense surtout les verbes-outils *être* (*est*, *suis*, *es*, *était* et *être* lui-même), *avoir* (*ai*, *as*, *avoir*, *avais*), *aller* (*va*, *vais*, *aller*, *vas*), *faire*, *falloir*, *venir*, *savoir*, *vouloir*, *voir*, *dire*, *penser*, *pouvoir* ainsi qu'un seul verbe exprimant une réelle activité : *passer* (« tu passes dans la journée ? »). Les verbes concernent les première, deuxième et troisième personnes du singulier et sont majoritairement conjugués au présent, à l'imparfait ou à l'infinitif. **Le futur et les formes du pluriel sont donc absentes**. **Les substantifs sont moins nombreux** : ils constituent 10% de ce répertoire des unités les plus fréquentes et concernent les domaines de la temporalité et des sentiments, tels que *bisous*, *soir*, *nuit*, *soirée*, *journée*, *heure*, *amour*, *temps*, *week-end*, etc. Notons que l'**on ne retrouve pas dans cette liste les substantifs génériques** comme *truc*, *bidule*, *machin*, pourtant typiques de contextes linguistiques relâchés. Les adjectifs, pour leur part, concernent 6% du corpus étudié.

8. Cela concerne 150 unités.

On notera également dans ce tableau la prépondérance du pronom *on* (8.254 occurrences) par rapport à *nous* (1.278 occurrences) et la prépondérance du pronom *ça* (11.644 occurrences), alors que *cela* n'apparait pas dans la liste.

Dans Rivenc (1971), l'auteur présente les 1.000 mots les plus fréquents d'un corpus du français parlé, qui se répartissent comme suit : 33% de mots grammaticaux, 33% de substantifs, 22% de verbes et 11% d'adjectifs. La Figure 8.4 résume les deux distributions.

FIGURE 8.4 – *Classements grammaticaux du lexique (1.000 mots les plus fréquents du français parlé – 1.000 mots les plus fréquents du corpus de sms)*



Nous exécutons la même analyse sur **les 1000 unités les plus fréquentes du corpus de sms** en optant pour l'inclusion des adverbes dans la classe des mots grammaticaux. L'écart avec le corpus du français parlé est faible. Cependant, on note une différence majeure : **les verbes restent une classe plus représentée que les substantifs**. Malgré le biais entrainé par la différence entre les dates des collectes (1971-années 2000), nous aurions tendance à attribuer ce phénomène à la différence entre les types de corpus (oral/eSMS).

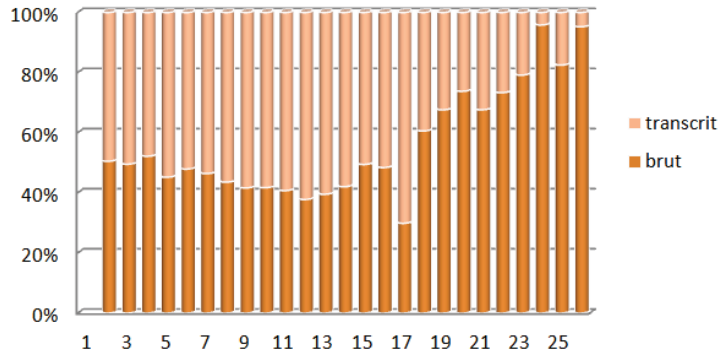
Étant donné la nature de notre corpus, nous ne pouvons pas compléter cette section d'un panorama relatif aux phrases (absence de ponctuation non réintégrée dans la transcription). En raison de cette même nature qui accorde une grande importance à l'économie, nous choisissons plutôt de donner les quelques chiffres concernant la longueur des unités. **La longueur moyenne** (en nombre de caractères) **d'un mot du corpus brut est de 3,62**. **Elle est de 3,82 pour le corpus transcrit**[9]. Il n'existe donc pas une très grande différence entre les deux ensembles d'unités. La répartition des mots en fonction du nombre de caractères est détaillée en Figure 8.5[10].

Cette répartition met en avant deux phénomènes majeurs. Premièrement, la proportion d'unités de taille inférieure à 18 caractères est approximativement la même dans les deux corpus ; ensuite, on note la nette supériorité, dans le corpus brut, des unités de longueur supérieure à 18 caractères. Ceci peut paraitre curieux étant

9. On notera la différence significative de ce résultat avec des chiffres basés sur d'autres genres de corpus, par exemple des corpus d'articles scientifiques : 10,79 caractères en moyenne (Loiseau *et al.*, 2006, p. 632).

10. Les valeurs exactes se trouvent en Annexe A.4

FIGURE 8.5 – *Moyennes du nombre de caractères par mot dans les corpus de sms brut et transcrit*



donné que l'on s'attend, dans les sms bruts, à une pratique abrégée (ou simplifiée) des mots. Nous l'expliquons cependant par la pratique courante d'agglutination des mots, que le logiciel WordSmith ne segmente pas et considère comme une seule et unique unité (par ex. *menervpa* pour *m'énerve pas*).

Nous clôturons cette partie lexicométrique générale avec un détail région par région des mots-clés. Nous appelons « mots-clés » les unités lexicales qui apparaissent statistiquement davantage dans un corpus donné que dans un corpus de référence. Vu qu'il n'existe pas de corpus de référence pour les sms en francophonie, nous avons pris le parti d'utiliser l'ensemble des corpus que nous possédons et, pour chaque corpus de le confronter aux trois autres aires de la francophonie (par exemple, pour étudier les mots-clés du corpus belge, comparer le lexique belge au lexique réunionnais-québécois-suisse). Les résultats des 100 mots-clés statistiquement plus importants pour chaque région sont détaillés en Annexe A.6. Ce tableau met en avant plusieurs phénomènes. Premièrement, on retrouve bien évidemment des noms de lieu tels que Liège, Bruxelles, Neuchâtel, Montréal, etc., propres à chaque région. Deuxièmement, et comme on le verra en profondeur dans la partie 8.3, les mots-clés de chaque région montrent de nombreux régionalismes (*savoir*, *pis*, *bec*, *cellulaire*, *gsm*, *salutations*, etc.) propres aux différentes aires de la francophonie. Troisièmement, le corpus réunionnais est fortement marqué par le créole et une graphie qui lui est propre, un « code créole » pour des mots appartenant souvent aussi au lexique du français « standard » (par ex. *fé* pour *fais*). De la même façon, le lexique du Québec est particulièrement marqué par les anglicismes. Enfin, on remarque des habitudes moins facilement explicables, telles que le sous-emploi du *ne* de négation à La Réunion ou le sur-emploi de *coucou* et *hello* en Suisse, des phénomènes que l'on tentera d'approfondir plus tard.

8.1.2 Cooccurrences

L'étude des cooccurrences revêt une importance majeure en lexicologie (L'Homme, 2000), comme en analyse du discours (Fiala, 1987) et en traitement automatique des langues (Daille, 2001). En effet, « il s'agit d'un système de particularités expressives liées aux conditions sociales dans lesquelles la langue est actualisée, c'est-a-dire à des usages » (Rey et Chantreau, 1989, préface p. IX) cité dans Dubreil (2008). L'étude des cooccurrences dans les sms, un médium écrit de la langue en usage, est donc pertinente. Elle peut révéler des phénomènes originaux ou confirmer des tendances perçues par ailleurs.

Derrière le terme *cooccurrence* se cache ce que certains auteurs nomment *collocation* : « ce terme recouvre en fait des phénomènes linguistiques très variés relevant de la phonologie, pragmatique, sémantique, syntaxique et même statistique [...] La collocation est la plupart du temps définie comme une association syntagmatique restreinte trouvant sa place entre le syntagme libre et le syntagme figé. » (Dubreil et Daille, 2005). Nous avons adopté, pour *cooccurrence*, la définition suivante, qui a le mérite de la clarté (là où souvent les auteurs sont assez confus et loquaces sur la question [11]) : « une suite de mots statistiquement significative et syntaxiquement définie dont le sens n'est pas toujours compositionnel » (Watrin, 2007). Mel'čuk (itre) l'explique de la sorte : « la cooccurrence de A et B : A n'apparaît pas dans les textes (ou n'apparaît que très peu) sans B ou vice versa ». En outre, cette définition convient parfaitement dans le cadre de notre recherche puisque nous présentons les résultats d'une méthode d'extraction des cooccurrences qui suit largement cette définition.

À l'aide du système élaboré par Watrin et François (2011), nous repérons les co-occurrences pour les structures nominales [12] de notre corpus. Le logiciel répertorie 31.906 unités polylexicales correspondant à des structures nominales. Nous trions cette liste en l'épurant des éléments mal reconnus (par exemple la structure verbale *fais de beaux rêves*) et de diverses coquilles dues à des erreurs dans la transcription des sms. Nous approfondissons ici les 100 unités dont le score est le plus significatif. La liste de ces 100 unités se trouve en Annexe A.7.

Les cooccurrences présentes dans les sms du corpus suivent la typologie suivante :

1. NA : *semaine prochaine*, *bisous doux* ;
2. A^x N : *petit coeur*, *meilleure amie*, *gros gros bisous* ;

11. Notamment en raison des diverses approches du phénomène mais également des divers critères d'identification des cooccurrences, entre autres en fonction de la définition de la notion de figement.

12. Est entendue comme structure nominale, toute structure comprenant un nom (et non pas dont le nom constitue la base). Ce choix d'étudier uniquement les structures nominales est arbitraire ; nous l'expliquons par le fait que, au vu des résultats obtenus pour un seul type de structure, nous ne pouvions envisager l'étude de toutes les structures.

3. NN : *clé usb*, *adresse e-mail* ;
4. NenN : *appel en absence* ;
5. NdeDetN : *grève de la STIB* ;
6. NàDetN : *mal à la tête*, *boite aux lettres* ;
7. NdeN : *bébé d'amour*, *minutes de retard* ;
8. NAdvA : *bisous tout doux*, *kiss tout plein*.

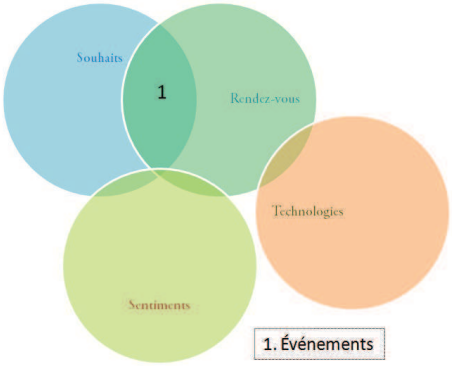
Le type le plus fréquent est AN (55%), suivi de NdeN (17%) et de NN (13%).

Dans cet ensemble de cooccurrences statistiques, nous retrouvons toute l'étendue du continuum du figement allant de suites de mots tout à fait libres (par ex. *bonne après-midi*, *gros gros bisous*), les unités apparaissant ensemble de façon significative dans ce corpus, à des suites tout à fait figées, véritables *collocations* (par ex. *coup de fil* ou *grand chose*). Entre les deux, on retrouve des combinaisons semi-figées (qui permettent un degré de liberté assez important) telles que *rayon de soleil*, *fête des mères* et *cadeau de Noël*.

Il est également possible de répertorier ces unités selon les thématiques auxquelles elles renvoient. Nous avons discerné 4 thématiques générales [13] : souhaits, sentiments, rendez-vous et technologie. Les souhaits portent sur des unités de temporalité (*bonne nuit*, *bonne soirée*, *bonne journée*, *bon week-end*) et sur des événements précis (*bon appétit*, *beaux rêves*, *bonne fête*, *bon amusement*). Au sein de la classe des événements, les souhaits d'encouragement sont nombreux (*bonne chance*, *bon courage*, *bonne merde*). Les sentiments réfèrent à des actions (*gros bisou*, *kiss kiss*, *gros becs*, *gros poutou*, *gros câlin*) ou à des appellations (*petit cœur*, *petit chou*, *bébé d'amour*, *meilleure amie*, *homme de ma vie*, *mi amore*). Les rendez-vous concernent des événements d'ordre général (*marché de noël*, *nouvel an*) ou sont liés à des unités de temporalité (*prochaine fois*, *semaine prochaine*, *fin de journée*, *vendredi soir*). Les unités liées à la technologie portent essentiellement sur la communication (*adresse e-mail*, *petit message*). La Figure 8.6 illustre la structure de cette organisation en 4 thématiques et les intersections entre thématiques. Ces intersections prouvent la redondance des 4 thématiques principales dans les cooccurrences. En effet, les souhaits, la prise de rendez-vous, la technologie et la communication de sentiments constituent des thématiques récurrentes qui s'additionnent et se mêlent dans un même message. Les cooccurrences, au sens de « suite de mots statistiquement significative », amorcent cette constatation qui sera approfondie par la suite.

13. Certaines unités n'entrent dans aucune de ces catégories (par ex. *mal à la tête*, *salle de bain*, etc.) mais elles sont peut nombreuses.

FIGURE 8.6 – *Thèmes principaux abordés par les cooccurrences issues du corpus de sms francophone*



8.1.3 Néologismes

8.1.3.1 Introduction et méthodologie

L'étude des néologismes s'est longtemps focalisée sur des productions écrites (ou retranscrites de l'oral) de type journalistique, littéraire et politique (Adaci, 2008 ; Cabré et de Yzaguirre, 1995 ; Tournier, 2002). Il existe à notre sens une frilosité [14] de la néologie envers la production informelle, relâchée. Cette attitude, probablement liée au conflit originel opposant linguistes descriptifs et normatifs, mène, au sujet de la néologie également, à un « fossé entre linguistes "laxistes" et défenseurs d'une langue prétendument toujours en péril et vouée au déclin » (Klein, 2006). L'intérêt pour la production spontanée touche plus précisément à la question de la différence entre créativité lexicale, dont chaque locuteur est (potentiellement) moteur, et évolution lexicale de la langue, avec des mots nouveaux entrant dans des dictionnaires de référence, des mots diffusés parce qu'ils répondent à un besoin d'un groupe social plus ou moins large (Sablayrolles, 2009). Nous partons du principe que l'étude des « candidats » à la néologie issus de la production spontanée n'est en réalité qu'une extrémité d'un continuum menant à l'entrée dans un dictionnaire ou à la standardisation institutionnalisée, en passant par une utilisation plus ou moins attestée par les médias et la littérature.

On ne trouve dans Sablayrolles (2000), au sein des six corpus utilisés pour le travail comparatif, qu'un seul corpus s'inscrivant pleinement dans la catégorie « spontanée » (le corpus scolaire) et un autre mêlant probablement la parole spontanée à la parole formelle (« lexies lues ou entendues çà et là ») [15]. D'ailleurs, comme

14. Nous employons le terme de « frilosité » parce qu'il évite une position extrême que nous ne désirons pas adopter, puisqu'il est des auteurs dont l'intérêt porte sur la créativité dans un contexte spontané (Boulanger, 1984 ; Cabré *et al.*, 2008).

15. De la même façon, dans Merle *et al.* (1987), cité dans le même ouvrage, les auteurs emploient comme corpus d'étude « tout ce qui leur tombe sous la main » (Sablayrolles, 2000, p. 250), mais il s'agit en réalité surtout de journaux, magazines et émissions qui tendent vers la formalité.

explicité un peu plus tôt dans Sablayrolles (2000, pages 250-251), les motivations des chercheurs en matière de néologie les conduisent souvent et logiquement vers des corpus spécialisés (dans un domaine particulier, qui a trait plutôt à la terminologie : vocabulaire de l'élevage des huitres, des remontées mécaniques, de la porcelaine, du rugby, de l'informatique ; cf. Humbley 1990), restreints (le style d'un auteur particulier ou une époque littéraire définie ; cf. Bar 1973) ou vers la néographie. Et c'est surtout la néographie qui a fait l'objet de l'intérêt scientifique au regard de la Cémo ; la créativité lexicale, pour sa part, y est encore peu étudiée.

Quelques études récentes portent sur les néologismes issus de l'Internet en général. Crystal (2001, pages 87-95) offre une typologie et des exemples intéressants en matière de néographismes, de composés, de dérivés, etc. Veale et Butnariu (2010) s'inscrivent dans une autre classe d'études qui vise à extraire automatiquement des néologismes à partir de l'Internet. On trouve encore ci et là des recherches portant sur la créativité lexicale dans le chat et la messagerie instantanée (par ex. Ruan (2012). L'étude des innovations lexicales dans les sms en particulier est abordée par Panckhurst (2009) (dans le « discours électronique médié » puis dans les sms : les anglicismes et les néographies), dans Baron (2008) (qui se limite à quelques exemples, surtout des acronymes et des sigles) et dans Fairon *et al.* (2006b) (troncations, sigles et acronymes). Certains de ces auteurs se focalisent surtout sur les néographismes, les sigles et ce qui tient plutôt du français familier ; chez d'autres, les exemples sont peu nombreux. Nous avons précédemment décidé d'exploiter plus systématiquement la voie de la néologie dans les sms en nous basant sur le corpus de sms belge collecté en 2004 (Cougnon et Fairon, 2010). En effet, il nous semblait important d'analyser les pratiques réelles de la population, la néologie dans « l'usage quotidien ». Nous avons ensuite élargi cette étude (Cougnon et Beaufort, 2011) grâce aux deux corpus de sms réunionnais et suisse, collectés respectivement en 2008 et 2009. Nous détaillons dans cette partie les deux études en question et les complétons à l'aide du corpus québécois.

Notre recherche a porté sur les créations lexicales formelles : des « mots qui n'existaient pas dans un état antérieur de la langue et qui sont obtenus, pour la plupart, par dérivation ou composition » (Sablayrolles, 2000, p. 43). Notre méthode exclut donc malheureusement les néologismes sémantiques et les néologismes formels avec homographes présents dans le dictionnaire de référence. Nous avons analysé les créations apparues de 2004 à 2010 (période qui couvre les différentes collectes de sms). Pour la mise en œuvre de notre méthodologie, nous avons, comme explicité dans la partie *Préliminaires*, tiré parti de la retranscription des sms en graphie standard, en appliquant sur ces transcriptions le dictionnaire Delaf, qui a servi de corpus d'exclusion. Sablayrolles (2008) montre l'insuffisance d'une méthode de repérage de la néologie uniquement basée sur les dictionnaires comme corpus d'exclusion. En effet, le tri manuel, les choix méthodologiques et l'intuition (Sablayrolles, 2000) viendront combler quelques difficultés que posent les dictionnaires : l'absence d'une lexie du dictionnaire n'en fait pas pour autant un néologisme (on

verra qu'il y notamment aussi des mots étrangers et des régionalismes que nous trions). De plus, nous employons un dictionnaire qui n'a pas été mis à jour depuis 2001, ce qui a pour mérite de ne pas exclure une série d'unités nouvellement entrées dans les dictionnaires et que nous voudrions tout de même repérer. Ce nouveau critère formel méthodologique précise la notion de néologisme dans le cadre de cette recherche : **il s'agit d'une « unité lexicale » n'existant pas dans le dictionnaire de référence Delaf avant 2001. Ainsi, pour nombre d'entre elles**[16], **le sentiment de nouveauté n'est pas forcément de rigueur**. Nous sommes tout à fait consciente de cette perte graduelle de néologicité et, par souci de clarté, lorsqu'une unité apparait effectivement dans un dictionnaire de référence plus récent (Rey *et al.*, 2011), nous la notons à l'aide du symbole ⊗ (par ex. *aprèm*⊗).

Une fois la liste de « candidats » établie, nous avons fléchi les formes (par ex. la conjugaison du verbe *phoner*) en ayant recours à des filtres morphologiques[17]. Grâce à Unitex, l'outil d'exploitation de corpus que nous avons présenté précédemment, nous avons appliqué cette nouvelle liste au corpus francophone total afin d'obtenir les résultats les plus complets possibles, notamment parce que les formes fléchies n'auraient pas été retenues au premier passage puisqu'elles auraient été confondues avec des homographes présents dans le dictionnaire. Par exemple, le verbe conjugué *texte* dans *il me texte tout le temps* n'aurait pas été retenu dans la liste des candidats car il est homographe du substantif *un texte* ; il est donc identifié au deuxième niveau lorsque nous cherchons les flexions du néologismes *texter*.

Après obtention de la liste des sms contenant (au moins) un néologisme, chaque « candidat » a été analysé en cotexte, en étudiant l'ensemble des environnements linguistiques du corpus dans lesquels la lexie se manifestait. Nous avons uniquement gardé les formes qui apparaissaient clairement comme des « candidats » à la néologie, et non comme des néographismes, des jeux sur la forme ou autres phénomènes qui mériteraient chacun une analyse à part entière, et qui ont par ailleurs déjà fait l'objet de diverses études (par ex. Fairon *et al.* 2006b, pages 31-48), mais que nous ne traitons pas ici.

Notre intérêt pour la néologie sera dirigé par les questions suivantes :
– Un corpus de sms est-il plus empreint de néologismes lexicographiés ou, au contraire, de créativités lexicales individuelles ?

16. On pensera par exemple à *petit déj*⊗, *appart*⊗, *aprèm*⊗ et *chatter*⊗.

17. Les filtres morphologiques, une option proposées par le logiciel Unitex, sont un type particulier d'expressions régulières qui permettent d'interroger le texte non pas sur des catégories grammaticales ou des unités exactes, mais sur des parties d'unités comportant des caractéristiques précises. Ainsi, nous avons par exemple utilisé <<^**phon**>>, qui permet de répertorier toutes les occurrences des formes commençant par *phon*, dont *phonais*, *phonerai*, etc. Ces filtres ne permettent pas de récupérer l'intégralité des formes fléchies concernées mais recouvrent une grande partie de celles-ci. En outre, ce système engendre du bruit (on obtiendra par exemple *phonétique*) que nous nettoyons manuellement à l'étape suivante.

– Ces « SDF » (sans dictionnaires fixes,Sablayrolles 2009), s'ils sont attestés, appartiennent-ils au langage courant,au français familier en général ou sont-ils spécifiques des sms et de la Cémo ?
– Les créations lexicales endogènes attestées dans les sms sont-elles influencées par les langues étrangères ?
– Les « candidats » à la néologie de 2004 sont-ils toujours d'actualité en 2008, 2009 et 2010 ?
– L'innovation lexicale diffère-t-elle d'une région à l'autre de la francophonie et d'un sous-groupe de la population à l'autre ?

8.1.3.2 Approche qualitative

Il ressort d'un aperçu global de la répartition des sms contenant au moins un néologisme et ceux n'en comportant aucun (Figure 8.7) que 3,2 % des sms sont concernés par ce phénomène. 1.622 messages comportent donc une innovation lexicale. Les unités en question, au nombre de 93 (types), sont répertoriées, par région, dans les tableaux A.18, A.19, A.20, A.21 et A.22, en annexe.

FIGURE 8.7 – *Effectifs et pourcentages des sms contenant (1) ou non (0) un néologisme*

Néologisme

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	49657	96,8	96,8	96,8
	1	1622	3,2	3,2	100,0
	Total	51279	100,0	100,0	

Les tableaux reprennent le « candidat » à la néologie, le type de néologisme (le procédé de création dont il est issu), le nombre d'occurrences dans le corpus pour ce « candidat » et un exemple de cotexte du « candidat » dans le corpus. En ce qui concerne les procédés néologiques, nous avons opté pour une typologie générale simple reprenant les types majeurs suivants [18] :

1. Dérivation (que nous abrègerons en dériv) (ceci comprend les dérivés d'éléments allogènes) ;
2. Composition (compo) ;
3. Agglutination (agglut) ;
4. Conversion (conv) ;
5. Troncation [19] (tronc) ;

18. Les néologismes par emprunt (sans modification graphique) ont été retirés de cette liste puisque les éléments allogènes constituent un sous-chapitre à part entière de ce travail.

19. On ne parlera donc pas ici de syncopes puisqu'elles participent plutôt, dans le contexte des sms, à de la néographie, que nous n'étudions pas ici. Nous sommes consciente qu'en règle générale, la néographie ne devrait pas être écartée d'une étude néologique. C'est bien le contexte particulier de la Cémo qui nous encourage à le faire.

6. Mot-valise (m-v) ;
7. Siglaison (sigl), qui comprend le sigle lui-même et l'acronyme ;
8. Adaptations graphiques (adapt) des éléments allogènes au système phonétique de la langue emprunteuse.

Une analyse transversale des « candidats » à la néologie issus des tableaux fait ressortir plusieurs tendances. On remarquera tout d'abord **le nombre élevé d'hapax** (65 hapax pour 93 lexies au total). Ce phénomène est la conséquence de plusieurs facteurs, tels que la taille limitée du corpus ou encore le contexte ludique dans lequel ces formes sont apparues. Malgré la prépondérance des hapax, notre corpus total a montré la récurrence d'un certain nombre de créations lexicales : ainsi *aprèm*⊗ ou *aprem* (troncation et agglutination de *après-midi*, très répandue dans ce que l'on a appelé précédemment le français non conventionnel), avec 724 occurrences, *mdr* (siglaison à partir de *mort de rire*) qui apparait à 571 reprises ou encore *appart*⊗ (troncation de *appartement*) qui apparait 63 fois.

Concernant la typologie des néologismes, c'est **la dérivation** qui **est le premier procédé de création dans le corpus** ; par suffixation, avec *dodoter* (« dormir, faire dodo »), *dédramatiseur*, *bisette* (« petite bise »), *e-mailiser* (envoyer un courriel) ou *dégageable*, et par préfixation, avec le préfixe re- de réitération dans des exemples tels que *rebisous*, *rebooter*, *retexter* et le dé- de privation avec par exemple *désaoulé* ou *décacher*. Le corpus illustre un cas de dérivation où le suffixe est emprunté à l'anglais : *craquous* (« craquant », sur le calque de *fabulous*, *marvellous*, etc.).

La composition n'est pas un procédé très usité dans les sms (il concerne moins d'une dizaine de lexies). On recense *physiothérapeute*⊗, *megatop*, *aquaplouf*, *circocabaret* et *autogonflant*. Si ce procédé est plutôt rare, c'est probablement parce qu'il nécessite deux éléments formants, c'est-à-dire un nombre de caractères important. Pourtant, cette hypothèse voudrait alors que la troncation, un phénomène qui abrège le mot et diminue donc le nombre de caractères utilisés, soit plus populaire dans le corpus, ce qui n'est pas le cas. On recense moins de dix cas de troncations tels que *appart*⊗, *déj* (« déjeuner »), *mamme* (« mamelle » avec adaptation graphique), *bioanal* (« bioanalytique ») et *préloc* (*prélocation*, au sens de « prévente », par ex. de billets de concert). **Le gain d'espace ne semble pas (toujours) régir l'emploi de néologismes dans les sms**.

On notera également la quantité très importante de conversions (substantif-verbe) d'éléments allogènes : *looker*, *loveuse*, *débriefing*⊗, *chatter*⊗, *speecher*, *texter*, *sm-ser*, *frencher* et *watcher* sont des exemples pertinents mais non exhaustifs de ce phénomène.
Nous avons encore été surprise de rencontrer, dans une liste de 93 néologismes, pas moins de 4 mots-valises, procédé d'ordinaire peu répandu ; il s'agit par exemple de *bichique* (variante orthographique de l'agglutination de *biche* et de *chic*), *bisounours*

(issu de *bisou* et de *nounours*), *moisversaire* (jour de chaque mois durant lequel on fête un événement, créé sur le modèle de *anniversaire*) et *cybernamoureutisme* (issu de *cybernétique* et *amoureutisme*, lui-même néologisme dérivé de l'adjectif *amoureux*).

Enfin, si l'on analyse les domaines thématiques ou champs notionnels concernés par la création lexicale dans les sms, on note des néologismes répondant à des **technologies** nouvelles : les usagers du sms *textent*, *textotent*, *chattent*, *pluggent*, *deletent*, usent de *textos*⊗ et de *courriels*⊗. Certaines nouvelles réalités arrivent en francophonie en même temps que leur dénomination. Lors de la communication par sms, il est souvent question de qualificatifs concernant l'humain : un destinateur, un destinataire ou une personne tierce que l'on s'amuse à décrire ou à critiquer ; et on voit donc apparaitre des créations lexicales qui se réfèrent à des traits humains : on parle par exemple de *footeux*⊗ (amateurs de football), de *grunges*⊗ (conversion, personnes qui s'habillent et vivent comme les musiciens et chanteurs du style pop rock grunge), *bouratcha* (« bourré, ivre »), de *loveuse* (« amoureuses ») et de *craquous*. Enfin, le domaine des **sentiments** est particulièrement exploité dans la communication par sms. La néologie reflète cette tendance, notamment au travers d'hypocoristiques tels que *bichique*, des dérivés de *bise* ou *bisou* que sont *bisette* et *bisouille*, des composés tels que *bisounuit*, mais aussi de verbes tels que *dodoter* (« faire un petit dodo ») et *frencher* (« embrasser à la française, avec la langue »). De manière générale, on conclura que les champs notionnels abordés dans les néologismes touchent majoritairement au domaine comprenant la nature humaine et les loisirs, contrairement aux corpus journalistiques et politiques, par exemple, pour lesquels ce domaine n'arrive qu'en troisième position (Sablayrolles, 2000, p. 327).

8.1.3.3 Variation diatopique

Les néologismes extraits de notre corpus peuvent également être analysés sous un angle diatopique. En effet, tout d'abord, nous remarquons des différences de résultats en fonction de chaque aire géographique. Il existe des créations lexicales répertoriées dans trois corpus, mais pas dans un quatrième, telles que les célèbres *mdr* ou *aprèm*⊗. En effet, celles-ci n'apparaissent jamais au Québec, alors que la forme *lol* (équivalent anglais de *mdr*) apparait à 312 reprises. Cette absence pourrait s'expliquer par la taille réduite du corpus québécois qui ne permet pas de représenter l'intégralité des emplois, et on peut également s'interroger sur l'omniprésence de l'équivalent anglais dans ce corpus de petite taille. De la même façon, *texto*⊗ apparait partout sauf en Suisse, bien qu'en Belgique son équivalent anglais soit largement préféré (5 occurrences pour *texto*⊗ contre 776 occurrences pour *sms*). Le corpus réunionnais est le seul qui atteste la dérivation de *texto*⊗ qu'est *textoter*, et partage avec le corpus québécois la conversion *texter*. *Courriel*⊗, équivalent de *email* et abréviation de *courrier électronique*, ne se trouve également que dans le corpus québécois. Ces constatations sont révélatrices d'une différence dans les variétés du français étudiées : les nouvelles réalités sont

nommées différemment dans les diverses aires de la francophonie, en fonction des coutumes locales et de la politique linguistique du pays. De la même façon, nous trouvons *préloc* en Suisse, apocope évidente de *prélocation* qui ne se trouve pas dans le corpus belge où l'on emploie plus volontiers sa variante *prévente*. Tel est également le cas du verbe *soupercommuter*, hapax, composition à partir des lexies belges *souper* (« diner ») et *commu* (« cuisine et living communautaires dans les dortoirs étudiants » en Belgique), qu'il n'est pas étonnant de ne trouver dans aucun autre corpus que le corpus belge.

D'un point de vue grammatical, en nombre d'occurrences et tous corpus confondus, **nous constatons que plus de la moitié des néologismes sont des substantifs, la deuxième catégorie la plus fréquente étant l'adjectif, suivi du verbe**[20]. Or, ici aussi, si on adopte un point de vue diatopique, les aires géographiques se distinguent les unes des autres. En effet, dans le corpus réunionnais et québécois, le verbe passe en deuxième position, tout comme dans le corpus suisse, où l'on ne rencontre d'ailleurs aucun néologisme adjectival.

Nous recensons également des phénomènes néologiques évoluant dans le temps. Il serait dangereux d'interpréter l'absence dans les corpus de 2008 et 2009 de certains néologismes apparus dans le corpus de 2004 comme une disparition réelle, puisque ces absences peuvent résulter, comme nous l'avons déjà dit, de la différence de taille de corpus ou, comme nous venons de le voir, de la variation diatopique. Toutefois, nous pouvons mettre en évidence deux phénomènes significatifs. Nous remarquons que certains néologismes de 2004 sont entrés dans le dictionnaire par la suite. Ainsi, nous trouvons sans surprise les plus récurrents *aprèm*⊗, *appart*⊗, *chatter*⊗, *texto*⊗, mais nous y découvrons également des hapax de 2004 : *débriefing*⊗ et *squatteur*⊗. Ensuite, soulignons que certaines lexies très récurrentes et présentes dans trois aires de la francophonie, avec une fréquence identique, n'ont jamais été lexicographiées ; c'est le cas du sigle *mdr*. Il s'agit peut-être de créations lexicales trop spécifiques à la communication médiée par ordinateur pour être considérées par la lexicographie générale.

8.1.3.4 Variation diastratique

Nous allons à présent examiner la répartition des emplois de néologismes en fonction du profil des auteurs des messages. Nous effectuons donc des tests sur les variables *âge*, *sexe*, *langue première*, *occupation*, *éducation*, *nombre de sms par semaine*, *chat*, *forum* et *email*.

Comme nous le verrons plus en profondeur dans les parties détaillées sur les éléments allogènes ou les régionalismes, la variable *âge* est difficile à traiter étant donné sa distribution non normale (cf. Annexes A.8). Nous avons donc recours

20. Notons que cet ordre de fréquence N–A–V suit les tendances générales énoncées notamment dans Sablayrolles (2000, p. 318).

à une transformation de la variable. Afin de créer une distribution normale (cf. Annexe A.9), nous regroupons les âges comme suit : 1 (10 à 15), 2 (16 à 17), 3 (18 à 19), 4 (20 à 22), 5 (23 à 27), 6 (28 à 40) et 7 (41 à 70). Le résultat du test de Levene (Annexe A.10) prouve qu'il y a une égalité des variances entre les groupes (sig. = 0,096). Le test t prouve une dépendance (la valeur p[21] est inférieure à 0,05) entre les variables *âge* et *emploi d'un néologisme*. Au vu des variances des deux groupes, on peut dire qu'il existe une répartition des âges plus étendue pour les sms ne comprenant pas de néologisme que pour les messages en comprenant (pour lesquels les données sont plus ramassées). Nous pouvons donc conclure que la distribution (en Annexe A.9) des **sms comprenant des néologismes concerne une classe d'âge plus restreinte** (les valeurs sont plus ramassées) que celle des sms n'en comprenant pas (la distribution est plus homogène ou variée) : les plus jeunes font preuve de davantage de créativité.

La seconde variable quantitative (*nombre de sms par semaine*) montre une distribution normale. Un test t peut donc directement lui être appliqué (cf. Annexe A.11). Le résultat du test prouve que les médianes des groupes sont inégales (p = 0). Au vu des variances des deux groupes, on peut dire qu'il existe une un peu plus grande diversité du nombre de sms envoyés par semaine pour les sms comprenant un néologisme que pour les messages n'en comprenant pas, et pour lesquels les données sont plus ramassées.

Est-ce que le fait d'être un homme ou une femme modifie l'usage des néologismes dans les sms ? Nous allons employer un test khi-carré, dont le détail se trouve en Annexe A.12. Le khi-carré nous donne une valeur p inférieure à 0,05, ce qui signifie qu'il y a dépendance entre les deux variables *néologisme* et *sexe*. Si l'on regarde de plus près les résidus ajustés, on découvre que **le taux de femmes usant de néologismes est plus important que celui que l'on aurait attendu pour une distribution équilibrée**. Au contraire, **le taux d'hommes est inférieur au taux attendu**.

Nous effectuons le même test pour étudier le lien entre niveau d'études du locuteur et emploi ou non de néologismes dans les sms. Le détail du test se trouve en Annexe A.13 et montre une valeur p inférieure à 0,05, ce qui signifie que les variables *néologisme* et *niveau d'études* sont dépendantes. Le détail des résidus ajustés montre deux nettes tendances : **le taux d'étudiants et de diplômés (dernier diplôme obtenu) du secondaire supérieur usant de néologismes est plus important que celui que l'on aurait attendu**, alors que **le taux de diplômés du post-secondaire usant de néologismes est inférieur au taux attendu**[22].

21. Il s'agit de la probabilité de faire erreur en rejetant H0.

22. Nous avons également exécuté un test khi-carré pour évaluer la dépendance entre les variables *types de profession/occupation* et *emploi de néologismes*. Celui-ci nous amène à la conclusion suivante : le taux d'étudiants et de chômeurs usant de néologismes est plus important que celui que l'on aurait attendu et le taux d'employés dans le secteur des services usant de néologismes est

L'analyse suivante, basée sur la langue première des auteurs des messages, révèle un résultat plus intéressant. En effet, le test khi-carré est appliqué aux deux variables avec le regroupement pour la variable *langue* suivant : 1 (français) et 2 (autres langues). Le résultat du test khi-carré disponible en annexe A.14 montre une dépendance des deux variables ($p < 0{,}05$). Seulement, là où nous aurions attendu un taux plus élevé chez les auteurs de langue première française, qui auraient davantage tendance à mieux manipuler la langue française, y compris en matière d'innovation lexicale, nous observons le phénomène contraire : **le taux de locuteurs de langue première française (uniquement) usant de néologismes est moins important que celui que l'on aurait attendu** alors que le taux de bilingues ou de locuteurs de langue première autre que le français et usant de néologismes est supérieur au taux attendu. On pourrait donc s'interroger sur les contextes linguistiques propices à une plus grande créativité lexicale et **se demander si la compétence en une langue n'est pas un facteur moins important** pour la créativité lexicale **que ne l'est la « désinhibition » des locuteurs non-natifs** ou bilingues **face à une langue qui leur semble peut-être moins monolithique** et « inaltérable ». Des études plus approfondies sont évidemment nécessaires afin de confirmer ou infirmer cette hypothèse.

Nous clôturons l'analyse statistique des néologismes avec une étude de la dépendance entre la variable *néologisme* et les diverses pratiques de la Cémo (chat, forum et courrier électronique). Le résultats des trois tests khi-carrés (disponibles en Annexes A.15, A.16 et A.17) prouvent l'indépendance de ces trois variables avec la variable *néologisme*. **Il n'y a donc pas de taux d'emploi des néologismes plus importants en fonction de l'utilisation du chat et du courrier électronique, ni de la fréquentation des forums de discussion**.

8.1.4 Conclusions

L'approche lexicométrique du corpus de sms a permis de défricher le terrain de l'étude lexicale des sms en livrant de larges tendances. L'étude des types, tokens et hapax a d'abord mis en avant la relative pauvreté lexicale du corpus de sms, notamment par rapport à un autre corpus d'écrit plus normé, ainsi qu'une plus grande richesse lexicale pour la population plus âgée.

La répartition des catégories grammaticales du lexique le plus courant dans les sms montre une supériorité du verbe par rapport au substantif, ce qui est inhabituel, notamment par rapport à la répartition type pour un corpus du français parlé où le substantif prévaut. La fréquence élevée des verbes dans le corpus est également soulignée par Biber et Conrad (2009, 207), pour qui « la fréquence élevée de verbes

inférieur au taux escompté. Toutefois, nous ne préférons pas prendre en considération ce résultat étant donné les problèmes liés à l'enquête (catégorisation de cette typologie) que nous avons relevés au chapitre précédent.

dans les sms n'est pas surprenante, étant donné l'importance des propositions verbales, y compris dans le cas de messages très courts » [23]. Ce « nouveau registre » que forme la communication par sms est ainsi, selon les auteurs, à rapprocher de registres plus traditionnels, tels que la conversation.

L'étude des cooccurrences sur les structures nominales a mis en lumière la prévalence des structures de type AN, NdeN et NN. Ces structures renvoient à des thématiques constantes, que sont les domaines des souhaits, des rendez-vous, de la technologie et, surtout, des sentiments, qui s'entremêlent dans un même message.

Nous avons vu que les créations lexicales dans les sms étaient clairement influencées par les langues étrangères (en particulier l'anglais) dont de nombreux mots sont dérivés en suivant les règles traditionnelles de la dérivation du français. Les néologismes répertoriés peuvent aussi être regroupés en 3 groupes thématiques : les nouvelles technologies, les portraits humains et les sentiments. Hormis le cas de *mdr*, il ne semble pas exister de lexique néologique spécifique aux sms ou à la Cémo.

Environ 9 % des néologismes présents dans les corpus de sms francophones sont entrés dans un dictionnaire de référence depuis 2001. Au sein de l'ensemble des néologismes non lexicographiés, on atteste encore de néologismes partagés par plus d'un auteur de sms et d'hapax, qui constituent 69 % du corpus de néologismes, mais qui ne s'expliquent pas forcément par la créativité lexicale individuelle [24].

Nous concluons, à la suite de Sablayrolles (2000, p. 405), que « l'extension des néologismes dans le temps et dans l'espace connait également des variations remarquables. Certains restent individuels [*craquous*], d'autres circulent dans toutes les bouches [*appart*⊗]. Certains restent confinés dans un cercle étroit [*rebooter*], d'autres franchissent allègrement les frontières [*mdr*]. Les domaines où les néologismes apparaissent sont de plusieurs types ». En effet, nous avons trouvé de nombreux cas de langage de spécialité. Nous avons également relevé des cas limites en matière de style, pour lesquels la frontière entre néologisme et français familier et néologisme dans le français familier était floue.

Au niveau des différents groupes sociaux, nous avons mis en avant plusieurs conclusions majeures, dont les suivantes :
– au niveau de l'âge, la population usant de néologismes est plus concentrée que celle n'en usant pas : cette concentration montre une préférence des plus jeunes en matière de créativité lexicale ;
– le taux de femmes usant de néologismes est plus important que ce que l'on aurait

23. « The high frequency of verbs in text messages is not surprising, given the reliance on clauses even in very short messages ».

24. Le peu d'occurrences de néologismes (93) dans le corpus ne permet par d'assimiler les hapax du corpus à des unités éphémères issues d'un seul locuteur.

attendu si la distribution était équilibrée ;
– le taux de diplômés du post-secondaire usant de néologismes est moins important que ce que l'on aurait attendu ;
– le taux de locuteurs de langue première française usant de néologismes est moins important que ce que l'on aurait attendu, alors que les bilingues et non francophones atteignent, au contraire, un taux plus important ;
– il ne semble pas y avoir de lien entre les variables *néologismes* et *utilisation du chat*, *des forums* et *du courrier électronique*.

8.2 Éléments allogènes

Le contact des langues dans la communication par sms constitue un sujet peu exploré dans la littérature. En effet, l'étude de ce contact a longtemps été limitée à des milieux restreints (des « communautés »), via des enquêtes de terrain (enquêteur « participatif ») ou des corpus oraux enregistrés. Depuis, on a vu émerger un certain intérêt pour le contact des langues dans les nouveaux médias de communication, tels que le courrier électronique (Durham, 2003 ; Hinrichs, 2006 ; Huang, 2004) et, plus récemment, les sms (Deumert et Oscar Masinyana 2008, Deumert et Vold Lexander 2012 et Doehler 2011). Mais ces recherches, particulièrement au sujet des sms, sont rares, leurs conclusions ne sont pas facilement réutilisables dans d'autres contextes et sont peu généralisables. Nous essayons, dans cette partie, de pallier le manque de systématicité des résultats dans le domaine des sms, en offrant, sur base d'un corpus large et diversifié, des réponses qualitatives et quantitatives aux questions suivantes :

1. Quelles sont les langues en présence dans le corpus ? L'anglais y est-il prédominant, y compris dans des régions montrant d'autres langues en présence ?
2. Quelles sont les fonctions des mélanges de langues dans un même message ?
3. Comment les mélanges de langues se manifestent-t-ils structurellement dans les messages ? Quels types d'unité sont concernés ?
4. Peut-on constater des tendances de mélanges différentes en fonction du profil des auteurs des messages ?

8.2.1 Approche qualitative

8.2.1.1 Typologie

La liste d'unités lexicales inconnues du dictionnaire [25] et présentes dans les corpus de sms permet d'identifier une série de formes simples et composées appartenant à des systèmes linguistiques autres que le système du français [26]. En effet, malgré la demande explicite faite par les organisateurs aux participants de n'envoyer que des messages en français, 18 autres langues sont présentes dans le corpus de sms, mêlées au français. Les messages totalement rédigés dans une autre langue ont été écartés de l'analyse, seuls sont pris en considération les messages dont au moins une des

25. Nous avons expliqué précédemment que notre méthodologie d'application d'un dictionnaire du français standard au corpus de sms permettait de dégager une liste d'unités inconnues de ce dictionnaire. Ce dictionnaire n'a pas été mis à jour depuis 2001. Donc, si l'emprunt de langue, assimilé au lexique français, a intégré ce lexique avant 2001, nous ne le retrouvons pas dans la liste de mots inconnus et il n'est pas repris dans l'analyse. Seuls sont repris les emprunts de langue après 2001.

26. Notons que cette méthode ne permet pas de prendre en considération les homographes du français appartenant à un autre système. Cela étant, nous croyons ce phénomène assez rare, une analyse sur 500 messages n'en ayant identifié aucun ; nous pensons donc que le biais engendré est faible, même s'il n'est probablement pas inexistant.

unités lexicales appartient au français [27]. En 2011, nous avons entrepris d'étudier ce phénomène dans le corpus belge [28]. Nous étendons ici l'enquête à l'ensemble du corpus francophone.

Nous présenterons dans un premier temps les données auxquelles nous avons apposé l'étiquette d'« élément allogène ». Cette étiquette, très vaste, hyperonyme pour référer à l'emploi de plusieurs codes linguistiques au sein d'un même message, renvoie autant au phénomène d'emprunt (de langue), défini comme « un mot, un morphème ou une expression qu'un locuteur ou une communauté emprunte à une autre langue, sans le traduire » (Moreau, 1997, p. 136), qu'aux phénomènes de mélanges de langues dans un cadre discursif, tels que le code-switching et le code-mixing. Nous regarderons ensuite la façon dont ces unités pénètrent l'écrit francophone des sms (analyses qualitative et quantitative).

L'analyse globale des éléments allogènes montre que **90% des élément allogènes du corpus appartiennent à l'anglo-américain** et que **les 20 éléments dont le nombre d'occurrences est le plus élevé proviennent également de l'anglo-américain** (cf. Tableau 8.28, p. 179). Nous relevons 677 éléments allogènes uniques (types) pour 3.467 occurrences (tokens).

On distinguera donc l'emprunt de langue du processus de mélange de langues (lié au discours). Lors de l'emprunt de langue, il n'y a pas de « langue matrice » (Myers-Scotton, 1993) (dorénavant LM), car nous ne nous situons pas dans le discours d'un locuteur donné ; il y a une langue de base qui emprunte et une deuxième langue à laquelle on emprunte. La situation présente deux systèmes linguistiques en contact, qui influent l'un sur l'autre (généralement pas l'un autant que l'autre) si bien que certaines unités du lexique de l'un sont copiées dans le lexique de l'autre (quelquefois après adaptation graphique, phonologique ou dérivation). Notons encore que seul l'emprunt de forme (cf. exemple 7) est étudié, l'emprunt de sens n'étant pas détectable automatiquement. Précisons encore que nous abordons ici uniquement les emprunts aux langues vivantes (les références latines ayant été écartées).

(7) Hey ! Je vais quand même te raconter les **news**. J'ai rencontré un gars. (smsFQ)

Au sein de la classe des emprunts de langue (cf. la liste qui figure en Annexe A.23 établie manuellement à l'aide de Rey *et al.* (2011) et de Bob (2011), pour les emprunts dans une situation moins conventionnelle), nous notons une série d'unités

27. Ainsi, un message où le français est peu présent, tel que « Haha ! how did u know ? Non, sérieux I've been waiting to see u to tell u but i can't wait anymore. i want u to delete some1 from all u contacts » (smsFQ), est pris en considération dans nos résultats, même si le français n'y est pas la langue matrice.

28. Pour information, l'article est le suivant : Louise-Amélie Cougnon (2011). « 'Tu te prends pour the king of the world ?'. Language contact in text messaging context ». In Cornelius Hasselblatt, Peter Houtzagers and Remco van Pareren (éds.). Language Contact in Times of Globalization. Studies in Slavic and General Linguistic, 38, Rodopi, Amsterdam/New York, pp. 45-59.

lexicales renvoyant à une réalité étrangère ; la réalité à laquelle l'unité étrangère renvoie est également nouvelle pour la langue accueillante (*churros*, *nugget*, *pashmina*, etc.). Elle est appelée xénisme, lorsqu'elle est empruntée telle quelle et peu ou pas intégrée, c'est-à-dire lorsqu'elle est encore perçue comme étrangère par la communauté linguistique emprunteuse (cf. l'exemple 8, avec *aïd*).

(8) Coucou je te souhaite une très bonne fête "**mabrouk l'aïd**". Je te souhaite sincèrement tout le bonheur que tu mérites à toi ainsi qu'à toute ta famille (smsFB)

En ce qui concerne les mélanges de langues dans un cadre discursif, nous nous situons dans une approche interactionnelle (Blom et Gumperz, 1972) et distinguons trois cas particuliers : le code-switching, le transfert et le code-mixing. Comme nous l'avons dit, la langue de base des sms dans laquelle apparaissent des formes étrangères (dorénavant FE) est le français. Les FE sont mélangées avec la LM de diverses manières. Le premier phénomène de mélange se nomme code-switching (en français, « alternance de codes » ou « alternance de langues » ; cf. exemple 9). Bien qu'il soit souvent employé pour signifier tout mélange de langue dans un même discours, le sens plus spécifique de code-switching est le suivant : « la juxtaposition, à l'intérieur d'un même échange verbal, qui a toutes les marques d'un échange unilingue, de séquences de longueur variable où le discours appartient à deux langues, variétés ou registres de langues différents et dont les règles grammaticales sont respectées » (Cougnon, 2007).

(9) C'est gentil de préciser mais je me doutais bien qu'il ne lui avait pas parlé. Attente insupportable mais restons **open ! Be aware ! But he makes me crazy ! Yes he does !** (smsFB)

L'étiquette de code-switching s'oppose à celle de transfert, en ce que cette dernière renvoie à une situation d'échange verbal lors duquel la mélange de langues est limité à une construction bien définie et est suivie d'un retour prédictible à la LM. Cela s'écarte du code-switching si l'on prend en compte le processus de négociation [29] qui peut avoir lieu lors de tout acte de langage. Auer (1998) explique que lors du transfert, il n'existe aucun signe de négociation, il s'agit d'un passage vers une langue 2 qui n'est pas fait pour durer. La construction bien définie du transfert peut se limiter au mot-simple ou composé (transfert lexical ; cf. ex. 10), au syntagme (transfert de syntagme ; ex. 11) ou à un segment plus long, coïncidant ou non avec une proposition (transfert de segment ; ex. 12). **Le transfert est le phénomène le plus récurrent du corpus**.

(10) Hé vieux ! On a pire cartonné ! ø/ Hallucinant. :-D Toi ça **geht's** ou bien ? Tu fous quoi ? (smsFS)

(11) Salut **mi amor**. Moi ça va ! J'ai eu 7/20 en sciences (la merde). Y a plus que 3 jours sans se voir, ça va passer vite (smsFB)

29. Lorsque le nouveau code employé invite le locuteur lui-même et l'interlocuteur à l'utiliser.

(12) Et non ! Je dormais :-) ... Je bosse à la Maladière toute la journée ! Si tu veux boire un bon café et faire tes courses, **you're welcome** ! Bec et bonne journée (smsFS)

Le transfert permet des mélanges de langues répétés présentant une maitrise des deux systèmes linguistiques mélangés (cf. ex. 13).

(13) Tu peux toujours me ramener après ? **Btw** je t'emmène à 6h20 ish parce que **we got to get there early this week too** ! (smsFQ)

On opposera encore code-switching avec code-mixing, ou « mélange de codes » ; il s'agit d'un phénomène de contact des langues qui ne respecte pas les règles grammaticales de l'une ou des deux langues (cf. ex. 14, *You a eat* et *J'ai check*).

(14) **Asalam**. Ca va ? Bien révisé **today** ? **Hope** tu t'es reposé quand même un peu. **You** a **eat** ? Quoi ? J'ai **check** mail, **you** pas m'a écrit :(j'ai tes mails des photos. Bébé, **now** qu'elle dort. Moi aussi au lit, pour **read** bouquin recette lol.(smsFR)

8.2.1.2 Composition des éléments allogènes

La Figure 8.8 présente une classification des éléments allogènes selon leur composition en mots simples, composés, locutions et expressions. Notre analyse permet de confirmer les résultats de Doehler (2011) : **les mots isolés sont les plus fréquents**. Souvent, le mot simple peut être confondu, en contexte discursif, avec un transfert « monophrastique » (*Good* dans l'exemple 15), la FE ayant valeur d'énoncé (Doehler, 2011).

(15) **Good**. On s'en donne des **news**. :) (smsFQ)

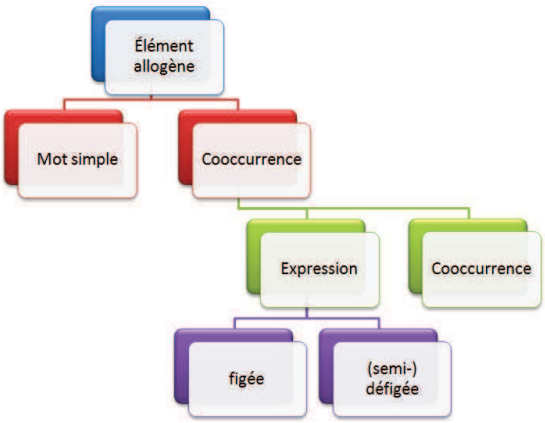
Les expressions figées constituent un ensemble non négligeable d'éléments allogènes du corpus de sms. On citera par exemple *geen probleem*, *good job*, *see you*, *hasta pronto*, *saha chribetkom*, *God bless (you)*, *by the way*, *here I come*, *to be continued*, *my god*, *good luck*, etc. (cf. ex. 16).

(16) **Good job** ! :) Comment tu sais que c'est le même ? Ils sont 2. (smsFQ)

Les cooccurrences [30] sont détectées automatiquement par le logiciel *AntConc* et permettent de mettre en avant des pratiques linguistiques récurrentes et statistiquement significatives.
Nous avons analysé les cooccurrences pour les 10 éléments allogènes les plus récurrents du corpus (cf. liste de la Figure 8.28, p. 179). Nous filtrons la liste proposée

30. Suites de mots dont l'apparition dans un texte est statistiquement plus forte qu'avec d'autres mots. En cela, il ne s'agit donc pas systématiquement de suites cohérentes pouvant mener à une sorte de figement ou à une collocation.

FIGURE 8.8 – *Classification des éléments allogènes selon leur type de composition*



en éliminant de celle-ci les éléments non pertinents[31]. Les résultats des dix cooccurrences les plus significatives pour chaque élément allogène sont présentés dans la Figure 8.9. *Today* ne montre que peu de co-occurrences statistiquement significatives et fonctionne plutôt de manière isolée. *Big*, pour sa part, apparait de manière significative avec les variantes de *bise*. Le nombre d'occurrences (60) de l'adjectif *new* est assez élevé, et les substantifs auxquels il s'associe sont très variables. On obtient le même genre de phénomène avec *my* (il se combine avec une large variété d'unités françaises ou anglaises) mais il montre une nette propension à s'associer davantage avec certains mots tels que *love* (24 occurrences) et *friend* (8 occurrences). L'adjectif *good* (47 occurrences), pour sa part, s'associe uniquement, d'après la liste des cooccurrences les plus récurrentes, à un autre élément allogène, et ce, pour marquer le souhait : *good luck* (16 occurrences), *good night* (14 occurrences), *good day* (6 occurrences) et *good morning* (2 occurrences). On retrouve également la forme *good* dans les expressions figées :*good news* (4 occurrences), *sounds good* (4 occurrences), *good job* (3 occurrences) et *all good* (2 occurrences).

Le cas du substantif *news* (201 occurrences) est tout à fait intéressant. Il apparait surtout en complément direct introduit par l'indéfini *des* (*avoir*, *donner*, *vouloir*, *attendre* et *prendre des news*), mais on l'observe également en tant que sujet (*les news vont vite*, 3 occurrences). Enfin, *sorry* (250 occurrences) présente également une distribution hétérogène. Il fonctionne la plupart du temps, associé à un second élément, soit sont redoublement (*sorry sorry*), soit un adverbe (*encore sorry*), en tant que groupe isolé. Mais il est également souvent la base d'une proposition entière (*sorry pour*, *sorry mais*).

31. Statistiquement, nous obtenons des suites de type *la new* où l'élément apparait souvent adjacent à un déterminant ou à une conjonction ; ces suites d'unités n'ont pas de cohérence sémantique et sont donc jugées peu pertinentes pour notre objectif de recherche.

FIGURE 8.9 – *Cooccurrences pour les 10 éléments allogènes les plus fréquents (lignes et colonnes par ordre décroissant)*

kiss	sorry	news	today	big	now	you	new	my	good
kiss kiss	sorry sorry	avoir des news	pas today	big kiss	now je	love you	new year	my love	good luck
gros kiss	sorry mais	donner des news	se voir today	big bisou	mais now	you are	new numéro	my friend	good night
big kiss	sorry pour	vouloir des news	avoir cours today	big biz	now on	see you	new an	my tel(éphone)	good day
kiss tout plein	sorry bisou	quelles news	pour today	big bises	now dodo	to you	new CD	my père	good news
kiss je t'aime	sorry NPERS	les news de	comment (ça) va today		dormir now	for you	new GSM	my best friend	sounds good
kiss bonne journée...	vraiment sorry	les news vont vite			rentrer now	are you	new version	my god	good job
lol kiss	sorry mon grand	news pour			écrire now	want you	new année	my dear	good morning
réponds kiss	sorry ma belle	good news			seulement now	you too	what's new	my bed	all good
merci kiss	encore sorry	attendre des news			répondre now	miss you		my angel	
fais de beaux rêves kiss	sorry biz	prendre des news			enfin now	with you		my way home	

8.2.1.3 Fonctions des mélanges de langues

Dans cette section, nous avons pour but d'établir la liste des fonctions des mélanges de langues dans un contexte de sms francophones. « L'alternance est fonctionnelle dans la mesure où elle accomplit un travail de signalisation » (Doehler, 2011, p. 52). Notre hypothèse est la suivante : **l'emploi de l'anglo-américain dans les sms parait répondre au besoin d'économie propre à ce mode de communication**, la variante anglaise étant plus courte en nombre de caractères que la variante française de l'unité concernée[32] (Bautista, 2004) On note très fréquemment l'emploi des formes *happy*, *ask*, *maybe*, *next*, *now*, *send*, *sorry* et *today*. Cette pratique de l'emploi de FEs anglo-américaines lorsqu'elles sont plus courtes que leurs variantes francophones contamine également les registres plus familiers et la néologie avec des formes récurrentes telles que *thanks* (pour « thank you »), *gonna* (pour « going to »), *wanna* (pour « want to »), *aka* (pour « also known as ») et *asap* (signifiant « as soon as possile »).

Pourtant, cette pratique n'est pas constante : nous pensons sur base de l'observation d'un petit nombre d'exemples que **la volonté ou le besoin inconscient de transmettre un effet de style** (« faire cool ») **est quelquefois prioritaire au besoin de « faire court »**. D'autres fonctions semblent aussi motiver l'usage de l'anglo-américain. En outre, **pour les autres langues** (notamment romanes), nous pensons que **la volonté d'abréger ne régit pratiquement jamais l'emploi de FEs** (cf. *beijo beijo* dans l'exemple 17).

32. Voir l'exemple 17 où la forme *now* permet de maintenir le message en dessous des 160 caractères, ce que n'aurait pas permis la forme *maintenant*.

(17) Coucou amour de ma vie, ca va bien ? Moi ai fait moitié de sociologie du livre, **now** m'y remets... Et toi tu fais ? Tu me manques bcp fort :-(**beijo beijo**. (smsFB)

Pour vérifier cette hypothèse, nous nous aidons de la terminologie de Blom et Gumperz (1972) en ce qui concerne l'alternance situationnelle et métaphorique. Une alternance situationnelle est définie comme un changement de langue répondant à un contexte social ou linguistique (propriétés de l'interlocuteur, thème de la conversation [33], etc.) ; en d'autres termes, une alternance situationnelle est le produit d'une ou plusieurs circonstances sociolinguistiques. Une alternance métaphorique, par contre, n'est pas le produit de circonstances, elle est un outil en vue d'un objectif spécifique (attirer l'attention, flatter, etc.). Inspirée par Mackey (1968) et Auer (2000), nous listons une première série de types d'alternance situationnelle (Cougnon, 2007).

1. Changement de thème ;
2. Contraste séquentiel (entre une conversation en cours et une séquence explicative ou une remarque sur la conversation) ;
3. Changement de style individuel [34] ;
4. Changement de style situationnel (passage d'un discours informatif à un discours évaluatif, impératif, etc.) ;
5. Changement d'interlocuteur [35].

Les cinq types d'alternance situationnelle sont illustrés dans le corpus et **le changement de thème est le type le plus récurrent**. L'exemple 18 illustre un (1.) changement de thème entre deux réponses (*ben non* et *Vamos*) à des questions posées dans des messages antérieurs ou implicites.

(18) Ben non, j'ai largué **face to face** la dame après 6 longs jours, 1 record : −) ! **Vamos** pour demain mais j'ai cours jusque 6h15, grande folle ! Fais-moi signe quand t'es là ! (smsFB)

Le changement de thème peut évidemment porter sur deux déclaratives sans qu'il n'y ait de questions préalables (le locuteur mène sa réflexion seul, cf. ex. 19).

(19) Avec de l'espoir, le plus tard que je vais être au métro, c'est 6 h 30. **My stuff is already packed all I'll need to do is get changed and apply makeup** (smsFQ)

33. Notons que dans le contexte spécifique des sms, la conversation est limitée au seul message analysé puisque nous ne possédons pas les messages précédents, suivants, ni les messages en réponse de l'interlocuteur.

34. Cette notion de style inclut notamment les oppositions intimité/distance, formalité/informalité et solidarité/individualisme qui encouragent l'emploi d'un code plutôt qu'un autre.

35. Nous ne mettons pas cette fonction de côté car il arrive que dans un même message différents individus soient visés par l'acte de communication.

Les trois autres types, bien que moins courants, sont observables dans le corpus de sms. Ainsi, le (2.) contraste séquentiel entre la conversation en cours et une séquence explicative (cf. ex. 20, la séquence explicative introduite par *because*[36]).

(20) Les chouchoux on est encore au concert du Rideau, faudra vraiment que vous veniez une fois ! ! On ira pas vous rejoindre **because** demain je bosse... mais on s'appelle biz biz (smsFB)

On remarque qu'il s'agit d'un mélange de langue sur la base du transfert d'une seule lexie. Le recours à une seule FE suffit à marquer le contraste. Pour le (3.) changement de style individuel, on est plus rarement dans le cas du transfert d'une seule lexie[37] ; on rencontre d'avantage une alternance interphrastique comme dans l'exemple 21, où la forme néerlandaise *goeidag hoe gaat het met jouw* pour « bonjour comment vas-tu ? », d'un niveau assez formel (notamment comparée à une forme plus brève telle que *dag hoe gaat het* ? « salut comment ça va ? »), précède un franc *merde* qui tranche avec le style néerlandais.

(21) **Goeidag hoe gaat het met jouw** merde j'ai tellement étudié néerlandais regarde bien Star Academy mais ne te couche pas trop tard (smsFB)

Le (4.) changement de style situationnel se fait généralement via une suite de mots, voire une expression entière (cf. ex. 22, contraste entre une proposition déclarative et une impérative).

(22) c'est bon chez toi ? M'énerve je me suis trop bêtement bloqué ici j'arrive pas à trouver un lift mais clair je vais trouver. **Dammi un attimo** ! (smsFB)

Enfin, on note quelques rares cas de (5.) changement d'interlocuteur, comme dans l'exemple 23, où le message s'adresse à un seul destinataire ; le passage à l'italien élargit le nombre de destinataires (*a tutti*)[38].

(23) Bah j'étais droit de me dire la même chose. Non je vais pas y aller je crois, de toute façon elle va mettre le corrigé sur internet et ça sert à rien ces exercices ! On se retrouve pour les statistiques alors ? ! Je dis aux autres ! **Baci, a tutti** ! (smsFS)

La liste des alternances métaphoriques est, elle, inspirée de Gumperz (1989), Giancarli (1999) et Zongo (1993) et complétée par nous-même. Les auteurs des messages du corpus de sms recourent à des FEs afin de :

1. Marquer une citation ;

36. Notons que *because* a été intégré au français familier et est attesté dans certains dictionnaires de référence. Il fait donc partie des zones floues entre emprunt de langue et élément allogène d'un mélange de langues.

37. On en trouvera quelques cas ; cependant, moins il y a de contexte, plus il est difficile d'identifier le style des segments dans chacune des langues. Par ex. « **Enculado** je suis sur la piste » pourrait être interprété comme un changement de style individuel.

38. Notons qu'il est possible d'interpréter ce passage en italien d'une différente manière : *a tutti* peut être compris comme une « italianisation » de l'expression française abrégée *à tout'*.

(24) Pour NPERS je m'excuse d'utiliser ce téléphone-là : tu m'as dit un jour "**Ein Mädchen kann warten bis der Richtige kommt,aber bis dahin ein schöne Zeit mit dem Falschen verbringen**" ça te dit quelque chose ? (smsFR)

2. Attirer l'attention sur un élément important du message [39] :

(25) Yooo what bus are we taking ? Just to know **l'heure maximale où je peux arriver**. :p (smsFQ)

3. Marquer une interjection (26) ou un élément phatique [40] (27) :

(26) Pour une bonne nouvelle, c'est une bonne nouvelle ! **Great** ! On va pouvoir boucler le rapport ! On verra ça demain. Bonne soirée, NPERS (smsFB)

(27) Hello NPERS. Je rentre du cercle. Super amusé ! Comment va ? Mon souper ? Je m'attendais à ce que tu sortes avec nous après ; **Que pasa** ? Merci d'être venu ; je suis heureux d'avoir rassemblé tout le monde. Biz (smsFB)

Un très grand nombre de sms commencent par des salutations non francophones telles que *hola*, *dag*, *hi*, *salam* ou *saluti* [41]. On trouve également de nombreuses FEs lors du maintien de contact : *how are you* ?, *que tal* ?, *que pasa* ?, *Wat heb je allemaal gedaan* ?, *labes* ?, etc. Il semble que cet emploi assez régulier montre une fonction plus vaste que la simple politesse étant donné que ces FEs consomment souvent un nombre excessif de caractères (par rapport, notamment, à la variante francophone). **Nous pensons que dans la majorité des cas cet emploi revêt une fonction stylistique**.

4. Réitérer (répétition ou répétition légèrement modifiée) pour clarifier, amplifier, faire ressortir ou éclaircir ses propres idées :

(28) Doux amour de ma vie, tu vas bien ? Je suis à Bruxelles ! Je t'aime tout plein. Fais attention à toi pm. Tu me manques ! **I love you very very much** ! **Te quiero mi amor**. (smsFB)

Généralement, dans le contexte des sms, il s'agit d'amplifier un sentiment, souvent amoureux. Dans l'exemple 28, on remarque la présence du sentiment évoqué en langue française (*Je t'aime tout plein*), réitéré deux fois (*I love you* et *Te quiero*), pour insister sur la force et/ou la véracité dudit sentiment.

5. Exprimer de la solidarité :

39. Remarquons que la langue matrice du message en exemple peut être l'anglais.

40. Les éléments phatiques dans la conversation sont les éléments qui initient (par ex. *hola viejo amigo*) ou maintiennent (par ex. *you're welcome*) le contact entre le locuteur et l'interlocuteur (y compris d'un message à l'autre).

41. Dans Cougnon et François (2010), nous avions entrepris de répertorier la distribution des salutations les plus courantes pour la Belgique francophone. Le résultat de cette enquête, sous forme de tableau, se trouve en Annexe A.24.

(29) **Dag schatje**, ça fait longtemps qu'on cherchait l'occasion de vous revoir, et là, c'est tout trouvé ! Demain, se tient La Plus Grande Soirée Disco d'Europe au Sportpaleis d'Anvers (**volle vlaamse quiquine** ! ! !) ! Ça te dirait de venir avec NPERS et moi nous apprendre tes dernières techniques de chasse ? (smsFB)

On note que *volle vlaamse quiquine* se réfère de façon péjorative à des jeunes filles flamandes présentes lors de l'événement cité et cette attitude péjorative met donc en avant une solidarité masculine, supposée amusante, dont le but est de convaincre le destinataire à participer à l'événement.

6. Manifester son appartenance ethnique :

(30) **Salam alikoum** c'est NPERS comment va ? Ah oui pour les CD de NPERS tu les as ? Si oui dis-moi quand tu sauras me les donner et où **Beslama** réponds sur mon numéro. (smsFB)

L'exemple 30 illustre un cas d'utilisation de FE pour afficher une appartenance ethnique commune. De plus, là où le segment *Salam alikoum* montre une connaissance commune de l'arabe classique, koinè pour l'ensemble des pays arabophones, la deuxième FE, *beslama*, semble plus importante pour ce qui est du marquage d'identité, puisqu'elle met en lumière le partage d'une identité plus précise, marocaine

7. Marquer une stratégie hédoniste (exploiter les connotations affectives des variétés linguistiques utilisées) :

(31) C'est cool si ça été. Moi j'ai un peu foiré histoire. Demain j'ai néerlandais, j'espère que je vais réussir. Étudie bien anglais. T'as intérêt à réussir. **Big kiss**. **I miss you**. **I love you too**. **Good night**. Au 17 ou au 20 (smsFB)

L'exemple 31 démontre que chaque langue peut être utilisée, selon le locuteur, en fonction de la connotation que celle-ci a pour lui-même et/ou pour le destinataire. On remarque ici l'emploi de l'anglais pour tout ce qui a trait aux sentiments et du français pour ce qui concerne l'informatif et le pratique [42]. Cette connotation affective de la langue peut être partagée par plus d'un individu, par exemple par une communauté. C'est le cas de l'anglo-américain pour la communauté francophone [43] qui considère cette langue, au-delà du concept de langue « véhiculaire », comme porteuse d'un effet de style dit « cool » (cf. ex. 32 où *good* n'est pas utilisé pour abréger le message puisqu'il remplace l'adjectif français *bon* au nombre de caractères inférieur).

(32) Oh chérie ! Tu vas bien ? PUTAIN je suis DÉFONCÉE ! Grave ! Je pensais à toi tu sais bien ! Bref j'espère qu'il n'est pas fâché ! Passe un **good** week-end puce à dimanche bisou NPERS (smsFB)

42. Comme cela est expliqué juste après, la tendance est généralement inversée.
43. Et, en réalité, pour l'ensemble des autres communautés linguistiques (Crystal, 2008).

Les langues germaniques sont d'ailleurs généralement plus souvent associées à une fonction de style (« faire cool »,« être à la mode », « être dans le coup », etc.). En effet, l'allemand, le néerlandais et l'anglais sont rarement employés pour communiquer un sentiment qui concerne l'interlocuteur, mais bien plus pour améliorer l'image que le locuteur renvoie de lui-même (cf. ex. 33).

(33) Re. Ça va toujours ? Là je suis toujours au kot de ma **zus**. Vendredi jevais pieuter chez NPERS. Samedi c'est réunion en partie avec les pionniers. Et ta petite épaule je l'aime bien :-)(smsFB)

À l'inverse, les langues romanes semblent majoritairement employées pour leur connotation affective, en particulier les langues italienne et espagnole (cf. ex. 34).

(34) Même le ciel étoilé dessine ton visage, avec dans tes yeux, les étoiles les plus brillantes de la voie lactée :o) **bellissima Chepita**, **mi amor** ! Ma tendre...(smsFB)

8. Marquer la familiarité :

(35) Je suis bien décidée, je vais chez Wella, parce que elle avait bien compris, ça sera pour mardi. Merci **grote zuster** (smsFB)

9. Jouer avec la langue [44].

(36) OK ça roule ! **Met plezier fieu**. Prépare le **zooie** j'apporte la **water**. À plus (smsFB)

8.2.1.4 Conclusions sur les éléments allogènes d'après l'analyse qualitative

« Dans un contexte plurilingue les auteurs de sms usent de la totalité de leur répertoire linguistique, y compris les langues de prestige et les langues minoritaires. Souvent, on perçoit une séparation fonctionnelle des langues, avec des thématiques ou des contextes communicationnels régulièrement associés à des langues et des variétés de langue spécifiques. De tels schémas conventionnels sont pourtant flexibles et constamment en voie de mutation puisque la créativité linguistique de l'individu est un aspect central et très valorisé de ce nouveau genre » [45]. (Deumert et Vold Lexander, 2012)

44. Dans cette exemple, le jeu de mots porte sur les éléments *zooie* et *water* qui, combinés, forment le nom d'un plat typiquement belge (flamand)

45. *The analysis shows that in multilingual contexts texters make use of the totality of their linguistic repertoire, including majority as well as minority languages. Often we see a functional separation of languages with particular topics or communicative stances being associated regularly with specific languages or language varieties. Such conventionalized patterns, however, are fluid and constantly changing as individual linguistic creativity is a central and highly valorized aspect of this new genre.*

En effet, nous avons montré que le panorama des fonctions liées à l'emploi de langues étrangères dans les sms était vaste et que ces fonctions ne se limitaient pas à un besoin de faire court. Nous avons également conclu que les fonctions associées à une même langue étrangère n'étaient pas fixées même si des tendances globales se dessinaient. L'emploi d'une langue varie, y compris dans le discours d'un seul individu.

En conclusion, il semblerait que l'emploi et la fonction de chaque langue dans les sms ne suivent pas des schémas prédictibles et figés, contrairement à ce que l'on a pu conclure pour d'autres situations de communication plurilingue, et qu'ils soient majoritairement dépendants de 5 facteurs :

1. L'espace disponible dans un message de 160 caractères ;
2. Le rapport socio-/ethnico-psychologique établi entre le locuteur et le destinataire ;
3. Le thème et le style du discours ;
4. Les connotations associées à chaque langue, dans la communauté dont fait partie le locuteur ou pour le locuteur lui-même ;
5. La volonté du locuteur de créer un « effet » dans l'acte de communication.

8.2.2 Variation diastratique

L'objectif d'une approche quantitative est de révéler des phénomènes non reconnus à l'aide d'une analyse manuelle. Nous tentons de dégager des tendances qui permettent de mettre en avant des pratiques spécifiques à certaines sous-catégories de la population. Nous prenons le parti, dans un premier temps, d'analyser les variations en fonction de chaque région, puis nous regardons les tendances pour le corpus entier. Dans chaque partie, nous détaillons [46] la ou les variables étudiées (éléments allogènes, régionalismes, etc.) ; nous nous posons ensuite des questions sociolinguistiques auxquelles nous tentons de répondre avec l'aide des statistiques.
Nous regroupons, pour l'ensemble de l'analyse statistique dans toutes les régions, sous l'étiquette succincte *Emprunt* tous les éléments allogènes rencontrés. Cette fusion des tous les éléments allogènes n'est donc pas conceptuelle mais répond plutôt au besoin de faire court dans les tableaux, les figures et les explications statistiques.

8.2.2.1 Belgique

Description de la variable *Emprunt* : les tableaux présentés dans la Figure 8.10 révèlent deux informations importantes. Premièrement, **9,2 % des messages du corpus belge comportent un emprunt** (au moins). Ce chiffre (Valide : 1) parait important (notamment au regard des chiffres pour La Réunion et la Suisse). Deuxièmement, les langues les plus courantes sont **l'anglais, l'espagnol et l'italien**, le néerlandais arrivant en 4^e position seulement. Le détail de ce classement se trouve en Annexe A.25.

FIGURE 8.10 – *Effectifs et classement des emprunts dans le corpus belge*

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	24050	90,8	90,8	90,8
	1	2444	9,2	9,2	100,0
	Total	26494	100,0	100,0	

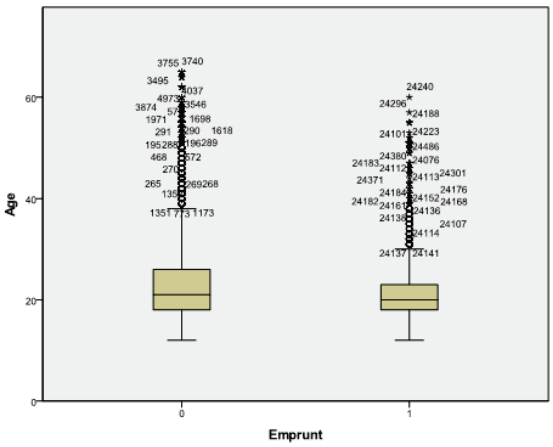
Langue	Position
Anglais	1
Espagnol	2
Italien	3
Néerlandais	4
Allemand	5
Arabe	6
Autres	7

1. Est-ce que le fait d'être plus ou moins âgé influence l'emploi d'un emprunt dans les sms ? Nous nous trouvons ici en présence d'une variable qualitative et

46. Les tests et hypothèses statistiques seront davantage détaillés dans cette partie sur les éléments allogènes. Pour les phénomènes étudiés suivants, afin de moins complexifier la lecture du texte, nous avons pris le parti de les résumer.

d'une variable quantitative. Il est donc nécessaire de regarder la distribution des données quantitatives [47].

FIGURE 8.11 – *Box-plot des emprunts du corpus belge en fonction de l'âge des auteurs*



Un rapide aperçu du box-plot présenté en Figure 8.11 permet de constater que les deux médianes [48] (pour Emprunt (0) : absent ; Emprunt (1) : présent) sont plus ou moins identiques. Un test doit donc être effectué pour confirmer cette hypothèse. Étant donné la distribution asymétrique des données *âge*, il y a peu de chances que les conclusions à partir de celles-ci soient fiables. Nous prenons donc la décision d'échantillonner la variable âge en ne prenant en compte que les sms des locuteurs ayant moins de 30 ans, et de tester notre hypothèse sur ceux-ci. Le box-plot de cet échantillon se trouve en A.26, p. 343. Cette nouvelle distribution étant plus *normale*, nous tentons donc un test t en définissant les hypothèses de départ :

- H1 : les médianes des emprunts en fonction des âges ne sont pas égales (M0 ≠ M1) ;
- H0 : les médianes des emprunts en fonction des âges sont égales (M0 = M1).

Le détail du test t se trouve en A.27, p. 344. On remarque que la valeur p est inférieure à 0,05, donc on rejette H0, les médianes des deux groupes ne sont pas égales, **il existe une influence de l'âge sur l'emploi des emprunts dans les sms** : la population usant d'éléments allogènes est plus restreinte que celle qui n'en utilise pas.

2. Est-ce que le fait d'être un homme ou une femme détermine l'usage des emprunts dans les sms ? Nous nous trouvons ici en présence de deux variables

47. Il est à noter pour l'ensemble des analyses statistiques que les profils sont démultipliés en fonction du nombre de messages que chaque auteur a envoyé et ne correspondent pas aux distributions présentées au chapitre 2.

48. La médiane est une mesure de dispersion qui calcule la moyenne des écarts par rapport à la moyenne. Cela permet de voir si tout l'effectif est proche de la moyenne ou s'il existe de grandes variations. Que les dispersions soient différentes pour deux modalités de la variable ou non, on prend la peine de chercher plus loin afin de confirmer l'observation du box-plot.

qualitatives. Nous tentons donc un test khi-carré (dont le détail se trouve en A.28, p. 344) en définissant les hypothèses de départ :
– H1 : il y a une dépendance entre les deux variables ;
– H0 : les deux variables sont indépendantes.
Le khi-carré lui-même nous donne une valeur p inférieure à 0,05, ce qui signifie que l'on rejette H0. Il y a donc dépendance entre les deux variables *emprunt* et *sexe*. L'analyse des résidus standardisés ajustés [49] (Figure 8.12) met en avant **un taux de femmes utilisant des emprunts dans les sms plus élevé (résidu = 3,8) que celui que l'on aurait pu attendre**. Au contraire, le taux d'hommes est moins élevé (résidu = -3,8) que celui que l'on aurait attendu.

FIGURE 8.12 – *Analyse des résidus standardisés relatifs au khi-carré (emprunts-sexe)*

Tableau croisé Sexe * Emprunt

			Emprunt		
			0	1	Total
Sexe	F	Effectif	14273	1546	15819
		Résidu ajusté	-3,8	3,8	
	M	Effectif	9777	897	10674
		Résidu ajusté	3,8	-3,8	
Total		Effectif	24050	2443	26493

3. Est-ce que le niveau d'études du locuteur [50] influe sur l'emploi ou non des emprunts dans les sms ? Nous nous trouvons ici en présence de deux variables qualitatives. Nous tentons donc un test khi-carré (dont le détail se trouve en A.29, p. 345) en définissant les hypothèses de départ (H1 et H0). Le khi-carré lui-même nous donne une valeur p inférieure à 0,05, ce qui signifie que l'on rejette H0. Il y a donc dépendance entre les deux variables *emprunt* et *études* en Belgique. L'analyse des résidus standardisés ajustés (Figure 8.13) montre que **le taux de diplômés du post-secondaire employant des emprunts dans les sms est largement inférieur à celui auquel on aurait pu s'attendre**. Pour les autres types de diplômés, le taux est supérieur.
4. Est-ce que le nombre de sms envoyés par semaine par un auteur et l'emploi ou non des emprunts dans les sms de cet auteur constituent deux variables dépendantes ? Le box-plot présenté en Figure 8.14 semble montrer une égalité des médianes entre emprunt et non emprunt en fonction du nombre de messages envoyés.

   En raison de la distribution acceptable (courbe normale) de la variable *Nombre de sms par semaine*, nous exécutons un test t (cf. Annexe A.30, p. 345) pour confirmer ou infirmer ce constat. La valeur p est égale à 0,877 ($> 0,05$) donc

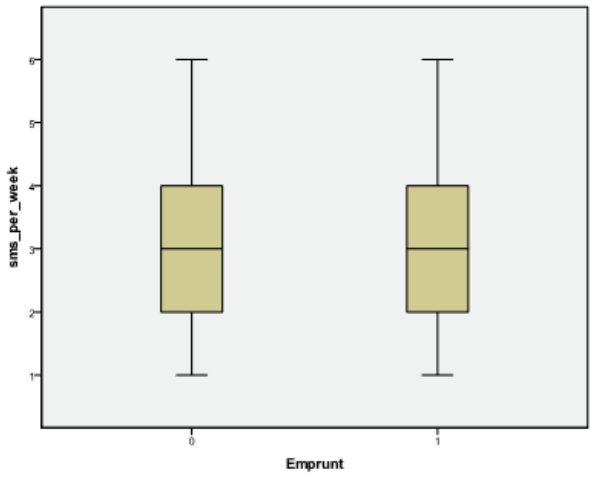
49. Le résidu standardisé ajusté est un indice statistique qui permet de mesurer la significativité de la dépendance entre deux modalités de réponse au sein d'un tableau de contingence.
50. Pour rappel, les niveaux d'études sont les suivants : Niveau (1) : primaire ; Niveau (2) : secondaire inférieur ; Niveau (3) : secondaire supérieur ; Niveau (4) : études post-secondaires.

FIGURE 8.13 – *Analyse des résidus standardisés relatifs au khi-carré (emprunts-études)*

Tableau croisé éducation * Emprunt

			Emprunt		
			0	1	Total
éducation	1	Effectif	296	43	339
		Résidu ajusté	-2,2	2,2	
	2	Effectif	2380	311	2691
		Résidu ajusté	-4,4	4,4	
	3	Effectif	7381	832	8213
		Résidu ajusté	-3,4	3,4	
	4	Effectif	13993	1257	15250
		Résidu ajusté	6,4	-6,4	
Total		Effectif	24050	2443	26493

FIGURE 8.14 – *Box-plot des emprunts du corpus belge en fonction du nombre de sms envoyés par semaine*



on ne rejette pas l'hypothèse nulle : les médianes des groupes sont égales. Dans le corpus belge, **on n'utilise pas plus d'emprunts lorsqu'on envoie beaucoup de sms par semaine et inversement**.

8.2.2.2 La Réunion

Description de la variable *Emprunts* : les tableaux présentés dans la Figure 8.15, p. 170, révèlent deux informations importantes. Premièrement, seuls **4,2 % des messages du corpus réunionnais comportent un emprunt** (au moins). Ce chiffre parait assez faible. Ceci peut sans doute être expliqué par la confusion des mots créoles avec les constructions néographiques propres à la Cémo, telle qu'elle a été décrite précédemment. Cette confusion n'a pas permis de retenir les emprunts en langue créole. Seules les autres langues étrangères sont donc prises en considération pour une région où le créole est pourtant omniprésent. La présence du créole ne doit

toutefois pas être sous-estimée dans le corpus, seule sa position est indéterminée et non définissable par nous-même. Deuxièmement, les langues les plus courantes pour ces emprunts sont **l'anglais, l'espagnol et l'allemand**. Le détail de ce classement se trouve en Annexe A.31, p. 346.

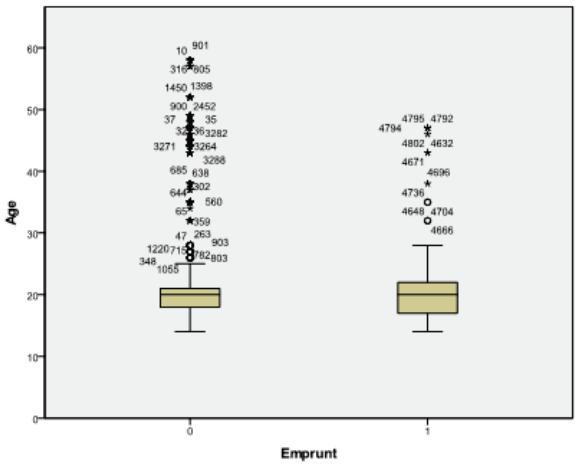
FIGURE 8.15 – *Effectifs et classement des emprunts dans le corpus réunionnais*

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	4607	95,8	95,8	95,8
	1	200	4,2	4,2	100,0
	Total	4807	100,0	100,0	
Manquante	Système manquant	1	,0		
Total		4808	100,0		

Langue	Position
Anglais	1
Espagnol	2
Allemand	3
Arabe	4
Néerlandais	5

1. Est-ce que le fait d'être plus ou moins âgé influence l'emploi d'un emprunt dans les sms ? Nous nous trouvons ici en présence d'une variable qualitative et d'une variable quantitative. Il est donc nécessaire de regarder la distribution des données quantitatives.

FIGURE 8.16 – *Box-plot des emprunts du corpus réunionnais en fonction de l'âge des auteurs*



Un rapide aperçu du box-plot présenté en Figure 8.16 met en avant deux phénomènes : premièrement, on aurait tendance à croire que les deux médianes sont plus ou moins identiques. Un test doit donc être effectué pour confirmer cette observation. Étant donné la distribution non normale des données *âge*, il y a peu de chances que les conclusions à partir de celles-ci soient fiables. Nous prenons donc la décision d'échantillonner la variable *âge* en ne prenant

en compte que les sms des locuteurs ayant moins de 30 ans (car les tranches d'âge supérieures ne sont pas assez représentées dans le corpus), et de tester notre hypothèse sur ceux-ci. Malheureusement la distribution obtenue après cet échantillonnage n'est pas plus acceptable : elle reste non normale. Nous recourons donc à un test U de Mann-Whitney, test non paramétrique, qui nous permettra de comparer les tendances centrales des deux groupes. Nous définissons les hypothèses de départ :

- H1 : les médianes des emprunts en fonction des âges ne sont pas égales (M0 ≠ M1) ;
- H0 : les médianes des emprunts en fonction des âges sont égales (M0 = M1).

Le détail du test U se trouve en Annexe A.32, p. 346. On remarque que p = 0,022 ; le test de Mann-Whitney nous incite donc à rejeter l'hypothèse nulle, à savoir H0 puisque les médianes des âges en fonction des emprunts ou non emprunts sont inégales. **Le corpus réunionnais montre une tendance à utiliser différemment les emprunts d'une classe d'âge à l'autre**, du moins pour ce qui est de la population de moins de 30 ans.

2. Est ce que les hommes et les femmes recourent de la même façon aux emprunts ? Nous nous trouvons ici en présence de deux variables qualitatives. Nous tentons donc un test khi-carré, dont le détail se trouve en A.33, p. 346, en définissant les hypothèses de départ :
   - H1 : il y a une dépendance entre les deux variables ;
   - H0 : les deux variables sont indépendantes.

   Le khi-carré nous donne une valeur p inférieure à 0,05 (0,029), ce qui signifie que l'on rejette H0. **Il y a donc dépendance entre les deux variables *emprunt* et *sexe***. L'analyse des résidus standardisés ajustés (Figure 8.17) met en avant un taux de femmes utilisant un emprunt plus élevé que celui que l'on aurait attendu. Au contraire, le taux d'hommes employant des emprunts dans le sms est moins élevé que celui que l'on aurait attendu.

FIGURE 8.17 – *Analyse des résidus standardisés relatifs au khi-carré (emprunts-sexe)*

Tableau croisé Sexe * Emprunt

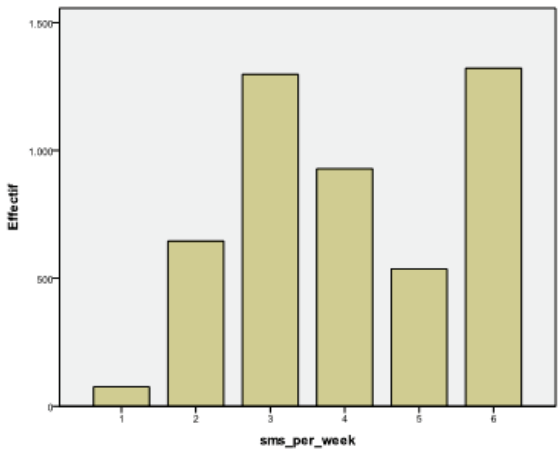
			Emprunt		
			0	1	Total
Sexe	F	Effectif	3263	153	3416
		Résidu ajusté	-2,2	2,2	
	M	Effectif	1344	43	1387
		Résidu ajusté	2,2	-2,2	
Total		Effectif	4607	196	4803

3. Est-ce que le niveau d'études du locuteur influe sur l'emploi ou non des emprunts dans les sms ? Nous nous trouvons ici en présence de deux variables qualitatives. Nous tentons donc un test khi-carré, dont le détail se trouve en A.34, p. 347, en définissant les hypothèses de départ (H1 et H0). Le khi-carré nous donne une valeur p supérieure à 0,05 (0,276), ce qui signifie que l'on ne

rejette pas H0. **Les variables *emprunt* et *études* sont donc indépendantes à La Réunion**.

4. Est-ce que le nombre de sms envoyés par semaine influe sur l'emploi ou non des emprunts dans les sms ? Le box-plot présenté en Figure 8.14, p. 169, semble montrer une égalité des médianes entre emprunt et non emprunt en fonction du nombre de messages envoyés. En raison de la distribution non normale, illustrée dans la Figure 8.18 de la variable *Nombre de sms par semaine*, nous exécutons un test U de Mann-Whitney (cf. Annexe A.35, p. 347) pour confirmer ou infirmer ce constat.

FIGURE 8.18 – *Distribution du nombre de sms envoyés par semaine (information donnée par les auteurs des messages composant le corpus réunionnais)*



Le test de Mann-Whitney nous incite donc à rejeter l'hypothèse nulle. On retient H1 : les médianes des emprunts en fonction des sms envoyés par semaine ne sont pas égales (M0 $\neq$ M1). **Dans le corpus réunionnais, on a tendance à utiliser plus ou moins d'emprunts selon le nombre de sms envoyés par semaine**.

8.2.2.3 Québec

1. Description de la variable *Emprunts* : les tableaux présentés dans la Figure 8.19, p. 173, révèlent deux informations importantes. Premièrement, **17,7 % des messages du corpus québécois comportent un emprunt** (au moins). Comme on pouvait le prévoir pour une région en contact intense avec l'anglais, ce chiffre est élevé. Deuxièmement, les langues les plus courantes sont **l'anglais, l'espagnol et l'italien** ; l'anglais et l'espagnol forment 99 % des emprunts dans le corpus. Le détail de ce classement se trouve en Annexe A.36, p. 348.
2. Est-ce que le fait d'être plus ou moins âgé influence l'emploi d'un emprunt dans les sms ? Nous nous trouvons ici en présence d'une variable qualitative et

FIGURE 8.19 – *Effectifs et classement des emprunts dans le corpus québécois*

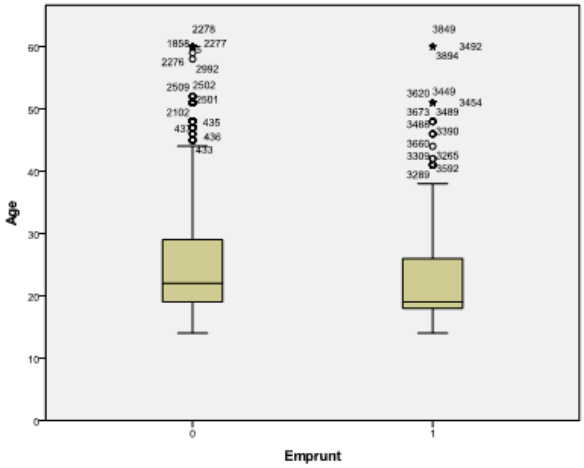
Emprunt

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	3210	82,3	82,3	82,3
	1	692	17,7	17,7	100,0
	Total	3902	100,0	100,0	

Langue	Position
Anglais	1
Espagnol	2
Italien	3
Allemand	4
Néerlandais	5

d'une variable quantitative. Il est donc nécessaire de regarder la distribution des données quantitatives.

FIGURE 8.20 – *Box-plot des emprunts du corpus québécois en fonction de l'âge des auteurs*



Le box-plot présenté dans la Figure 8.20 met en avant deux phénomènes : premièrement, on note que les deux médianes sont sensiblement différentes. Un test doit donc être effectué pour confirmer cette hypothèse. Deuxièmement, étant donné la distribution non normale des données âge, il y a peu de chances que les conclusions à partir de celles-ci soient fiables. Nous prenons donc la décision, comme ailleurs, d'échantillonner la variable âge en ne prenant en compte que les sms des locuteurs ayant moins de 30 ans, et de tester notre hypothèse sur ceux-ci, comme nous l'avons fait pour la Belgique. Malheureusement la distribution obtenue n'est pas plus acceptable (non normale). Nous recourons donc à un test U de Mann-Whitney, en définissant les hypothèses de départ comme exécuté précédemment (H0 et H1). Le détail du test U se trouve

en Annexe A.37, p. 348. On remarque que la valeur p est inférieure à 0,05, donc on rejette H0, les médianes des emprunts en fonction de l'âge ne sont pas égales M0 ≠ M1. **On a tendance à utiliser différemment les emprunts selon l'âge, au Québec**, en tout cas en ce qui concerne la population de moins de 30 ans.

3. Est-ce que le fait d'être un homme ou une femme fait varier l'emploi des emprunts dans les sms ? Nous nous trouvons ici en présence de deux variables qualitatives. Nous tentons donc un test khi-carré (dont le détail se trouve en A.38, p. 348) en définissant les hypothèses de départ :
   – H1 : il y a une dépendance entre les deux variables ;
   – H0 : les deux variables sont indépendantes.
   Le khi-carré nous donne un résultat inférieur à 0,05, ce qui signifie que l'on rejette H0. **Il y a donc dépendance entre les deux variables *emprunt* et *sexe* au Québec**. L'analyse des résidus standardisés ajustés (Figure 8.21) met en avant un taux de femmes utilisant un emprunt légèrement plus élevé que celui que l'on aurait attendu. Au contraire, le taux d'hommes employant des emprunts dans le sms est légèrement moins élevé que celui que l'on aurait attendu.

FIGURE 8.21 – *Analyse des résidus standardisés relatifs au khi-carré (emprunts-sexe)*

Tableau croisé Sexe * Emprunt

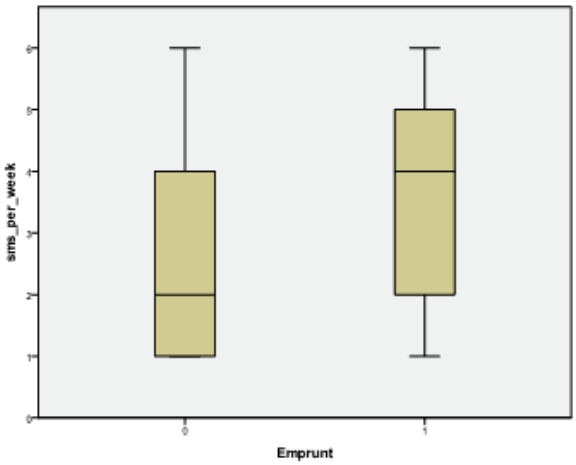
Résidu ajusté

		Emprunt	
		0	1
Sexe	F	-2,3	2,3
	M	2,3	-2,3

4. Est-ce que le niveau d'étude du locuteur influe sur l'emploi ou non des emprunts dans les sms ? Nous nous trouvons ici en présence de deux variables qualitatives. Nous tentons donc un test khi-carré (dont le détail se trouve en A.39, p. 349) en définissant les hypothèses de départ (H1 et H0). Le khi-carré offre un résultat supérieur à 0,05, ce qui signifie que l'on ne rejette pas H0. **Il y a donc indépendance entre les deux variables *emprunt* et *études* au Québec**.

5. Est-ce que le nombre de sms envoyés par semaine influe sur l'emploi ou non des emprunts dans les sms ? Le box-plot présenté en Figure 8.22 semble montrer une inégalité des médianes entre emprunt et non emprunt en fonction du nombre de messages envoyés.
   En raison de la distribution non normale de la variable *Nombre de sms par semaine* (A.40, p.349), nous exécutons un test U de Mann-Whitney (cf. A.41, p. 349) pour confirmer ou infirmer ce constat. La valeur p est inférieure à 0,001 donc on rejette l'hypothèse nulle : les médianes des groupes ne sont pas

FIGURE 8.22 – *Box-plot des emprunts du corpus québécois en fonction du nombre de sms envoyés par semaine*



égales. Dans le corpus québécois, **on utilise plus ou moins d'emprunts en fonction du nombre de sms que l'on envoie par semaine**.

8.2.2.4 Suisse

1. Description de la variable *Emprunts* : les tableaux présentés dans la Figure 8.23 révèlent deux informations importantes. Premièrement, seuls **3,6 % des messages du corpus suisse comportent un emprunt** (au moins). Ce chiffre est faible et semble corroborer les conclusions énoncées par Doehler (2011) pour un corpus plus restreint [51]. Deuxièmement, les langues les plus courantes pour ces emprunts sont **l'anglais, l'espagnol et l'italien** ; la place de l'italien n'est pas surprenante étant donné la situation linguistique du pays. Étonnamment, par contre, l'allemand n'apparait qu'en 4e position (le détail de ce classement se trouve en A.42, p. 350) [52].

2. Est-ce que le fait d'être plus ou moins âgé influence l'emploi d'un emprunt dans les sms ? Nous nous trouvons ici en présence d'une variable qualitative et d'une variable quantitative. Il est donc nécessaire de regarder la distribution des données quantitatives.

   Un rapide aperçu du box-plot présenté en Figure 8.24 met en avant deux phénomènes : premièrement, on aurait tendance à croire que les deux médianes sont plus ou moins identiques. Un test doit donc être effectué pour confirmer cette observation. Etant donné la distribution non normale des données âge, il y a peu de chances que les conclusions à partir de celles-ci soient fiables.

51. « La Suisse, bien qu'officiellement plurilingue, est régie par le principe de territorialité : les régions sont - à peu d'exceptions près - majoritairement unilingues. [...] Ce fait se reflète largement dans les messages recueillis » (Doehler, 2011).

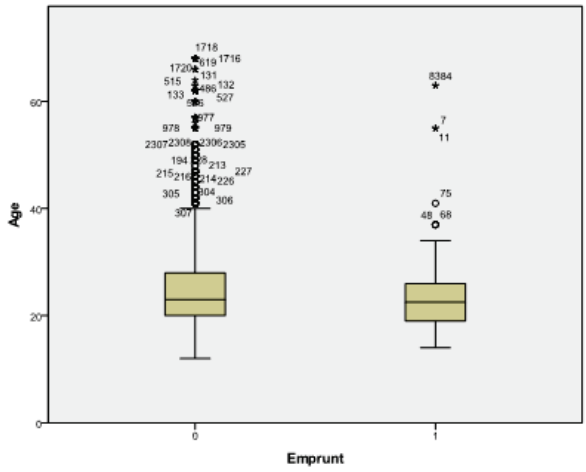
52. Notons que l'allemand et le suisse allemand ont été distingués lors du traitement des FE.

FIGURE 8.23 – *Effectifs et classement des emprunts dans le corpus suisse*

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	3412	96,4	96,4	96,4
	1	129	3,6	3,6	100,0
	Total	3541	100,0	100,0	

Langue	Position
Anglais	1
Espagnol	2
Italien	3
Allemand	4

FIGURE 8.24 – *Box-plot des emprunts du corpus suisse en fonction de l'âge des auteurs*



Nous prenons donc la décision d'échantillonner la variable âge en prenant uniquement en compte les sms des locuteurs ayant moins de 30 ans, et de tester notre hypothèse sur ceux-ci. Malheureusement la distribution obtenue après cet échantillonnage n'est pas plus acceptable (non normale). Nous recourons donc à un test U de Mann-Whitney. Nous définissons les hypothèses de départ (H0 et H1) comme réalisé précédemment. Le détail du test U se trouve en Annexe A.43, p. 350. On remarque que p = 0,741 ; le test de Mann-Whitney nous incite donc à ne pas rejeter l'hypothèse nulle, à savoir H0 (les médianes des emprunts en fonction des âges sont égales M0 = M1). **Le corpus suisse ne montre aucune tendance à utiliser plus ou moins d'emprunts en fonction de l'âge de l'auteur du message**, en tout cas pour la population des moins de 30 ans.

3. Est-ce que le fait d'être un homme ou une femme change l'usage des emprunts dans les sms ? Nous nous trouvons ici en présence de deux variables qualitatives. Nous tentons donc un test khi-carré, dont le détail se trouve en A.44,

p. 350, en définissant les hypothèses de départ (H1 et H0). Le khi-carré révèle un résultat supérieur à 0,05 (0,152), ce qui signifie que l'on ne rejette pas H0. **Il y a donc indépendance entre les deux variables *emprunt* et *sexe* en Suisse**.

4. Est-ce que le niveau d'étude du locuteur influe sur l'emploi ou non des emprunts dans les sms ? Nous nous trouvons ici en présence de deux variables qualitatives. Nous tentons donc un test khi-carré, dont le détail se trouve en A.45, p. 351, en définissant les hypothèses de départ (H1 et H0). Le khi-carré montre un résultat inférieur à 0,05 (0,015), ce qui signifie que l'on rejette l'hypothèse nulle. **Dans le corpus suisse, l'emploi des emprunts est donc lié au niveau d'étude**, directement ou indirectement car il serait plutôt lié à l'âge (les variables niveau d'études et âge étant elles-mêmes liées). L'analyse des résidus standardisés ajustés (Figure 8.25) met en avant un taux de diplômés du post-secondaire utilisant un emprunt légèrement moins élevé que celui que l'on aurait attendu. Au contraire, le taux de diplômés de secondaire supérieur employant des emprunts dans le sms est légèrement plus élevé que celui que l'on aurait attendu.

FIGURE 8.25 – *Analyse des résidus standardisés relatifs au khi-carré (emprunts-études)*

Tableau croisé éducation * Emprunt

			Emprunt		
			0	1	Total
éducation	0	Effectif	135	8	143
		Résidu ajusté	-1,3	1,3	
	2	Effectif	265	7	272
		Résidu ajusté	1,0	-1,0	
	3	Effectif	942	50	992
		Résidu ajusté	-2,8	2,8	
	4	Effectif	2070	64	2134
		Résidu ajusté	2,5	-2,5	
Total		Effectif	3412	129	3541

5. Est-ce que le nombre de sms envoyés par semaine influe sur l'emploi ou non des emprunts dans les sms ? En raison de la distribution non normale du nombre de messages envoyés par semaine en Suisse, nous exécutons un test U de Mann-Whitney (cf. Annexe A.46, p. 351) pour étudier les médianes. Le résultat du test de Mann-Whitney ($p = 0{,}044$) nous incite à ne pas rejeter l'hypothèse nulle. Les médianes des emprunts et non emprunts en fonction des sms envoyés par semaine sont égales ($M0 \neq M1$). **Dans le corpus suisse, on a tendance à utiliser de la même façon les emprunts, quel que soit le nombre de sms envoyés par semaine**.

Le tableau 8.26 récapitule les liens entre variables et les comparaisons de médianes dans chaque région.

FIGURE 8.26 – *Hypothèses liant (H1) ou non (H0) les variables et les médianes de variable dans chaque région*

	Belgique	La Réunion	Québec	Suisse	Global
Question 1 âge	H1	H1	H1	H0	H1
Question 2 sexe	H1	H1	H1	H0	H1
Question 3 niveau d'étude	H1	H0	H0	H1	H1
Question 4 Nombre de sms envoyés	H0	H1	H1	H0	H0

Le choix de réaliser des études distinctes par région n'est pas arbitraire. Nous ne voulions pas donner une idée globale mais bien un aperçu des pratiques de chaque région/période. Le tableau récapitulatif montre que les résultats sont plus ou moins variables d'une région à l'autre. Les hypothèses de la Suisse sont par exemple totalement opposées à celles du Québec. Il semble donc impossible de dresser des lignes de conduite des auteurs des messages. Notre corpus n'est-il pas assez représentatif ? Les corpus sont-ils trop disproportionnés dans leurs tailles ? Lorsque nous étudions les différents corpus comme un tout (cf. Annexes A.47 à A.53), nous concluons que l'âge, le sexe et le niveau d'études influent sur l'utilisation ou non d'un emprunt. Que déduire de ces constatations ? Nous décidons de ne pas généraliser nos conclusions à l'ensemble du corpus de sms mais de prendre en considération les conclusions de chaque corpus, par région, en gardant à l'esprit que **l'âge, le niveau d'études et le sexe des auteurs des messages du corpus sms4science semblent avoir un effet sur l'utilisation ou non d'un emprunt**, dans la majorité des régions.

8.2.2.5 Corpus global : langue de l'élément allogène

Le regroupement de l'ensemble des corpus offre des conclusions sur l'ensemble des données et permet certains calculs impossibles pour chaque région étant donné le non respect de certaines conditions pour la validité des tests statistiques, en particulier posséder un nombre suffisant de données par catégorie/variable. Le sujet que nous avons décidé d'aborder ici est **la langue de l'emprunt**. La Figure 8.27 propose le classement moyen des langues des emprunts pour l'ensemble des régions.
Le corpus de sms contenant un ou plusieurs emprunts est constitué de 3.462 messages (pour 986 auteurs). Le Tableau 8.28 présente le classement des 20 mots simples les plus empruntés dans le corpus sms4science francophone. Ce tableau met en avant **la prédominance de l'anglais dans les emprunts en contexte de sms francophones**. Nous avons vu dans la partie qualitative qu'à chaque langue était associée une fonction (émotive, pragmatique, etc.). Nous allons maintenant voir si à chaque langue sont associées des sous-catégories de la population (des variables de profil).

1. Est-ce que la variable *âge* influe sur le choix de la langue d'emprunt ? Nous nous trouvons ici en présence d'une variable qualitative et d'une variable quantitative. La variable qualitative montrant plus de deux réponses possibles, c'est

FIGURE 8.27 – *Classement des langues d'emprunt pour le corpus global sms4science francophone*

Langue	Position
Anglais	1
Espagnol	2
Italien	3
Allemand	4
Néerlandais	5
Arabe, portugais, grec, catalan, turc, etc.	6

FIGURE 8.28 – *Classement des emprunts les plus fréquents pour le corpus global sms4science francophone*

Position	Occurrences	Forme
1	547	kiss
2	250	sorry
3	201	news
4	173	today
5	136	big
6	120	now
7	61	you
8	60	new
9	52	my
10	47	good
11	45	thanks
12	38	the
13	36	please
14	31	study
15	27	call
16	23	happy
17	22	problem
18	21	best
19	21	day
20	21	nice

le test Anova qui est envisagé (Tableau 8.29).

On remarque que p > 0,05 (0,252) ; le test Anova (détails en A.54, p. 355) nous incite donc à ne pas rejeter l'hypothèse nulle, à savoir H0 (les médianes des langues empruntées en fonction des âges sont égales : M0 = M1). **Le corpus global ne montre aucune tendance à utiliser plus ou moins une langue d'emprunt en fonction de l'âge des auteurs de sms**.

2. Est-ce que la variable *sexe* influe sur le choix de la langue d'emprunt ? Nous nous trouvons ici en présence de deux variables qualitatives. Nous tentons donc

FIGURE 8.29 – *Test Anova pour la comparaison des médianes (variable qualitative à plus de deux modalités)*

ANOVA

	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	,303	3	,101	,188	,905
Intra-groupes	1658,222	3083	,538		
Total	1658,525	3086			

un test khi-carré, dont le détail se trouve en A.55, p. 355, en définissant les hypothèses de départ (H1 et H0). Le khi-carré révèle un résultat inférieur à 0,05 (0,003), ce qui signifie que l'on rejette H0. Les variables *langue de l'emprunt* et *sexe* sont donc dépendantes. Le corpus de sms francophones montre donc une tendance à utiliser davantage une langue qu'une autre en fonction du sexe de l'auteur du message. L'analyse des résidus standardisés ajustés met en avant **un taux d'hommes utilisant les emprunts allemands et *divers* (position 6) plus élevé que celui que l'on aurait attendu**. Au contraire, le taux de participantes employant des emprunts allemands et *divers* est moins élevé que celui que l'on aurait attendu.

3. Est-ce que la variable *niveau d'études* influe sur le choix de la langue d'emprunt ? Nous allons ici effectuer deux tests différents en fonction de deux questions de recherche qui nous intéressent particulièrement. Premièrement, est-ce que le fait d'avoir fait des études supérieures change l'emploi des langues d'emprunt ? Pour le savoir, nous regroupons d'abord les données en deux catégories[53]. Le test khi-carré (cf. A.56, p. 356) effectué révèle que $p > 0,05$ (0,218) ; l'hypothèse nulle est donc maintenue. Dans le corpus global, **il ne semble pas y avoir de lien entre études supérieures et langue d'emprunt**. Deuxièmement, il nous intéresse de savoir si le fait d'avoir fait des études supérieures influe sur l'emploi de l'anglais dans les emprunts. Pour ce faire, nous regroupons les observations en deux catégories[54]. Le test khi-carré (Annexe A.57, p. 356) prouve l'indépendance des deux variables ($p = 0,061$). **Le fait d'avoir accompli des études supérieures ne semble pas influer sur l'emploi de la langue anglaise** dans les emprunts de notre corpus.
4. Est-ce que la variable *langue première* influe sur le choix de la langue d'emprunt ? Cette question parait essentielle car il semble *a priori* évident qu'un locuteur aura majoritairement tendance à utiliser spontanément au quotidien une langue qu'il maitrise très bien. En raison du type de données présentes (deux variables qualitatives), nous réalisons un test khi-carré (Annexe A.58, p. 356). L'analyse des données de la présente recherche montre qu'il n'y a pas de dépendance ($sig = 0,722$) entre les variables langue de l'emprunt (anglaise ou autre) et langue première (française ou autre). La présentation des don-

53. La catégorie 1 regroupe les observations en deçà des études supérieures ; la catégorie 2 correspond aux études supérieures.

54. La catégorie 1 correspond à l'anglais alors que la catégorie 2 renvoie aux autres langues.

nées a toutefois mis en avant un phénomène intéressant : si la langue première d'un individu ne semble pas conditionner le choix de la langue d'emprunt, elle parait influer sur le fait même d'user d'un emprunt ou non. En effet, un test khi carré sur les variables *emprunt ou non emprunt* et *langue première* (Annexe A.59, p. 357) prouve qu'**il existe une dépendance entre la variable langue première et choix de la langue d'emprunt**. Au regard des résidus standardisés ajustés, **on peut conclure, de façon très étonnante, que le taux de locuteurs de langue première française usant d'emprunts est bien plus élevé que celui que l'on aurait attendu**, alors que le taux des locuteurs avec une langue première autre est très en deçà de ce que l'on aurait attendu (Annexe A.60). On pourra donc se demander si le contexte plurilingue d'un individu est moins stimulant pour l'emploi d'un emprunt que ne l'est la maitrise de la langue matrice du message (dans notre cas, le français) ; on pourrait encore émettre l'hypothèse selon laquelle le contexte plurilingue d'un locuteur engendrerait chez lui une tendance à l'hypercorrection (l'attitude standard étant alors d'employer des systèmes linguistiques différents dans les sms).

8.2.3 Conclusions

Nous avons vu dans la partie qualitative qu'il existait une importante variété de fonctions liées à l'emploi d'éléments allogènes dans les sms. Nous avons également pu constater que les fonctions associées à une langue étrangère en particulier n'étaient pas fixées même si des tendances globales se dessinaient : l'anglais, pour faire court ou pour créer un effet de style « cool » et les langues romanes (italien et espagnol) pour transmettre une forme d'affectivité. Nous avons ensuite utilisé l'analyse quantitative pour discerner et affiner ces observations.

Ainsi, nous avons mis en avant une tendance générale à utiliser différemment les éléments allogènes en fonction de l'âge, du sexe et du niveau d'études du locuteur, mais non en fonction du nombre de sms qu'il envoie par semaine. Nous avons également constaté, via l'analyse quantitative du corpus global, que, pour notre corpus, l'usage ou non d'un élément allogène est fonction de la langue première du locuteur et que l'emploi d'une langue étrangère spécifique est fonction du sexe du locuteur. Enfin, nous observons la nette prédominance quantitative de l'anglais sur les autres langues et ce, malgré les langues en présence dans chaque région étudiée (le néerlandais en Belgique, l'allemand et l'italien en Suisse).

En conclusion, il semblerait donc que l'emploi de chaque langue dans les sms soit majoritairement dépendant de 7 facteurs :

1. L'espace disponible dans un message de 160 caractères ;
2. Le rapport socio-/ethnico-psychologique entre le locuteur et le destinataire ;
3. Le thème et le style du discours ;
4. Les connotations associées à chaque langue, dans la communauté dont fait partie le locuteur ou pour le locuteur lui-même ;
5. La volonté de créer un « effet » ;
6. Le profil du locuteur (âge, sexe, niveau d'études et langue première) ;
7. La région dans laquelle réside le locuteur.

Doehler (2011, p. 65) a conclu que les emplois d'éléments allogènes dans les sms « ne relèvent pas d'opérations aléatoires, mais de choix qui sont instrumentaux dans la structuration de la dimension socio-interactive de la communication par SMS ». Conséquemment, rappelons que, même s'ils ne sont pas aléatoires, ces choix ne sont pas figés, et qu'ils varient notamment et surtout en fonction des 7 facteurs que nous venons d'énumérer.

8.3 Régionalismes lexicaux

Nous entamons dans cette partie une étude du lexique régional des variétés du français hors de France (métropolitaine) dans des terres où le français a une assise importante et se trouve en présence d'autres systèmes linguistiques (officiels ou récurrents dans la pratique de la population) : la Belgique, La Réunion, le Québec et la Suisse. L'étude scientifique des variétés géographiques (ou topolectales) du français hors de France est relativement récente, surtout dans une perspective descriptive (Francard et Latin, 1995 ; Klein et Rossari, 2003 ; Pöll, 2001). On estime en général que les premières conclusions objectives au sujet des régionalismes datent des années 70 avec Davau *et al.* (1972) ou Valdman (1979). Le français régional est à différencier du dialecte. En notre sens, le français régional est une manifestation délimitée (géographiquement, culturellement, etc.) d'une koinè, une langue commune, pratiquement virtuelle. Le dialecte, au contraire, n'est pas une variété de langue (quelque virtuelle qu'elle soit), c'est un système linguistique totalement distinct : seules le séparent de l'étiquette de langue ses diffusions culturelle, sociale et médiatique. Peu de recherches sur les variétés du français sont fondées sur l'écrit spontané des populations. En cela, la Cémo offre un objet d'étude particulièrement intéressant puisqu'il permet de revisiter des résultats et théories quasi exclusivement construites sur l'oral et sa transcription.

Étant donné l'absence d'accord entre les différentes disciplines liées à la linguistique sur la terminologie dans le domaine du régionalisme lexical[55], nous décidons de poser notre propre définition : une définition vaste et qui fait sens dans le cadre de notre étude. On s'entendra donc pour appeler **régionalisme lexical** les unités du lexique d'une aire francophone qui sont spécifiques à cette aire ou communes à toutes les aires excepté (au moins) une. On inclura donc aussi dans cet ensemble les *spécificités* de chaque région, c'est-à-dire les unités lexicales dont l'emploi est plus courant, plus répandu, que dans le reste de la francophonie.

Le monde scientifique a, jusqu'ici, très peu traité de la question des variétés de langue dans les sms. Si l'on met de côté Crystal (2008) qui cite le phénomène sans l'approfondir, et Fairon *et al.* (2006b) qui traitent de la question en se basant sur le corpus belge et étudiant donc uniquement la variété française de Belgique, la littérature est pauvre à ce sujet. Les travaux s'intéressant aux régionalismes concernent surtout le chat (Siebenhaar, 2005 ; Ziegler, 2005) ou la communication électronique en général (Eisenlohr, 2004) et ne visent pas la systématicité.

Nous avons nous-même investi, la voie des régionalismes dans les sms en tentant

55. Notamment quant au fait d'inclure ou non dans le référent « région » les régions intra-hexagonales, ou même, ce qui nous semble encore plus écarté de l'objectivité (qui permet d'accorder à chaque variété le même statut), quant à l'acception ou le refus de la notion de « français central » ou de « français de référence ». On citera par exemple la définition proposée par Boulanger (1985, p. 137).

d'une part une systématicité pour le corpus belge[56] et, d'autre part, en 2010, en nous associant pour envisager l'étude du français en Belgique et à La Réunion (Cougnon et Ledegen, 2010). Ces études sur la variation des français ne recouvrent donc pas l'intégralité du corpus francophone de sms, un objectif que nous visons ici.

Pour comprendre le profil des régionalismes lexicaux dans chaque région, un panorama de la situation linguistique est nécessaire. Or, nous ne pensons pas qu'un état détaillé de la question linguistique synchronique et diachronique dans chaque région étudiée soit nécessaire puisqu'elle a déjà fait l'objet de nombreuses recherches plus particulièrement axées sur ce sujet. C'est pourquoi, nous avons décidé de résumer, avant chaque typologie, la situation des quatre régions en fonction des traits importants pour notre analyse, sans aller plus avant dans les détails.

La méthodologie d'analyse des régionalismes a différé d'une région à l'autre en fonction de la disponibilité des données. Celle-ci sera donc également décrite dans chaque partie selon les spécificités de la région. Certains traits de cette méthodologie sont pourtant communs. Premièrement, une fois la liste des unités régionales constituée, nous avons employé l'outil d'exploitation de corpus Unitex pour identifier les entités concernées. La syntaxe de reconnaissance des unités, élaborée par nous-même, a la spécificité de permettre (comme cela a été le cas pour les néologismes) l'identification de formes « fléchies » (conjugaison des verbes et déclinaison des noms et adjectifs) des unités appartenant au dictionnaire de référence (par exemple, les verbes supports de certaines constructions : « avoir bon »), mais non de celles n'en faisant par partie (la majorité des unités régionales) qui doivent être fléchies manuellement. Nous n'avons pas réalisé ce travail manuel de flexion, complexe et couteux en temps, ce qui entraine un biais de non-reconnaissance d'un certain nombre de formes dans le corpus. Deuxièmement, quelques unités graphiques sont polysémiques ou se construisent de multiples façons et engendrent donc des ambiguïtés lors de la reconnaissance par Unitex. Ainsi, la spécificité de Belgique *dis*, interjection ou marqueur de discours selon le cas, se confond avec la forme verbale fléchie traditionnelle. Nous avons filtré manuellement l'intégralité des sms concernés afin de séparer les formes régionales du bruit engendré par les homographes.

Nous tenterons, dans cette partie, d'offrir des réponses qualitatives et quantitatives aux questions suivantes :

1. Quelle vitalité des régionalismes lexicaux dans les sms peut-on observer ?
2. Quels sont les contextes de leur emploi dans les messages ?
3. Quels types d'unités sont concernés ?

56. Pour information, l'article est le suivant : Louise-Amélie Cougnon (2008). « Le français de Belgique dans 'l'écrit spontané'. Approche d'un corpus de 30.000 SMS ». Dans *Travaux du Cercle Belge de Linguistique*. En ligne.

4. Observe-t-on des différences de profils sociodémographiques entre les usagers qui emploient des régionalismes et ceux qui n'en emploient pas ?

8.3.1 Régionalismes de Belgique

La Belgique possède trois langues nationales : le français, le néerlandais et l'allemand. Le français est parlé en Wallonie (à côté de l'allemand) et dans la Région de Bruxelles-Capitale, à côté du néerlandais. Le néerlandais est également parlé en Flandre. En Belgique, on assiste donc à la rencontre des cultures germanique et romane qui marque également l'histoire linguistique du pays : une origine latine (par la romanisation), un mélange de substrat préceltique et celtique (que l'on retrouve entre autres choses, dans la toponymie belge), et d'adstrat germanique (Blampain, 1997). Notons que le français de Belgique est influencé par le liégeois, le namurois, le wallo-picard et le wallo-lorrain (Pöll, 2001). L'adoption de la langue française comme langue de culture date du XIII^{e} siècle et a donné lieu à un bilinguisme franco-dialectal pendant plusieurs siècles. En 1830, le français est promu seule langue officielle du royaume (le flamand étant considéré comme une langue « nationale », pratiquée, mais non officielle). Cette situation durera jusqu'à la décennie 1873-1883 où une série de lois en matière linguistique redresseront l'équilibre linguistique, notamment pour ce qui a trait à la justice, à l'administration et à l'enseignement.[57]. Cependant, les contacts avec les parlers flamands, tout comme ceux avec les dialectes belgo-romans, ont marqué durablement le français de Belgique, notamment via bon nombre de régionalismes lexicaux (Piron, 1979).

Il existe une certaine « insécurité linguistique » (Labov, 1966) historique en Belgique (notamment en raison des « Chasses aux Belgicismes » de Hanse *et al.* (1971, 1974) durant les années 1970). Cette insécurité tend à disparaitre de nos jours, surtout auprès des générations les plus jeunes pour qui le modèle français est devenu moins important que le modèle américain. Ce phénomène peut, selon nous, s'expliquer par deux facteurs : d'une part, la mondialisation et la diffusion massive des nouveaux médias de communication ; d'autre part, l'attitude des scientifiques eux-mêmes envers le descriptif, non qu'ils initient aujourd'hui les recherches descriptives[58], mais plutôt que les études normatives et leurs conclusions sont aujourd'hui moins prégnantes à l'université comme dans les médias.

Deux sources principales sont à la base de l'étude des régionalismes lexicaux de Belgique : les unités précédemment répertoriées dans Fairon *et al.* (2006b), complétées lors de la recherche comparative de Cougnon et Ledegen (2010) par le dictionnaire Lebouc (2006), que nous avons d'abord numérisé[59].

57. La supériorité culturelle du français à l'époque s'explique par sa place privilégiée au sein des élites flamandes.

58. De telles études existent de longue date, notamment depuis Marcel Paquot, initiateur des premiers cours sur le français régional de Belgique à l'Université de Liège, et son successeur, Maurice Piron, notamment avec Piron (1955).

59. Un grand merci à Mario Marcon pour ce travail de numérisation des entrées lexicales. Le

Nous dénombrons 5.079 sms concernés par les régionalismes dans le corpus belge, soit un sixième de ce corpus. Ces sms comprennent 188 régionalismes distincts, dont le détail se trouve en Annexe A.61.

8.3.1.1 Approche qualitative

Un rapide aperçu des 188 belgicismes permet de constater que les régionalismes de notre corpus concernent des mots simples, des mots composés, des constructions à verbe support ainsi que des expressions.

Au sein des unités simples, nous dénombrons des verbes (*savoir*, *souper*, *diner*, *péter*, *bloquer*, etc.). Le célèbre *savoir*, que nous approfondissons un peu plus loin puisqu'il s'agit du belgicisme le plus employé dans les sms, se construit avec un infinitif comme complément (exemple 37) ; des adverbes et des pronoms s'insérent optionnellement entre les deux.

(37) Bébé je **sais** seulement te répondre maintenant pourquoi tu voulais que je te téléphone avant-hier je **sais** pas te téléphoner j'ai plus de tune Réponds sur mon GSM bébé tu **sais** me téléphoner tantôt quand je sors de l'école à 2h20 je t'aime (smsFB)

Parmi ces verbes, certains opèrent une restriction (*chatouiller*[60], exemple 39) ou extension (*clôturer*, exemple 38 ou *baptiser*) de sens, d'autres une mutation de sens (*bloquer*, au sens de « étudier » – exemple 40 – ou *brosser*).

(38) Merci pour les coordonnées du cours d'allemand malheureusement les inscriptions sont **clôturées** à bientôt (smsFB)

(39) Au moins t'es direct ! C'est déjà ça ! Lol. Non c'est fini depuis plus ou moins 2 semaines. Et au cas où le quizz te **chatouillerait**, c'est moi et j'ai pas de regrets...mdr. Simple quizz : cheveux plus ou moins bouclés ? (smsFB)

(40) Désolé bébé mais ça va pas être possible car maintenant je suis au tennis pendant 3h et puis je dois **bloquer**. J'ai beaucoup à . Désolée, on se verra encore sûrement dans pas longtemps, big bisou dans le cou (smsFB)

Une catégorie supplémentaire est, elle, constituée de formes nouvelles qu'on ne retrouve pas dans un dictionnaire de référence (*guindailler*, *kotter* - exemple 41 -, *prester*, *resonner*).

choix de ce dictionnaire peut être critiqué ; cependant, le choix de celui-ci résulte de sa plus grande accessibilité au moment de la recherche. Depuis est paru le dictionnaire de Francard *et al.* (2010), qui pourrait constituer la base d'une prochaine étude sur les spécificités du français en Belgique.

60. Dans le sens de « démanger » ou de « titiller la curiosité » (Lebouc, 2006)

(41) BON ANNIVERSAIRE choupette. Alors c'est le grand jour. Je te souhaite tout ce que tu désires pour cette 21ème année. Gros bisou. Je suis très contente de **kotter** avec toi. (smsFB)

Enfin, il existe un ensemble de verbes dont la construction est originale. Nous avons déjà parlé de *savoir* ; c'est également le cas de *sonner* (quelqu'un), de *dire quoi* (à la place de *dire ce qu'il en est*) et du verbe dépronominalisé *promener* (exemple 42).

(42) OK c'est pas grave de toute façon y a eu 1 changement et Alex et ses parents viennent **promener** avec nous... (smsFB)

Nous recensons également, parmi les unités simples, de nombreux substantifs : certains caractérisent une réalité n'existant pas ou peu dans les autres aires de la francophonie. Ainsi, *le carré* (« surnom d'un quartier célèbre de Liège en forme de rectangle »), *un cercle* (« association rassemblant des étudiants issus d'une même faculté, ou encore, par restriction, son lieu de rassemblement »), *une corona* (« épreuve traditionnelle en vue de l'obtention d'un couvre-chef, la *calotte*, marquant l'appartenance à une association étudiante de type cercle »), *un cougnou* (« pain brioché »), etc. D'autres constituent une nouvelle dénomination pour une réalité existant par ailleurs et portant déjà une étiquette : *américain* (« steak haché ou tartare »), *athénée* (« lycée »), *baise* et *betch* (« baiser » ou « bisou », cf. exemple 43), *baptême* (« bizutage »), *baraki* (« personne avec peu de manières, souvent frimeuse mais pas forcément vulgaire »), *bompa* (« grand-père » ou « papi »), chicon (« endive »), etc.

(43) Avec 1 sms comme ça, comment veux-tu que le moral soit bon ? Point positif : je fume pas, je bois pas de coca ! Allez, encore une petite clope, on est plus à ça près... Grosses **baises** (smsFB)

Un ensemble de noms composés vient compléter cette liste. Ainsi, les structures traditionnelles AN (*fancy fair*, *sale linge*, *mobile-home*), NàN (*bac à linge*) et NdeN (*chambre du conseil*, *heure de fourche*, *salle d'étude*, *journal de classe*, cf. exemple 44).

(44) Tu viens ou pas ? On est 3 dans la **salle d'étude** pour l'instant donc ça s'annonce bien. Enfin on verra dans 15 minutes. Viens quand même ! On fera le travail de chauffage. (smsFB)

Les adjectifs typiques du français de Belgique[61] sont souvent des participes passés (d'un verbe existant aussi dans la variété de langue) employés comme adjectifs. On citera comme exemples *blindé*, *busé*, *clôturé*. De rares cas sortent de cette règle (*crollé*, *mouru* et *togé*).

La variété française de Belgique est également marquée par une série d'adverbes que l'on peut retrouver ailleurs en francophonie, mais qui sont employés plus fréquemment en Wallonie et à Bruxelles. Ainsi, *direct* (exemple 45), *assez bien*, *au mieux*, *encore bien*, *tantôt* et *pour du bon*.

61. Et pouvant apparaitre ailleurs en francophonie, mais avec une fréquence moins importante

(45) Si je suis passé hier ! Mais je suis **direct** parti à Hochport avec NPERS ! J'ai maths là ! J'ai trop une grosse migraine ! Trop forte ! (smsFB)

Les expressions, phénomènes assez rares dans l'eSMS car souvent couteux en termes de caractères, sont pourtant assez courantes dans les sms belges lorsqu'elles sont typiques du français de Belgique. On recense par exemple *ad fundum* (« à fond ! »), *oui mais non*[62], (« bien sûr que non »), *aimer autant* (« préférer », exemple 46), *faire sans*, *faire son/sa* + complément[63].

(46) J'**aime autant** savoir que tout se soit bien passé qu'autre chose pour vous ! On vient de terminer de nettoyer mais au moins, c'est 1 bonne ambiance ! 1 bon week-end, bisous (smsFB)

Les constructions à verbe support (ou CVS[64]) sont, dans le contexte de l'eSMS, plus fréquentes que les expressions. On retiendra des constructions avec :
– prendre : *prendre attention* (« faire attention », cf. exemple 47) ;
– avoir : *avoir bon*, *avoir difficile*, *avoir dur*, *avoir le temps long* ;
– être : *être fâché sur*, *être occupé avec* (cf. note 63) ;
– faire : *faire la file*, *faire drôle*.

(47) Y suis repassée 5 minutes elle m'avait vue descendre avec toi et m'a renvoyé 1 SMS. Suis restée avec le petit et pas **pris attention** à eux ! Je m'ennuie. Bise (smsFB)

Notre deuxième axe de recherche porte sur les 25 régionalismes les plus courants du corpus (détail en Annexe A.67). L'emploi le plus fréquent dans les sms concerne, comme nous venons de le préciser, le verbe *savoir* construit avec un verbe à l'infinitif. Un aperçu de 15 lignes de concordances (sélectionnées au hasard) pour *savoir* permet de mieux comprendre cette construction (Figure 8.30).

Si l'ensemble des cas illustrés dans la Figure 8.30 fait de *savoir* un synonyme de *pouvoir*, il est coutume de penser que *savoir* ne se construit pas suivi d'un infinitif en français standard. Or, ce n'est pas toujours le cas et la frontière entre l'emploi standard de *savoir* et son emploi en français de Belgique est assez floue[65].

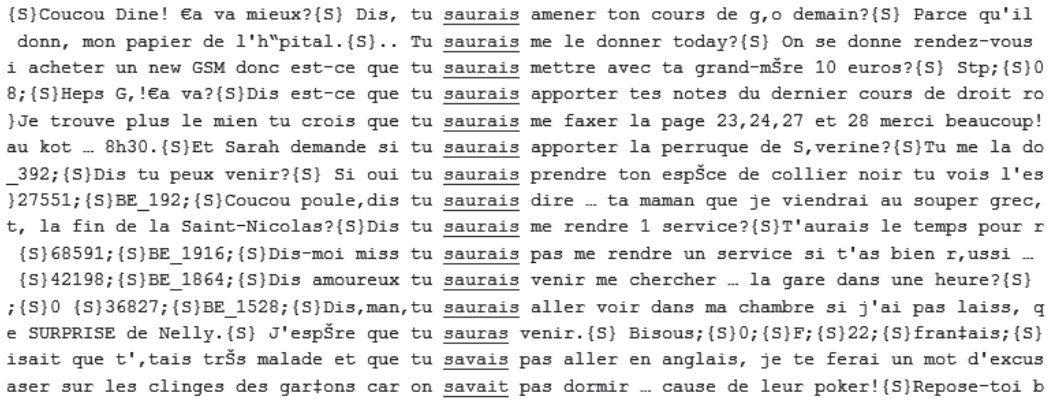
62. Répertorié chez Lebouc (2006) comme un adverbe.

63. Notons que *faire sans*, *faire son/sa* + complément et *être occupé avec* sont attestés dans des dictionnaires de référence du français standard et peuvent être assimilés à du français familier. Pourtant, dans le contexte belge, leur pratique est beaucoup plus importante qu'ailleurs, comme le corpus de sms le prouve. Le complément est souvent l'adjectif *malin/maligne* : *tu fais ton malin/ta maligne*

64. Ce sont les travaux de Maurice Gross et de son Laboratoire d'Automatique Documentaire et Linguistique qui ont caractérisé les constructions à verbe support en tant que phénomène original qui affecte autant le lexique que la grammaire. Selon leur théorie, on peut dire, sémantiquement parlant, que dans les CVS, le verbe est vide de sens, qu'il est le support de son ou ses compléments et qu'il n'ajoute pas ou peu de sens à celui déjà exprimé par son ou ses compléments.

65. En effet, prenons le cas de *savoir conduire* et *pouvoir conduire*. Dans une proposition telle que « Elle ne sait pas me conduire jusque Bruxelles », *savoir* est commutable avec *pouvoir* (sur l'axe paradigmatique). Pourtant, comment le français standard répondrait-il à la question « Pourquoi ne vient-elle pas à Bruxelles ? » ? S'il répond « elle ne peut pas conduire », il ne différencie pas l'aptitude

FIGURE 8.30 – *Exemple de concordance de la variété de savoir propre au français de Belgique*



L'interjection (ou marqueur du discours, selon la place et la fonction dans la phrase[66], cf. exemple 49) *dis* arrive en deuxième position dans le classement. La forme, **reprise dans les dictionnaires de référence du français**, mais **davantage employée en Belgique**, peut revêtir une fonction d'interpellation du destinataire ou d'interjection au sens propre (cf. exemple 48). Le plus souvent pourtant, *dis* introduit ou « annonce » une demande d'information (49), de service, de faveur, il est support d'une question. Dans certains cas, tels que les exemples 50 et 51, *dis* peut apparaitre comme un atténuateur, un signe d'embarras par rapport à la demande qui sera faite.

(48) **Dis** bordel.. on se calme ici.. c'est la fête ou quoi.. je suis vraiment désolé.. que faire pour me rattraper ? ? Mais stp on se connait à peine...tu vas pas commencer.. stp débloque-moi qu'on discute.. (smsFB)

(49) Hello poulette ! **Dis**, faut savoir faire exercice 5 page 2bis en maths ? J'en ai trop marre, c'est chiant **dis**.. Allez studie bien ! À demain ! Tchaw tchaw ! (smsFB)

(50) Salut petit papa ! **Dis** tu peux venir me chercher vers 9h OK ? Bisous (smsFB)

(la connaissance) et la possibilité (le moyen). La distinction se fera au moyen de paraphrases telles que « Elle n'a pas le permis » . Or, le dictionnaire de référence Rey *et al.* (2011) propose « être capable de », en tant que définition standard de *savoir* et donne comme exemple de référence : « à 4 ans, elle savait déjà lire et écrire ». Pour nous, il n'y a pas de différence entre « savoir + lire » ou « savoir + aimer » et « savoir + conduire ». Il s'agit dans les deux cas d'« être capable de » après un apprentissage, de la pratique, etc. La distinction entre l'adoption belge du verbe et son utilisation traditionnelle ne nous semble donc pas nette et le phénomène ressemble plutôt à un continuum entre français dit standard et variété. S'il existe des tendances majeures à l'utilisation de *savoir* avec cette structure et ce sens particuliers en fonction de la région, mais aussi, comme nous le verrons dans la partie quantitative, en fonction du sexe, la variation de la limite au-delà de laquelle *savoir* en tant que « être capable de » est remplacé par pouvoir nous semble en réalité individuelle.

66. Également repris dans l'expression *allez dis*.

(51) **Dis**, peux-tu, quand tu reçois ce SMS, me faire un signe pour me montrer que tout va bien ! Parce que t'avais pas l'air à l'aise au téléphone ! Voilà, passe une bonne nuit et à bientôt. Bisous. Moi (smsFB)

Le reste des items de la liste étudiée, excepté *tantôt* (« dans un passé ou un avenir proche »), *à tantôt* (« à tout à l'heure »), *gsm* et quelques autres, appartient au monde des étudiants (et des élèves) ; ainsi *kot*, *blocus*, *cokotteur*, *syllabus*, *cercle*, *farde*, *pété*, *rhéto*, etc. L'ensemble du répertoire des belgicismes décrit précédemment est d'ailleurs marqué par cette thématique. Il est vrai que **le folklore estudiantin** a fortement marqué le français de Belgique, mais c'est probablement la nature biaisée de notre corpus (nous avons vu que la population des étudiants y était fortement représentée) qui rend ces items si fréquents. Nous verrons si la thématique est aussi prégnante dans les autres régions ayant subi le même biais.

8.3.1.2 Variation diastratique

L'approche qualitative a permis de mettre en avant la diversité des formes régionales en usage dans les sms. Il est maintenant intéressant d'approfondir le sujet en examinant la quantité, la distribution et les dépendances de ces belgicismes.

Nous débutons avec un aperçu global de la quantité de sms contenant au moins un belgicisme (cf. Figure 8.31) : 17 % des sms sont concernés par ce phénomène. Il est intéressant de noter que **l'utilisation de régionalismes dans les sms du corpus belge étudié est deux fois plus importante** (en pourcentage du nombre de sms concernés) **que celle des éléments allogènes**.

FIGURE 8.31 – *Effectifs et pourcentages des sms contenant au moins un belgicisme*

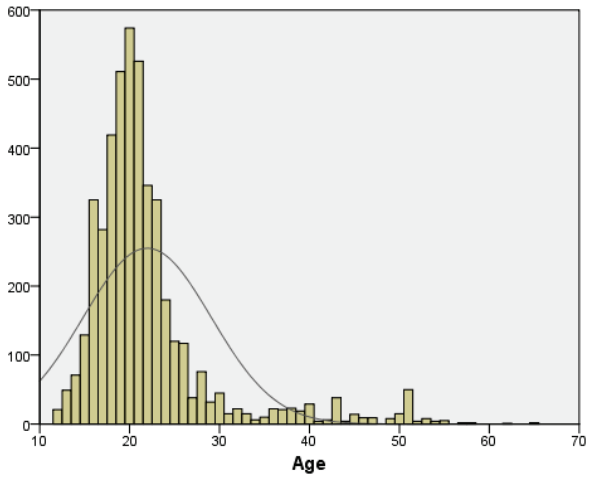
		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	24945	83,2	83,2	83,2
	1	5055	**16,9**	16,9	100,0
	Total	30000	100,0	100,0	

Examinons à présent la distribution des régionalismes en fonction des différentes variables : *sexe*, *âge* et *utilisation du service de chat*. Nous débutons avec l'analyse de la variable *sexe*. Comme nous l'avons vu lors de l'analyse des éléments allogènes, lorsqu'on désire travailler sur la base de deux variables qualitatives en statistiques, on utilise un test khi-carré. Le test, détaillé en Annexe A.62, présente une valeur p inférieure à 0,05 ce qui signifie qu'**il y a dépendance entre les variables *régionalisme* et *sexe*** dans le corpus belge. Lorsque nous réalisons un diagnostic case par case du tableau des résidus standardisés ajustés, nous affinons notre interprétation : le **taux de femmes usant de régionalismes est plus important** que celui auquel on aurait pu s'attendre. Au contraire, le taux d'hommes est

inférieur au taux escompté.

La dépendance entre *régionalismes* et *âge* peut être évaluée à l'aide d'un test t à condition que la distribution des âges soit symétrique. Puisque ce n'est pas le cas ici, nous ré-échantillonnons, tel que nous l'avons fait pour les éléments allogènes, à la population des moins de 31 ans (cf. AnnexeA.63). Le test en Annexe A.64 prouve une dépendance (la valeur p est inférieure à 0,05) entre les variables *âge* et *utilisation d'un belgicisme*. Cette dépendance prouve que la distribution des **sms comprenant des régionalismes concerne une classe d'âge plus restreinte** (les valeurs sont plus ramassées) que celle des sms n'en comprenant pas (la distribution est plus homogène ou variée).

FIGURE 8.32 – *Distribution, selon l'âge, des sms comprenant au moins un belgicisme*



Une analyse des variables *forum* et *chat* clôture la description générale des données. On peut en effet se demander si la pratique d'un mode de communication proche de celui du sms, qui est plus soutenue (on envoie en général un nombre important de messages chat) et moins contraignante (moins de limite d'espace) peut encourager l'emploi régulier de formes linguistiques régionales que l'on verrait réapparaitre dans les sms. Le test khi-carré pour la dépendance entre *belgicismes* et *utilisation du chat* aboutit à une valeur p supérieure à 0,05 ($p = 0{,}738$, cf. A.65). Les variables *régionalisme* et *utilisation du chat* sont donc indépendantes dans le corpus belge[67]. Pour ce qui est des forums, la valeur p est inférieure à 0,05 (cf. Annexe A.66) ; nous concluons donc à une dépendance des deux variables. Les résidus standardisés ajustés montrent que le **taux d'usagers des forums usant de régionalismes est moins important** que celui auquel on aurait pu s'attendre. Au contraire, le taux de non usagers des forums est supérieur au taux escompté.

67. Relativisons toutefois ces résultats. En effet, en 2004, l'utilisation du chat était beaucoup moins fréquente que celle du forum, une tendance qui s'est probablement inversée aujourd'hui.

Après avoir abordé le lien de trois variables importantes issues du corpus de profil avec l'ensemble des belgicismes du corpus de sms, nous analysons un régionalisme en particulier : le régionalisme le plus courant, *savoir*. Son utilisation est-elle le fait d'une classe particulière ? Pour y répondre, nous allons réaliser des tests statistiques à partir des variables *sexe*, *occupation* et *utilisation du courrier électronique*. Le but est de comprendre si le profil des usagers associés à *savoir* diffère de celui des usagers associés aux autres régionalismes. Le khi-carré (Figure 8.33) est appliqué aux deux variables qualitatives *savoir* et *sexe*. La valeur p est égale à 0,002 ($< 0{,}05$), ce qui signifie que les deux variables sont dépendantes. L'observation des résidus standardisés ajustés montre que le **taux de femmes usant de savoir (dans sa construction particulière au français de Belgique) est plus important** que celui auquel on aurait pu s'attendre. Au contraire, le taux d'hommes est inférieur au taux escompté.

FIGURE 8.33 – *Test khi-carré mesurant la dépendance des variables savoir et sexe en Belgique*

Tableau croisé Sexe * Savoir

			Belg1		
			0	1	Total
Sexe	F	Effectif	2109	700	2809
		Résidu ajusté	-3,2	3,2	
	M	Effectif	1393	367	1760
		Résidu ajusté	3,2	-3,2	
Total		Effectif	3502	1067	4569

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	10,002[a]	1	,002		
Correction pour la continuité[b]	9,776	1	,002		
Rapport de vraisemblance	10,107	1	,001		
Test exact de Fisher				,002	,001
Nombre d'observations valides	4569				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 411,01.

b. Calculé uniquement pour un tableau 2x2

Afin de juger de la dépendance entre l'utilisation du verbe *savoir* et le type d'occupation, nous utilisons également un test khi-carré (cf. Annexe A.68). La valeur p trouvée est égale à 0,448 ($> 0{,}05$). **Les deux variables *utilisation de savoir* et *type d'occupation* sont donc indépendantes**.

Enfin, nous appliquons le test khi-carré aux variables emploi de *savoir* dans les sms et pratique du courrier électronique (cf. Annexe A.69). La valeur p trouvée est égale à 0,248 ($> 0{,}05$). **Les deux variables emploi de savoir dans les sms et pratique du courrier électronique sont donc indépendantes**.

8.3.2 Régionalismes de la Réunion

L'Ile de la Réunion se situe dans l'Océan Indien à 700 km à l'est de Madagascar. Sa seule langue officielle est le français[68]. Elle est colonisée en 1665 et devient DOM (Département français d'outre-mer) en 1946. Cette situation de colonisation amène des mutations sociolinguistiques, évoluant d'une diglossie avec le français en langue dite « haute » et le créole en langue dite « basse » vers une attitude moins méprisante et un bilinguisme institutionnalisé (Beniamino, 1996 ; Bordal et Ledegen, 2007 ; Cougnon et Ledegen, 2010 ; Ledegen, 2004 ; Prudent, 2001). Sur le terrain, on atteste un mélange constant des deux langues française et créole, autant dans les échanges lexicaux, phonétiques et syntaxiques qu'au niveau même de la morphologie (Beniamino et Baggioni, 1993 ; Carayol, 1985 ; Chaudenson, 1974). C'est ce phénomène que Chaudenson *et al.* (1993, p. 391) nomme l'« osmoticité » entre les deux langues. D'un point de vue lexical, cette symbiose linguistique franco-créole qui existe à la Réunion cause le problème, pour le linguiste, « de savoir à quelle langue appartiennent les mots, quelles que soient par ailleurs les représentations des locuteurs »(Beniamino, 1996 ; Ledegen, 2011).

Nous avons utilisé Beniamino (1996) comme ressource de base pour la constitution du lexique. Nous n'avons malheureusement pas pu accéder aux données numérisées et avons donc transformé un document pdf en document texte. Nous avons ensuite trié les entrées pour les épurer d'un certain nombre de coquilles engendrées par la méthode de reconnaissance de caractères à partir d'un document pdf. Nous avons également trié les résultats de cette recherche puisqu'un grand nombre d'homographes au français standard sont insérés dans le dictionnaire (par ex. *passer* à la 1e et 3e personne du singulier et *passe* « Mouvement d'imposition des mains à but thérapeutique »). Enfin, le dictionnaire propose, pour les expressions, une entrée pour chaque unité de ces expressions. Ainsi, *jour : barre de jour, fait-jour, fête des dix jours, jour de l'an tamoul*, où *jour*, bien que n'ayant pas une signification propre au français de La Réunion, constitue une entrée du dictionnaire qui reprend ces expressions. Nous avons donc été également contrainte de trier manuellement ces formes[69].

Nous dénombrons 374 sms comportant au moins un régionalisme dans le corpus réunionnais. Il s'agit d'un nombre somme toute assez faible, ce qui est probablement dû à la mauvaise reconnaissance des formes induite par la décision prise par l'équipe locale de respecter le plus possible la graphie créole. Nous pensons qu'un certain nombre d'unités propres au français de La Réunion n'ont ainsi pas été reconnues car elles suivaient les propriétés graphiques de l'eSMS qui, comme nous l'avons

68. Même si son statut de langue régionale n'est pas effectif, le créole a toutefois une place de plus en plus grande dans l'enseignement : depuis 2001, les établissements scolaires du premier degré peuvent proposer soit un enseignement en langue vivante réunionnaise, soit un enseignement bilingue créole/français. Dans le secondaire, une option « Langue et culture régionales » est proposée (Source : entretiens avec Gudrun Ledegen).

69. Notons que cette méthodologie a été appliquée avec le plus grand soin, mais ne permet pas d'obtenir des résultats parfaits, étant donné que la vérification manuelle est elle-même imparfaite.

expliqué, se confondent facilement avec celles du créole. Le détail des 79 formes régionales identifiées se trouve en Annexe A.70.

8.3.2.1 Approche qualitative

Le tableau des 79 unités régionales de La Réunion met en avant la diversité grammaticale des formes existantes. On trouve, comme en Belgique, des mots simples, des mots composés et des expressions, mais aucun cas de construction à verbe support.

Au sein des unités simples, nous remarquons de nombreux substantifs : *bec* (« bisou »), *boug* (« bougre », exemple 52), *cari* (plat typique de La Réunion), *case* (« maison »), *dalon* (« copain », « camarade ») et *nuite* (« nuit »). Les noms composés sont également courants, surtout dans les structures AN et NN, et sont séparés par une espace blanche (*gros soleil*, « moment où le soleil est le plus fort ») ou par un trait d'union (*thon-banane*, « femme corpulente »). Enfin, il existe à La Réunion deux syntagmes nominaux figés (un type que l'on ne recense pas dans les sms belges par exemple) qui sont *ma fille* et *les bas*.

(52) Merci à toi aussi mon **boug** ! Mi sava dodo osi ! Bye bye ! (smsFR)

Les verbes constituent une classe importante des régionalismes à La Réunion : *battre* (« se cogner », exemple 53),*défouler* (« amuser »), *lancer* (« élancer »), etc. *Gagner* est un cas particulier car il revêt de nombreux sens. Beniamino (1996) recense trois grandes classes sémantiques, réparties approximativement comme suit : « pouvoir, parvenir à », « posséder, obtenir » et « attraper, éprouver ». **Dans les sms réunionnais, ces trois grands groupes sont effectivement représentés**, comme l'illustrent les exemples 55, 56 et 57.

(53) Fais descendre, le message n'est pas fini ! Tu veux vraiment savoir ? ... fais attention à pas **battre** la tête trop fort au plafond ! J'ai le jeu Zuma ! (smsFR)

(54) Là je suis sur la route ! Où je ne sais pas ! Il y a plein de jolis arbres avec des feuilles jaunes et vertes... Et pour gâcher le tout une bagnole qui **boucane** ! Tu fais quoi ? (smsFR)

(55) En fait [NPERS] la di kel i paie [NPERS] si li **gagne** convaincre [NPERS] de sorte ek el. Sé pou sa ke [NPERS] i ve ke [NPERS] i sorte ek [NPERS]. C'est un truc de oufs. (smsFR)

(56) Hihi... tu penses quand on va **gagner** nos bulletins du troisième trimestre ? (smsFR)

(57) Un instant ça se peut que j'vais **gagner** une crampe au pouce ! (smsFR)

Quelques adjectifs tels que *mol* (« peu animé, ennuyeux »), *chauffé* (réchauffé), *gaillard* (« en bonne santé ») et *vilain* (« abimé », exemple 58) sont également observés dans le corpus.

(58) C'est pas que je vous mens mais j'ai un **vilain** téléphone. (smsFR)

Les adverbes constituent, comme en Belgique, une classe importante de régionalismes dans les sms : *dans le fond* (« en bas »), *doucement doucement* (« peu à peu »), *jusqu'à l'heure* (« jusqu'à maintenant, jusqu'à aujourd'hui »), *sur le tard* (« tard dans la soirée », exemple 59).

(59) J'étais en amoureux hier soir ! Je suis invité ce soir et je taf demain. Peut-être **sur le tard** mais juste l'aller retour. (smsFR)

Enfin, nous notons deux expressions régionales : *battre (un) carré* et *sonner quelqu'un*. La première, communément employée dans la vie quotidienne, surtout à l'oral (Bordal et Ledegen, 2007), signifie « se promener, faire un tour ». La seconde, signifiant « appeler quelqu'un » ou « téléphoner à quelqu'un », est commune au français de Belgique.

Nous reprenons la classification émise par Bordal et Ledegen (2007, p. 123) comparant les régionalismes de La Réunion au français de référence. Il existe donc dans le corpus de sms :
– des sens absents du français de référence : *bec*, *battre*, *défouler*, *gagner*, *les bas*, ...
– des formes absentes du français de référence : *dalon*, *kalou* (« mortier »), *moukaté* (« être méprisable, sans valeur »), *zamal* (« chanvre indien »), ...
– des mots composés ou locutions verbales absents du français de référence : *battre (un) carré*, *thon-banane*, *gros marraine* (« épouse du parrain »), ...
– des formes existantes en français de référence mais présentant une très forte fréquence dans les sms réunionnais : par exemple *case* (« maison »), *tantine* (« épouse, femme en général »).

Ces catégories forment les catégories traditionnelles des régionalismes. Nous proposons ici d'en ajouter une cinquième, typique du français de La Réunion : des variations graphiques de formes existant en français de référence et dont le sens est le même que celui du français de référence, avec ou sans modification phonétique. C'est le cas de *lontan*, de *boug*, de *manzé* et de *nuite*.

Deux spécificités ressortent de l'analyse thématique de ces régionalismes. Premièrement de nombreux régionalismes ont pour référents des réalités qui font davantage sens à La Réunion qu'en métropole ou dans le reste de la francophonie (*gros soleil*, *cafre* « réunionnais(e) d'origine africaine », *tamarin* « Fruit dont les petites fleurs rougeâtres se transforment en gousses, renfermant chacune 5 ou 6 graines servant à assaisonner bon nombre de plats et boissons typiques de La Réunion », etc.). Deuxièmement, on recense un nombre important de termes hypocoristiques spécifiques (notamment par leur fréquence) au corpus de sms de La Réunion (*chouchou*, *coco*, *tantine* et surtout *doudou*).

Notre deuxième axe de recherche porte sur les 25 régionalismes les plus courants du corpus (détail en Annexe A.71). On retrouve en tête *case*, *doudou* et *gagner* dont nous avons déjà brièvement parlé. Nous choisissons d'approfondir l'adjectif

mol (18 occurrences) dont la construction nous interpelle. Le dictionnaire Beniamino (1996) le présente comme qualificatif de personnes ou d'ambiances. Or, sur les 18 occurrences du corpus, 6 formes se construisent avec des pronoms créoles (*wi*, *ou*, *li*, *moin*...), une forme apparait seule en début de message (comme réponse à une question posée dans un précédent message), une forme qualifie un corps (*mon kor*) et 11 formes apparaissent dans l'expression *c'est mol*, dont une avec complétive (*c'est mol de*). La construction *c'est mol*, la plus récurrente du corpus, ainsi que l'attribution de l'adjectif au mot *corps* (ne qualifiant ainsi ni une personne ni une ambiance) viennent donc compléter celles recensées par le dictionnaire.

8.3.2.2 Variation diastratique

Le tour d'horizon de la typologie des formes régionales étant clôturé, nous proposons maintenant un focus statistique. Les statistiques descriptives basiques révèlent que la quantité de messages comportant au moins un régionalisme (Figure 8.34) est de 3 %. Comparé à la Belgique, ce pourcentage est assez faible probablement parce que, comme nous l'avons expliqué, les graphies de l'eSMS se confondent volontiers avec celles du créole. De plus, étant donné que les zones « flottantes » (Ledegen, 2011) entre créole et français de La Réunion sont nombreuses, le transcripteur a probablement préféré laisser telles quelles de nombreuses formes dont il doutait. Ce biais étant pris en considération, il est tout de même intéressant de noter que **l'utilisation de régionalismes dans les sms réunionnais est moins importante** (en pourcentage du nombre de sms concernés) **que celle des éléments allogènes** (4,2 %).

FIGURE 8.34 – *Effectifs et pourcentage des sms contenant au moins une unité spécifique au français régional de La Réunion*

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	12287	97,1	97,1	97,1
	1	373	2,9	2,9	100,0
	Total	12660	100,0	100,0	

Examinons à présent la distribution des régionalismes (variable qualitative) en fonction du sexe de l'auteur du message (variable qualitative). On utilisera donc un test khi-carré. Le test, détaillé en Annexe A.72, présente une valeur p inférieure à 0,05 (p = 0,002) ce qui signifie qu'**il y a dépendance entre les variables régionalisme et sexe** dans le corpus réunionnais. Lorsque nous établissons un diagnostic case par case du tableau des résidus standardisés ajustés, nous affinons notre interprétation et pouvons conclure que le **taux de femmes usant de régionalismes est plus important** que celui auquel on aurait pu s'attendre. Au contraire, le taux d'hommes est inférieur au taux escompté.

La dépendance entre *régionalisme* et *âge* est évaluée à l'aide d'un test t à condition que la distribution des âges soit symétrique. Comme nous l'avons conclu lors de l'analyse des éléments allogènes, ce n'est pas le cas à La Réunion. Nous savons également qu'un rééchantillonnage de la population à la sous-population des moins de 31 ans n'aide pas dans cette région. Nous avons donc recours à un test de Mann-Whitney (cf. Annexe A.73) qui permet de comparer les tendances centrales (que nous appliquons à la sous-population pour que la méthode soit comparable à celle suivie pour les autres régions). On remarque que p = 0,005 ; le test nous incite donc à rejeter l'hypothèse nulle (H0) puisque les médianes des régionalismes en fonction des âges sont inégales.

Nous clôturons cette description générale des données avec une analyse des variables *forum* et *chat*. Le test khi-carré pour la dépendance entre *régionalisme* et *chat* à La Réunion aboutit à une valeur p supérieure à 0,05 (p = 0,124, cf. Annexe A.74). Les variables *régionalismes* et *utilisation du chat* sont donc indépendantes dans le corpus réunionnais. L'hypothèse émise pour la Belgique selon laquelle cette indépendance serait liée à l'ancienneté du corpus ne semble donc pas se confirmer. Le corpus réunionnais ayant été collecté en 2008, nous attendons les résultats des deux autres régions avant de nous avancer sur la question. Pour ce qui est des forums, la valeur p est inférieure à 0,05 (cf. Annexe A.75) ; nous concluons donc à une dépendance des deux variables. Les résidus standardisés ajustés montrent que le taux d'usagers des forums usant de régionalismes est moins important que celui auquel on aurait pu s'attendre. Au contraire, le taux des non usagers des forums est supérieur au taux escompté.

Contrairement à ce que nous avons entrepris dans la partie réservée aux belgicismes, il n'est pas possible, sur la base du corpus réunionnais, d'étudier la variation de certaines unités en fonction de variables sociodémographiques. En effet, nous possédons uniquement 140 sms comportant au moins un régionalisme et rattachés à un profil. De plus, ces 140 sms sont partagés par seulement 43 auteurs. Il serait donc risqué de dresser des tendances avec un échantillon si petit.

8.3.3 Régionalismes du Québec

La colonisation du Québec est une colonisation française qui prend fin avec les guerres coloniales (1689 – 1760) dont les Anglais sortent vainqueurs. À la suite de cette conquête anglaise, l'anglais devient la langue du domaine public, de l'économie et du commerce. Ce n'est pas avant 1867 qu'est imposé, avec la Loi constitutionnelle (Acte de l'Amérique Britannique du Nord), le bilinguisme aux niveaux parlementaire et fédéral. Le statut du français reste plus ou moins le même jusque dans les années 1960 avec la Révolution tranquille. À partir de ce moment, la classe moyenne francophone devient plus importante et les instances se questionnent sur l'usage du français dans tous les domaines de la vie et de l'État. La Loi sur la langue officielle, puis la Charte de la langue française (ou loi 101),

posent de nouvelles exigences : le français devient la langue officielle du Québec [70].

Depuis lors, le Québec maintient une politique linguistique soutenue qui protège et promeut la langue française, notamment via sa politique d'utilisation du français dans les technologies de l'information et des communications et des outils rigoureux constamment mis à jour : *Le grand dictionnaire terminologique* (le GDT [71]) et la *Banque de dépannage linguistique* [72].

Le lexique du français québécois est marqué par ce qui ressemble à nos yeux à des archaïsmes (mais ne sont pas perçus comme tels par les locuteurs locaux), par l'adstrat anglo-américain et surtout par l'innovation (néologismes, fortement encouragés de manière institutionnelle) (Plourde *et al.*, 2001 ; Pöll, 2001 ; Spencer, 2008 ; Valdman, 1979).

Le dictionnaire de québécismes que nous utilisons provient de la base de données lexicographiques du Québec (Poirier, 2011) [73]. Il a été épuré d'un certain nombre d'annotations. Les doublons (unités polysémiques) ont été simplifiés. La marque du féminin des adjectifs, initialement linéaire (par ex. *crapet*, *ette*), se manifeste dorénavant par une seconde entrée (*crapet* suivi de *crapette*), excepté lorsque cet adjectif et sa forme féminine sont reconnus par le dictionnaire de référence (par ex. *balancé*, *ée*), dans tel cas, une entrée unique est maintenue, la forme fléchie étant reconnue par le dictionnaire de référence.

Les régionalismes du Québec sont au nombre de 34 et apparaissent dans 186 messages (il s'agit donc d'un corpus de taille assez réduite) que nous décortiquons. Le détail des 34 formes régionales identifiées se trouve en Annexe A.76.

8.3.3.1 Approche qualitative

Un rapide aperçu des 34 québécismes répertoriés permet de constater que les régionalismes de notre corpus concernent les mots simples, les mots composés et les expressions.

Au sein des unités simples, nous dénombrons 4 verbes : *autographier* (« dédicacer »), *adonner* (« convenir (à) », cf. exemple 60), *souper* et *diner* (pour le repas de midi).

(60) Moi aussi ! Des idées d'endroits qui t'**adonnent** ? (smsFQ)

Les substantifs simples sont plus nombreux. *Char*, pour « voiture », est un régionalisme fréquent du corpus (16 occurrences, cf. exemple 61). On recense également,

70. La Charte sera modifiée à plusieurs reprises pour anticonstitutionnalité. La langue anglaise y retrouve quelques droits, notamment en ce qui concerne la scolarisation ou l'affichage public.

71. Pour en savoir plus : `http://www.oqlf.gouv.qc.ca/ressources/gdt.html`

72. Pour en savoir plus : `http://www.oqlf.gouv.qc.ca/ressources/bdl.html`

73. Nous remercions sincèrement Claude Poirier et Jean-François Smith pour le partage de leur ressource lexicale sous forme de nomenclature simple.

entre autres exemples, *bande* (« groupe de musique »), *dépanneur* (« épicerie de proximité »), *meeting* et *sofa* (régionalismes par leur fréquence d'emploi élevée), *beigne* (« pâtisserie traditionnelle en forme d'anneau »), *blonde* (« copine »), *épluchette* (« fête ») et *sloche* (« neige sale, gadoue »). Un seul nom composé apparait dans les sms : *avant-midi* pour « matin » (exemple 62).

(61) Vous êtes rendus où pis vous allez où, exactement ? En **char** ou à pied ? (smsFQ)

(62) Hey oki cool ! Jeudi es-tu à l'école ? Je vais être au palais pour l'**avant-midi**, probablement... (smsFQ)

Les adjectifs régionaux prennent également une place importante dans les sms québécois. On retiendra par exemple *abusif* (« exagéré »), *cave* (« idiot, bête, débile » [74], cf. exemple 63), *dispendieux* (« cher, pénible »), cf. exemple 64, *fin* (« gentille », cf. exemple 65), *nono* (« niais, innocent »), *safe* (« qui est en sécurité, sûr ») et *smatte* (« qui fait preuve d'intelligence, subtil »).

(63) Ouin, c'est ça, avec ma mère, je n'aurais pas la chienne, mais là , mon père ... ça sent le : estie qu'il est **cave** " à la fin du souper ." (smsFQ)

(64) Rien de TROP dispendieux... Mais quand même, on peut se permettre ! (smsFQ)

(65) Tu es **fin**, mais ça ne me dit pas c'est quoi ! (smsFQ)

Nous relevons également six expressions propres à l'usage courant des locuteurs québécois : *avoir de l'allure* (« avoir du jugement, du bon sens ou de l'entrain », cf. exemple 66), *avoir la chienne* (« avoir peur »), *être crampé (de rire)* (« rire aux éclats »), *en baboune* (« en moue », cf. exemple 67), *bien le bonjour* (plus rare en contexte spontané dans le reste de la francophonie) et *good luck* (« bonne chance », emprunt à l'anglais répertorié dans la base de données lexicographiques québécoise).

(66) Pis, pis, pis, ça **a de l'allure** ta soirée ? :) (smsFQ)

(67) Lol Non, j'ai ri ! J'étais crampée de sa réaction. Il s'est arrêté de marcher en plein milieu de la rue et le sourire lui est tombé **en baboune** ! Lol (smsFQ)

Enfin, nous retenons les adverbes : *à soir* (« le soir, ce soir »), *à matin* (« le matin, ce matin ») [75] et enfin *pis* pour « puis », qui n'est pas répertorié dans notre base de données de référence. Même s'il s'agit d'un adverbe familier, voire populaire, nous avons décidé de le garder dans notre analyse, étant donné qu'il apparait fréquemment (18 occurrences) dans le contexte des sms et qu'il n'est pas commun à toute la francophonie.

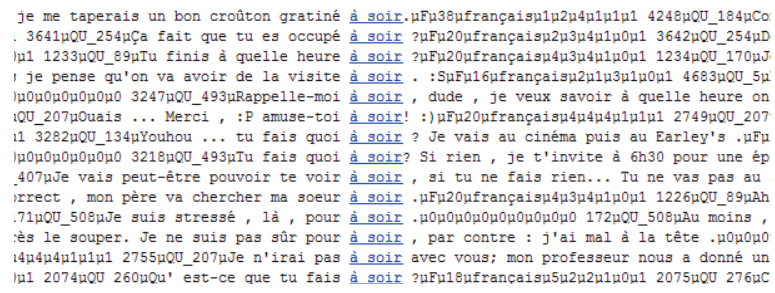
74. Notons que *cave* existe également en français argotique, avec le sens de « niais, dupe ».

75. Nous remarquons un rapprochement de ces deux formes québécoises avec les formes belges *au soir* et *au matin*.

Ce panorama de la typologie des québécismes mène à deux remarques générales. Premièrement, l'influence de l'anglais, langue pourtant maitrisée au niveau de la politique linguistique du Québec, est prépondérante. En effet, on remarque l'intégration dans la base de données officielle de québécismes les emprunts à l'anglais suivants : *meeting*, *pouding*, *safe*, *funny* et *good luck*. Ensuite, du point de vue thématique, les régionalismes concernent en grande partie les repas et l'alimentaire en général [76] : *pouding*, *beigne*, *bagel*, *souper* et *diner*.

La deuxième partie de cette analyse qualitative se focalise sur le contexte d'apparition, non pas du régionalisme le plus fréquent du corpus québécois (*souper*), qui apparait par ailleurs (Suisse, Belgique, etc.), mais du second (en nombre d'occurrences), l'adverbe *à soir*, synonyme de *ce soir*. La Figure 8.35 livre un aperçu du type des contextes d'utilisation de *à soir*.

FIGURE 8.35 – *Exemple de concordance pour l'adverbe à soir dans le corpus de sms québécois*



Les 32 occurrences de l'adverbe se répartissent comme suit (par ordre de fréquence) :
- il apparait en fin de proposition, en tant qu'adverbe de temps, pour fixer (proposition affirmative, plus rare : *C'est correct, mon père va chercher ma sœur à soir.*) ou s'enquérir (proposition interrogative, fréquente : *Tu fais quoi à soir ? Si rien, je t'invite à 6h30 pour une épluchette*) d'un rendez-vous ou d'un événement ;
- il apparait en début de proposition, en proposition affirmative, pour informer l'interlocuteur au sujet d'un événement : *à soir, c'est mon souper avec mon père . :-S* ou *à soir, je garde les petits gars de l'aide aux devoirs* ;
- il apparait en milieu de proposition : *Je n'irai pas à soir avec vous* ;
- il est introduit par une préposition : dans ce cas, la forme *à soir* est associée à un référent plus large (événements implicites liés à la soirée) que l'adverbe de temps *à soir* seul : *Je suis stressé, là, pour à soir.*

8.3.3.2 Variation diastratique

L'approche qualitative a permis de mettre en avant la diversité des formes régionales, les tendances en matière thématique et quelques constructions intéressantes en usage dans les sms. Il est maintenant important d'approfondir cette analyse en

76. Il s'agit peut-être d'un biais engendré par le corpus.

examinant la quantité et la distribution de ces québécismes.

Nous débutons avec un aperçu global de la quantité de sms contenant au moins un québécisme (Figure 8.36) : 3,6 % des sms sont concernés par ce phénomène. Il est intéressant de noter que **l'utilisation de régionalismes dans les sms du corpus québécois est cinq fois moins importante** (en pourcentage du nombre de sms concernés) **que celle des éléments allogènes**.

FIGURE 8.36 – *Effectifs et pourcentage des sms contenant au moins un québécisme*

Québécismes

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	4560	96,4	96,4	96,4
	1	171	3,6	3,6	100,0
	Total	4731	100,0	100,0	

Examinons à présent l'utilisation ou non des québécismes en fonction du sexe de l'auteur d'un message. Le test khi-carré, utilisé pour y parvenir et détaillé en Annexe A.77, présente une valeur p supérieure à 0,05 (p = 0,205) ce qui signifie qu'**il y a indépendance entre les variables *régionalisme* et *sexe*** dans le corpus québécois.

La dépendance entre *régionalisme* et *âge* est évaluée à l'aide d'un test t à condition que la distribution des âges soit symétrique. Comme nous l'avons conclu lors de l'analyse des éléments allogènes, ce n'est pas le cas au Québec et un rééchantillonnage ne permet pas non plus de parvenir à une courbe normale. Nous avons donc recours à un test de Mann-Whitney (cf. Figure 8.37) qui permet de comparer les tendances centrales (que nous appliquons à la sous-population des moins de 31 ans pour que notre méthode soit comparable à celle appliquée pour les autres régions). On remarque que p = 0,972 ; le test nous incite donc à accepter l'hypothèse nulle (H0) puisque les médianes des régionalismes en fonction des âges sont égales. **Dans le corpus québécois, les auteurs n'emploient pas plus de régionalismes lorsqu'ils sont plus jeunes et inversement**.

Une analyse des variables *forum* et *chat* clôture la description générale des données. Le test khi-carré pour la dépendance entre *québécismes* et *chat* aboutit à une valeur p supérieure à 0,05 (p = 0,244, cf. Annexe A.78). Les variables *régionalismes* et *utilisation du chat* sont donc indépendantes dans le corpus québécois. Pour ce qui est des forums, la valeur p est également supérieure à 0,05 (p = 0,292, cf. Annexe A.79) ; nous concluons donc que, dans le corpus québécois, l'emploi de régionalismes n'est pas dépendant de l'utilisation de services de chat et de forums de discussion.

Étant donné qu'il n'existe que 125 messages à la fois liés à un profil et comportant un régionalisme, nous ne pouvons pas aller plus en avant dans l'analyse d'un ré-

FIGURE 8.37 – *Test de Mann-Whitney pour la comparaison des médianes d'utilisation de régionalismes selon l'âge, au Québec*

Récapitulatif du test d'hypothèse

	Hypothèse nulle	Test	Sig.	Décision
1	Les médianes de Age sont identiques pour les catégories de Quebecismes.	Test de la médiane à échantillons indépendants	,972	Retenir l'hypothèse nulle.
2	La distribution de Age est identique sur les catégories de Quebecismes.	Test U de Mann-Whitney à échantillons associés	,046	Rejeter l'hypothèse nulle.

Les significations asymptotiques sont affichées. Le niveau de signification est .05.

gionalisme en particulier au Québec (les conclusions seraient fondées sur un petit nombre d'occurrences et la majorité des tests statistiques seraient ainsi invalidés).

8.3.4 Régionalismes de Suisse

Les langues nationales de la Suisse sont l'allemand, le français, l'italien et le romanche[77]. Le français pénètre en Suisse romande à partir du XIIIe siècle sous forme écrite (en concurrence avec le latin) et du XVIIe sous forme orale (en concurrence avec le patois) (Knecht, 1996). La Constitution fédérale de la Confédération[78] helvétique de 1938 (révisée en 1996) institue un découpage en trois régions linguistiques auxquelles s'applique le principe de territorialité[79](Pöll, 2001). On nomme Suisse romande ou Romandie les territoires de langue française. Étant donné la nature du corpus francophone suisse à la base de cette étude[80], nos messages ne concernent pas uniquement des locuteurs de la Suisse romande mais également des locuteurs francophones de Suisse. Les régionalismes en Suisse francophone sont marqués par le substrat dialectal, des archaïsmes, l'innovation et l'adstrat germanique. « La question de l'adstrat allemand ou alémanique est un chapitre délicat du français de Suisse. S'il va sans dire qu'un voisinage de plusieurs siècles exerce une influence, on s'étonnera cependant de la véhémence avec laquelle les germanismes, vrais ou supposés, sont combattus » (Pöll, 2001).

Notre travail sur les helvétismes est basé sur le repérage manuel (et donc plus précis que celui pour les autres régions mais forcément moins systématique) de ceux-ci par une locutrice suisse francophone vivant en Belgique[81].

77. Le romanche est considéré comme quatrième langue nationale, mais uniquement dans le canton des Grisons.

78. Fondée en 1291, la Confédération reste à dominance germanophone jusqu'en 1798. La naissance de la République helvétique (1798-1814) et l'occupation de l'armée française sont les deux événements qui engendrent une Suisse trilingue.

79. Chaque aire linguistique est théoriquement monolingue.

80. Il est constitué de tout message comprenant *a minima* une unité lexicale du français.

81. Nous remercions à nouveau Ariane Torné pour le remarquable travail de transcription, d'éti-

Les régionalismes de Suisse sont au nombre de 35 et apparaissent dans 551 messages que nous décortiquons. Le détail des 35 formes régionales identifiées se trouve en Annexe 8.38.

8.3.4.1 Approche qualitative

Le tableau des 35 unités régionales de Suisse (Figure 8.38) permet d'observer une certaine diversité dans les formes existantes. Tout d'abord, nous notons une spécificité régionale qui consiste en l'abrègement quasi systématique des noms de lieux[82] : ainsi, on dit *Neuch* (pour Neuchâtel), *Lauz* (pour Lauzanne) et *Tchaux* (pour Chaux-de-Fonds).

FIGURE 8.38 – *Répertoire des helvétismes identifiés dans le corpus de sms suisse*

aller à la douche	entre deux	pire + A
aula	faire la noce	pour finir
armée (de milice)	faire long	redire
autrement	gentiment	relais
avoir mal au cou	heure blanche	salutations
avoir trop de peine à	les sous	souper
bec	mensa	tout de bon
boguet	monstre	un (petit) coup
ça joue	natel	V + seulement
ces jours	ne pas être bien	
ces temps	ou bien	
c'est comment	panosse	
c'est tout bon	penser bien	

On trouve, comme dans les autres régions, des mots simples, des mots composés, des constructions à verbe support et des expressions. Au sein des unités simples, nous remarquons étonnamment peu de substantifs (11) : *bec* (« bisou »), *boguet* (« vélomoteur »), *mensa* (« cafétéria, cantine »), *panosse* (« serpillière »), *armée* (abréviation pour *armée de milice*, armée où les civils sont formés pour participer à des missions militaires, « service militaire »), *natel* (« téléphone portable »),

quetage et d'annotation qu'elle a exécuté sur le corpus suisse francophone. Nous remercions également Aurélia Robert-Tissot et le Prof. Elisabeth Stark pour la validation de ce travail peu aisé de défrichement sur le terrain des régionalismes suisses. Le choix de faire identifier les régionalismes d'une aire donnée par les locuteurs de cette aire est discutable puisqu'il peut entrainer un biais considérable ; c'était toutefois la meilleure solution dont nous disposions. Malgré notre volonté de compléter (ou même confronter) ces données avec un dictionnaire ou une base numérisée (notamment via une demande explicite faite auprès des responsables suisses de la Base de données lexicographiques panfrancophone), nous n'avons rien obtenu. Notre enquête se limite donc à la base constituée manuellement et vérifiée notamment grâce à Thibault et Knecht (2004).

82. Il s'agit d'une habitude de l'oral qui se perçoit de plus en plus à l'écrit, notamment au travers de la CéMO, comme en témoignent les nombreux sites Internet dont le nom même comprend l'abréviation latchaux.ch, neuch.ch.

salutations[83] (« salut, ciao, bye », cf. exemple 68), etc. Les noms composés sont moins courants, nous recensons uniquement *heure(s) blanche(s)* pour « heure(s) libre(s) ». De la même façon, nous n'avons trouvé qu'un seul syntagme nominal figé, *les sous*[84].

(68) Tchô ! Oui j'ai reçu ton SMS mais j'ai rendez-vous sur Skype avec mon frère vers 17-18h. Mais suis libre plus tard ou alors jeudi soir ! **Salutations** (smsFS)

Les verbes simples constituent une classe très réduite de régionalismes en Suisse. On recense uniquement *redire* (pour « confirmer ») et *souper* (pour « diner », cf. exemple 69) qui sont, par contre, assez fréquents (respectivement 86 et 61 occurrences). Notre méthodologie a permis d'identifier un seul adjectif, typique de Suisse par sa place dans la phrase (phénomène syntaxique) : *monstre* (« grand, énorme, conséquent », cf. exemple 70).

(69) Salut, je ne vais pas **souper** à la maison ce soir. Je rentre peut-être en cours d'aprèm par contre... (smsFS)

(70) Coucou, si ça dérange pas trop je vais pas venir dîner.... **Monstre** boulot .....(smsFS)

Les adverbes constituent par contre une classe importante de régionalismes suisses dans les sms : *autrement* (« dans le cas contraire »), *ces temps* (« ces temps-ci »), *entre deux* (« entre temps »[85], cf. exemple 71), *gentiment*[86] (« tranquillement, doucement », cf. exemple 72), *pire* (« très », régionalisme pour les classes d'âge les plus jeunes), etc.

(71) T'es que maintenant au lit ? T'étais où **entre deux** ? Moi j'arrive maintenant à la maison. Bonne nuit. :) Becs grand charmeur ! ;) (smsFS)

(72) Coucou ! Ça va pas mal, je commence **gentiment** à m'acclimater ! Et ça a l'air intéressant ! Et toi, ça va ? Si jamais j'ai congé mardi de midi à 2h et jeudi de 2h à 4h une semaine sur deux ! Si t'as le temps une fois, on peut aller manger ensemble ? ! Becs (smsFS)

Enfin, nous notons un grand nombre d'expressions régionales : par exemple, *ça joue* (« ça va, ça fonctionne », cf. exemple 73), *pour finir* (« ou pas », cf. note 86), *ou bien* (« ou pas ») et *c'est tout bon* (« c'est réglé », cf. exemple 74).

83. Notons que ce qui fait de *salutations* un régionalisme ne repose pas sur sa forme ni sur son référent mais sur sa fréquence. En Suisse, contrairement au reste de la francophonie, cette forme ne s'emploie pas dans des contextes formels (du type de la correspondance et du courrier électronique formels) mais de façon tout à fait informelle, par exemple comme on le voit ici, dans les sms.

84. Il s'agit, comme avec *salutations*, d'un régionalisme de fréquence puisque son sens est communément « l'argent ».

85. Notons que *entre deux* s'utilise uniquement lorsque les deux événements sont assez proches chronologiquement parlant. On ne trouvera par exemple pas : *ça fait dix ans que je l'ai plus vu.* **entre deux il a vieilli.*

86. On l'emploie également dans d'autres aires de la francophonie, mais sa pratique est plus importante en Suisse.

(73) Coucou, je passe encore vite la panosse et je vais vite me doucher pendant que ça sèche et je t'appelle **ça joue** ? T'es à la maison ? Je t'aime tellement fort ! (smsFS)

(74) Yo ça va ? J'ai réservé ce qu'on a décidé hier soir, ça m'a pris pas mal de temps mais c'est tout bon :-D je me réjouis d'y être ! J'arrive à 17h30, lac ou piscine ça te dit ? Biz (smsFS)

Les constructions à verbe support sont, dans le contexte de l'eSMS en Suisse, moins fréquentes que les expressions. On retiendra des constructions avec :
– *avoir* : *avoir mal au cou* (« avoir mal à la gorge »), *avoir trop de peine à* ;
– *être* : *ne pas être bien* (« être malade », exemple 75) ;
– *faire* : *faire la noce* (exemple 76, cf. note 86).

(75) Je suis dans le train. J'aurais bien voulu rester mais je préfère **ne pas être bien** à la maison ! ! =P bisou de microbe ;-) (smsFS)

(76) Hey ! Ouais je vais au bal ! J'espère que t'arrives à te libérer, il faut que tu sois là ! On **fera une** énorme **noce** pour la fin de ton armée ! Courage pour la fin ! (smsFS)

Deux spécificités ressortent de l'analyse thématique de ces régionalismes. Premièrement, contrairement à la Belgique, on ne retrouve pas ici le lexique d'une communauté particulière (comme c'était le cas avec le jargon des étudiants). Contrairement à La Réunion, les régionalismes rencontrés ne correspondent pas à des référents (flore, alimentation, etc.) faisant davantage sens en Suisse qu'ailleurs en francophonie. **Nous sommes en présence d'unités et d'expressions régionales du langage courant** définissant des réalités et des pratiques usuelles, qui ne sont pas étrangères au reste de la francophonie, mais pour lesquelles les Suisses ont choisi d'autres étiquettes. Deuxièmement, **l'influence de l'allemand se fait ressentir** avec des mots tels que *mensa* et des expressions telles que *avoir mal au cou* (***Hals**schmerzen haben*) et *un (petit) coup* (*einmal*, dans par exemple *komm einmal her*).

Notre deuxième axe de recherche porte sur les régionalismes les plus courants (à partir de 10 occurrences) du corpus (cf. Figure 8.39). On retrouve la forme de clôture du canal de communication, *salutations*, typique de la Suisse pour sa fréquence et l'originalité de son emploi informel. On dénombre également 5 expressions dont *ça joue*, signifiant « ça va » ou « ça marche » ; ce sens récurrent est tout à fait explicable par le type de thématiques abordées dans les sms, dont font partie la prise et l'arrangement de rendez-vous (exemple 77).

(77) Salut les gus ! ! Je vous propose qu'on se voit demain pour parler du C-D-F, vers 16h30. Est-ce que **ça joue** ? Dites-moi si c'est trop tôt ! ! J'attends de vos nouvelles ! ! (smsFS)

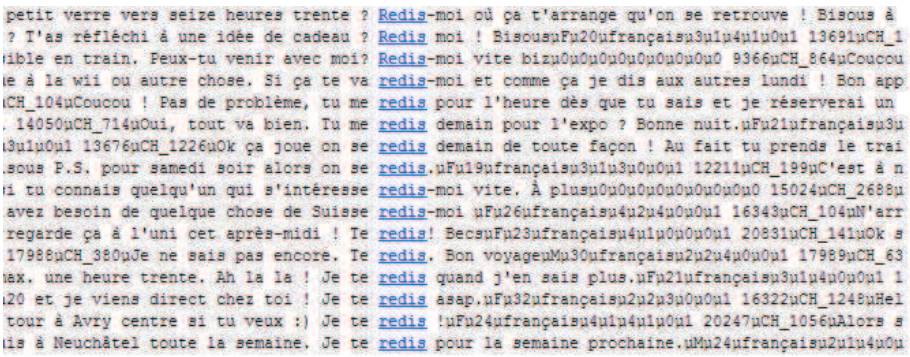
Le verbe *redire* apparait également dans cette liste avec 86 occurrences. Cet emploi récurrent mérite d'être détaillé. Nous entreprenons de décortiquer ses constructions

FIGURE 8.39 – *Helvétismes les plus fréquents dans le corpus suisse francophone*

ça joue, exp.	89
redire, V.	86
bec, N.	83
souper, N. et V.	61
natel, N.	34
c'est tout bon, exp.	19
salutations, N.	17
pour finir, exp.	12
ou bien, exp.	11
tout de bon, exp.	10

et contextes d'apparition dans les sms. L'exemple de concordance suivant livre un aperçu global de l'emploi de *redire* (Figure 8.40).

FIGURE 8.40 – *Exemple de concordance du verbe redire dans le corpus suisse*



L'analyse détaillée des 86 occurrences permet de catégoriser les constructions de ce verbe avec, par ordre de fréquence :

- *redire*, impératif, suivi d'un tiret et du pronom personnel *moi*, objet indirect : *redis-moi* ;
- *redire*, indicatif présent, séparé de son sujet par l'objet indirect *te*, *se*, *me*, qui le précède : *on se redit* ;
- *redire*, infinitif, inclus dans la formule semi-figée *merci de me*, qui le précède : *merci de me redire*.

Étant donné ces contextes d'apparition, il semble effectivement que *redire* occupe la même place en Suisse que *confirmer* dans d'autres régions.

8.3.4.2 Variation diastratique

L'approche qualitative a permis de mettre en avant la diversité des formes régionales en usage dans l'eSMS. Il est maintenant intéressant d'approfondir le sujet en examinant la quantité, la distribution et les dépendances de ces helvétismes.

Nous débutons avec un aperçu global de la quantité de sms contenant au moins un helvétisme (Figure 8.41) : 14 % des sms sont concernés par ce phénomène. Il est intéressant de noter que **l'utilisation de régionalismes dans les sms du corpus suisse est quatre fois plus importante** (en pourcentage du nombre de sms concernés) **que celle des éléments allogènes** pour la même région.

FIGURE 8.41 – *Effectifs et pourcentage des sms contenant au moins un helvétisme*

Helvétisme

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	3338	85,9	85,9	85,9
	1	550	14,1	14,1	100,0
	Total	3888	100,0	100,0	

Examinons à présent la distribution des régionalismes en fonction des variables *sexe*, *âge*, *chat* et *forum*, liées au profil des auteurs de message. Pour envisager l'étude des variables *sexe* et *helvétisme*, nous utilisons un test khi-carré. Le test, détaillé en Annexe A.80, présente une valeur p inférieure à 0,05 (p = 0,015) ce qui signifie qu'**il y a dépendance entre les variables *régionalisme* et *sexe*** dans le corpus suisse. Lorsque nous réalisons un diagnostic case par case du tableau des résidus standardisés ajustés, nous pouvons détailler cette conclusion : le **taux de femmes usant de régionalismes est plus important** que celui auquel on aurait pu s'attendre. Au contraire, le taux d'hommes est inférieur au taux escompté.

La dépendance entre *régionalisme* et *âge* peut être évaluée à l'aide d'un test t à condition que la distribution des âges soit symétrique. Nous avons vu que le corpus suisse ne permettait pas un test t pour analyser la variable *âge*, même après l'application d'un échantillonnage aux personnes de moins de 31 ans. Nous recourons donc à un test U de Mann-Whitney. Nous définissons les hypothèses de départ (H0 et H1) comme précédemment. Le détail du test U se trouve en Figure 8.42. On remarque que p = 0,371 (la valeur p est supérieure à 0,05) ; le test de Mann-Whitney nous incite donc à accepter l'hypothèse nulle (les médianes des régionalismes en fonction des âges sont égales M0 = M1). Le corpus suisse ne montre aucune tendance à utiliser plus ou moins de régionalismes en fonction de l'âge de l'auteur du message.

Une analyse des variables *forum* et *chat* clôture la description générale des données. Le test khi-carré pour la dépendance entre *helvétisme* et *chat* aboutit à une valeur p supérieure à 0,05 (p = 0,075, cf. A.81). Les variables *régionalismes* et *utilisation du chat* sont donc indépendantes dans le corpus suisse[87]. Pour ce qui est des forums, la

87. Notons donc que la tendance constatée en 2004, puis en 2008, est toujours d'actualité en

FIGURE 8.42 – *Distribution des sms comprenant au moins un helvétisme selon l'âge*

Récapitulatif du test d'hypothèse

	Hypothèse nulle	Test	Sig.	Décision
1	Les médianes de Age2 sont identiques pour les catégories de Helvétisme.	Test de la médiane à échantillons indépendants	,371	Retenir l'hypothèse nulle.
2	La distribution de Age2 est identique sur les catégories de Helvétisme.	Test U de Mann-Whitney à échantillons associés	,024	Rejeter l'hypothèse nulle.

Les significations asymptotiques sont affichées. Le niveau de signification est .05.

valeur p est supérieure à 0,05 (p = 0,594, cf. Annexe A.82) ; nous concluons donc à une indépendance des deux variables *utilisation d'un régionalisme* et *fréquentation des forums de discussion* en Suisse.

Notre deuxième axe de recherche porte sur le régionalisme le plus courant : le verbe *redire*. Son utilisation est-elle le fait d'une classe particulière ? Pour y répondre, nous allons examiner les variables *sexe*, *occupation* et *niveau d'éducation*. Le but est de comprendre si le profil des usagers associés à *redire* diffère de celui des usagers associés aux autres régionalismes. Le khi-carré représenté en Figure 8.43 est appliqué aux deux variables qualitatives *redire* et *sexe*. La valeur p est égale à 0,015 ($< 0{,}05$), ce qui signifie que les deux variables sont dépendantes. L'observation des résidus standardisés ajustés montre que le taux de femmes usant de *redire* (colonne 2) est moins important que celui que l'on aurait attendu. Au contraire, le taux d'hommes est supérieur au taux attendu. **Les hommes utiliseraient donc en moyenne davantage *redire* que les autres helvétismes**.

Afin de juger de la dépendance entre l'utilisation de *redire* et le type d'occupation, nous utilisons également un test khi-carré (cf. Annexe A.83). La valeur p trouvée est égale à 0,770 (la valeur p est supérieure à 0,05). **Les deux variables *utilisation de redire* et *type d'occupation professionnelle* sont donc indépendantes**.

Enfin, nous appliquons le test khi-carré aux variables *emploi de redire* dans les sms et *niveau d'éducation* (cf. Annexe A.84). La valeur p trouvée est égale à 0,034 (la valeur p est inférieure à 0,05). Les deux variables *emploi de redire* dans les sms et *niveau d'éducation* sont donc dépendantes. Pour détailler cette conclusion, nous observons les valeurs pour les résidus ajustés : les tests montrent que le taux d'emploi de *redire* par la population disposant d'un diplôme inférieur ou égal au diplôme de secondaire inférieur est plus important que celui qu'on aurait attendu. À l'inverse, le taux d'emploi de *redire* par la population disposant d'un diplôme d'étude post-secondaire est moins important que celui qu'on aurait attendu. Ainsi, **on peut conclure que *redire* est en moyenne davantage utilisé par les moins diplômés que par**

2009.

FIGURE 8.43 – *Test khi-carré mesurant la dépendance des variables redire et sexe en Suisse*

Tableau croisé Sexe * Helvétisme

			Helvétisme		
			1	2	Total
Sexe	F	Effectif	340	47	387
		Résidu ajusté	2,4	-2,4	
	M	Effectif	94	25	119
		Résidu ajusté	-2,4	2,4	
Total		Effectif	434	72	506

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	5,859[a]	1	,015		
Correction pour la continuité[b]	5,155	1	,023		
Rapport de vraisemblance	5,441	1	,020		
Test exact de Fisher				,024	,014
Nombre d'observations valides	506				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 16,93.

b. Calculé uniquement pour un tableau 2x2

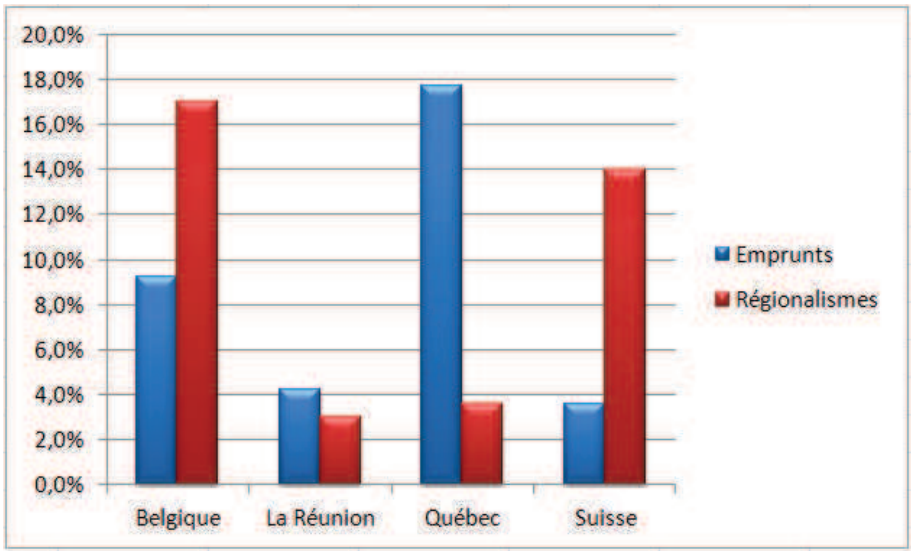
les plus diplômés.

8.3.5 Conclusions

Un des objectifs de cette partie accordée aux régionalismes lexicaux était d'étudier leur présence et leur fréquence. Cette présence est non négligeable. Nous avons relevé automatiquement (mais non exhaustivement) un pourcentage variable de messages contenant au moins un régionalisme : 17 % en Belgique, 3 % à La Réunion, 3,6 % au Québec[88] et 14 % en Suisse. Si l'on compare ces résultats avec les pourcentages d'éléments allogènes (Figure 8.44), on remarque que les éléments allogènes sont moins fréquents que les régionalismes en Belgique et en Suisse. Au Québec et à La Réunion, l'inversion de ce phénomène pourrait peut-être s'expliquer par une réalité de terrain qui se veut – au moins pour certaines régions ou villes, sinon pour un nombre important de locuteurs – bilingue, ce qui est moins le cas en Belgique et en Suisse, où règne majoritairement une diglossie (au sens de « coexistence de plusieurs systèmes linguistiques au niveau de l'état mais non de l'individu »).
D'aucuns pourraient justifier la fréquence d'emploi élevée des régionalismes par

88. Ce chiffre peu élevé est particulièrement étonnant et s'oppose aux tendances que l'on peut trouver dans la littérature, notamment dans Lamiroy *et al.* (2010). Nous ne voyons que deux explications possibles pour ce phénomène : le mauvais repérage informatique des québécismes ou la spécificité du corpus de sms par rapport aux autres types de corpus (parlé, littéraire, journalistique, blogs, etc.) sur lesquels sont basés les travaux de recherche traditionnels.

FIGURE 8.44 – *Différences de pourcentages entre taux de sms avec éléments allogènes et régionalismes*



des motivations économiques d'espace. C'est pourquoi, nous avons entrepris de comparer les régionalismes les plus courants (plus de 20 occurrences par région) à leur « variante » en français standard, en mettant en avant le nombre de caractères gagnés ou perdus. Le résultat de ce calcul se trouve en Annexe A.85. Comme nous pouvons le remarquer, un grand nombre de régionalismes participent à une économie d'espace dans le sms. Ceci ne signifie pas pour autant que le but de leur utilisation est l'économie. De plus, l'économie d'espace ne concerne pas tous les régionalismes : en effet, les formes fréquentes *dis* (marqueur de discours, optionnel), *souper*, *syllabus* et *ça joue* n'y participent pas. L'emploi de régionalismes n'est donc pas systématiquement une affaire d'économie.

Les variétés régionales du français témoignent, dans chaque aire étudiée, de particularités distinctes. Alors que le corpus belge est empreint d'un lexique en rapport avec le folklore estudiantin[89], le réunionnais utilise plutôt un vocabulaire relatif à des réalités propres à la région et pour des petits noms spécifiques. Le corpus suisse pour sa part montre des régionalismes de registre et de domaine courants, quelquefois influencés par des germanismes. Enfin, au Québec, nous avons mis en avant l'influence importante de l'anglo-américain sur les régionalismes ainsi que la prépondérance du domaine de l'alimentation.

Enfin, nous voulions savoir si l'emploi de régionalismes dépendait du profil des au-

89. Comme explicité précédemment, ceci est peut-être dû à la nature biaisée du corpus belge, dont les messages ont majoritairement été fournis par des étudiants ; cette explication est toutefois affaiblie par le fait que ce biais est justement présent dans toutes les régions (cf. Figure 6.13, p. 113) sans que les régionalismes s'y retrouvant appartiennent à cette même thématique.

teurs. Nous avons donc testé le corpus en fonction des variables *sexe*, *âge*, *utilisation du chat* et *fréquentation de forums* de discussion. La Figure 8.45 résume les résultats des tests statistiques.

FIGURE 8.45 – *Hypothèses relatives à l'emploi de régionalismes, liant les variables ou les médianes de variables dans chaque région (H1 = dépendance ; H0 = indépendance)*

	Belgique	La Réunion	Québec	Suisse
sexe	H1	H1	H0	H1
âge	H1	H1	H0	H0
chat	H0	H0	H0	H0
forum	H1	H1	H0	H0

Selon les observations basées sur corpus, il y a une propension à utiliser davantage de régionalismes lors de la pratique écrite dans les sms chez les femmes en Belgique, à La Réunion et en Suisse. Au contraire, il n'existe, dans aucune région, de dépendance entre l'utilisation d'un service de chat et l'emploi de régionalismes. Les hypothèses restantes (dépendance des variables *âge* et *forum de discussion*) varient en fonction de la région étudiée. Les pratiques diffèrent donc peut-être d'une région à l'autre ou bien la variation dans la date de collecte, la taille et la représentativité des corpus offre des données hétérogènes, qui engendrent des différences dans les résultats des tests statistiques.

8.4 Conclusions

Ce travail d'exploration lexicale des données a permis de livrer un panorama conséquent de la question linguistique dans les sms et de répondre à un certain nombre de questions de recherche plus précises.

L'analyse des données de sms a démontré de nombreuses prévalences.

1. Dans le domaine grammatical : la supériorité numérique des verbes par rapport aux substantifs ;
2. Dans la sphère structurelle : la fréquence élevée de collocations, pour les structures nominales, de type AN, NdeN et NN ;
3. Pour l'analyse thématique : le lexique des collocations se réfère largement au domaine des sentiments, des rendez-vous, des souhaits et de la technologie ;
4. Pour le domaine morphologique : la part importante de néologismes constitués par dérivation, y compris la dérivation des éléments allogènes (et une fréquence très peu élevée de créations lexicales par composition et troncation) ;
5. Pour l'analyse contextuelle et fonctionnelle : nous avons relevé une série de 7 facteurs régissant l'emploi d'éléments allogènes dans les sms francophones. Parmi ceux-ci, la contrainte d'espace due à la limite des 160 caractères, mais aussi, par exemple les connotations affectives associées à une langue.

L'anglais est particulièrement prégnant dans les sms ; les langues romanes, telles que l'italien et l'espagnol, le sont moins mais semblent répondre à des fonctions plus précises et touchant au domaine des sentiments (amour, haine, pitié, excuses, etc.). En outre, le transfert apparait comme le phénomène d'alternance de langues le plus courant dans les sms.

L'emploi des régionalismes est également très important dans les sms. Dans certains corpus régionaux, tels que les belge et suisse, le taux de régionalismes surpasse largement celui des éléments allogènes. Le gain d'espace ne régit pas toujours l'utilisation de ces formes régionales (surtout pour ce qui est des expressions figées et des marqueurs du discours, optionnels). Du point de vue des classes grammaticales, les substantifs sont les régionalismes les plus représentés dans les sms, mais les verbes sont également très présents, surtout en nombre d'occurrences. On se souviendra des exemples *savoir*, *sonner*, *diner*, *souper*, *battre*, *redire* et *adonner*. Les thématiques abordées par les régionalismes sont liées à la population sur-représentée dans le corpus (les étudiants) ou reflètent des réalités de terrain, qu'elles soient climatiques, gastronomiques ou politiques.

Grâce aux méthodes de la statistique, nous avons également vu que les différents phénomènes variaient selon la région ou le profil des auteurs des messages. Ainsi, l'approche lexicométrique a permis de déceler une différence dans la diversité lexicale que présentent les sms des jeunes et des moins jeunes. L'étude des néologismes a montré que les créations attestées dans le corpus pouvaient être le fait d'un

locuteur créatif donné (*cybernamoureutisme*), ou d'une ou plusieurs communautés (*sms*, *texto*, etc.) ; elles peuvent également constituer des invariants (*mdr*) apparents dans la plupart des aires géographiques étudiées.

Pour la quasi-totalité des phénomènes étudiés, on note une propension des femmes à employer davantage ces phénomènes linguistiques dans les sms que la normale. Dans le cas des néologismes et des éléments allogènes, on note également un sous-emploi des phénomènes par les diplômés post-secondaires. À l'inverse, les diplômés (ou des élèves encore aux études) du secondaire supérieur utilisent plus de phénomènes que ce à quoi on aurait pu s'attendre avec une distribution équilibrée. En matière de néologismes encore, on a noté une plus grande créativité chez les locuteurs non francophones que chez les francophones.

Enfin, l'étude statistique du lexique a montré une absence quasi totale d'influence de la pratique du chat, du forum de discussion et du courrier électronique sur l'emploi des phénomènes étudiés.

Chapitre 9

Variation orthographique

Dès que le vent soufflera
Je repartira
Dès que les vents tourneront
Nous nous en allerons
Renaud (1983)

Sommaire

9.1 Introduction 215
9.2 Variantes à dominantes morphogrammiques 219
9.3 Variation diastratique 224
9.4 Conclusions 227

9.1 Introduction

Comme nous l'avons vu dans le chapitre 5, depuis une dizaine d'années, l'utilisation du courrier électronique, de la messagerie instantanée et de l'esms s'est intensifiée de manière spectaculaire au point de devenir, au sein des tranches d'âge les plus jeunes, la pratique principale – et soutenue – de l'écrit. Or, une des propriétés essentielles de cette CéMO est son altération de la graphie. Pour répondre à des besoins d'économie de cout, d'espace et de temps, les usagers de la CéMO montrent une tendance manifeste à l'abréviation. Cette tendance suscite autant l'excitation et l'engouement, que la crainte et le scepticisme. C'est pourquoi il est nécessaire d'obtenir des données empiriques qui viendront confirmer ou infirmer les opinions variées.

L'étude des variations orthographiques dans les sms n'est rendue possible que dans un seul des quatre corpus dont nous disposons. En effet, notre but étant d'évaluer de manière quantitative les variations par rapport à l'orthographe conventionnelle, ce travail est basé sur une comparaison des deux corpus brut et transcrit, alignés au caractère près. Comme présenté dans la partie 6.1.4.2, seul le corpus belge est aligné au niveau du caractère. Les résultats de cette partie orthographique seront donc uniquement valables pour les données belges, ceci expliquant la brièveté de la partie par rapport aux parties dédiées au lexique et à la syntaxe.

Nous partons du principe (énoncé dans la partie 5.1.1.1, p. 20), qu'il est dangereux d'apposer l'étiquette de « variante orthographique inconsciente » là où un phénomène peut se confondre avec une stratégie de communication abrégeante typique de l'eSMS. Le panorama des tendances graphiques dans les sms (partie 5.2.2) a permis de montrer qu'il n'existait pas réellement de « cacophonie orthographique », mais bien des règles propres à la CéMO, définissables et identifiables, de sorte que l'on peut lire un sms tel que celui de l'exemple 78 avec plus de précautions.

(78) Salut [NPERS]. Je n'ai toujours pas recu ta lettre... Ce n'est pas normal ! Es ce que tu m'as envoyer la lettre ? ? (smsFB)

Nous ne pouvons pas interpréter avec certitude les raisons de l'emploi des formes *recu* sans cédille et *es ce* au lieu de *est-ce*, puisque cet emploi entraine un abrègement ou une simplification du message et peut donc participer à une stratégie d'abrègement consciente de la part du locuteur. Par contre, il est possible d'interpréter *envoyer* comme variante orthographique inconsciente, étant donné que cette forme n'entre dans aucune des catégories de stratégies propres à l'eSMS. Suite au constat de phénomènes tels qu'illustrés par l'exemple 78, pour lesquels la variation graphique n'abrège pas le message et ne simplifie pas l'encodage de caractères, nous avons précédemment entrepris d'examiner la présence d'écarts orthographiques non abrégeant, dans l'écrit sms (Cougnon *et al.*, 2012). Nous approfondissons cette recherche initiale.

La méthodologie d'identification des variantes orthographiques consiste en plusieurs points. Tout d'abord, nous excluons toutes les formes répondant aux stratégies linguistiques de l'eSMS présentées précédemment : nous gardons uniquement les formes dont la version brute présente un nombre de caractères égal ou supérieur à la forme standardisée. Lorsqu'il y a hésitation sur la nature de la variation (stratégie ou écart orthographique), nous faisons le choix méthodologique d'exclure la forme [1]. Parmi les formes brutes de taille identique ou supérieure, nous écartons également une série de phénomènes :

- les variations d'accentuation, étant donné qu'elles peuvent participer aux stratégies de simplification rédactionnelles (chute massive de l'accentuation) ;
- les formes plus longues qui répondent à une fonction expressive de l'eSMS (*oooh* pour *oh*) ;
- les variantes présumées liées à l'utilisation du clavier (confusion entre les lettres correspondant à une même touche sur le clavier traditionnel d'un téléphone portable ; par ex. *hare* pour *gare*) ;
- les variantes présumées liées au système de prédiction [2] (cf. section 5.2.1).

Ces choix méthodologiques engendrent des précautions que nous devons signaler

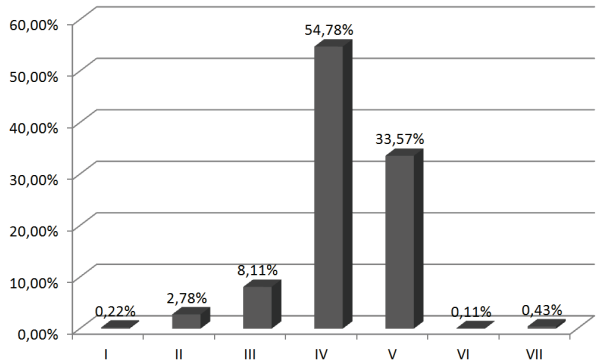
1. Il est, dans la majeure partie de ces cas, impossible de trancher quant à la nature de la variation.

2. On évite donc d'identifier comme variantes orthographiques inconscientes des variations de type *T amenés la twingo vers 19h30...* (smsFB) où *amenés* peut être interprété comme une proposition du T9 lorsque l'auteur du message voulait encoder *amène* ou *amènes*.

avant la lecture même des résultats : d'abord, il ne s'agira pas ici de connaitre le taux général de variantes orthographiques par rapport au taux général de formes correctement orthographiées mais de présenter la distribution des types de variantes au sein de l'ensemble des variantes qui ne répondent pas à des stratégies de simplification. Ensuite, le rejet de certaines formes, comme l'absence d'accent, entraine un biais que nous pensons important dans nos résultats, tel que la confusion a/à (code 420, cat. V), certainement la plus récurrente dans l'ensemble des homophones grammaticaux ; c'est également en raison du rejet de l'accent que le code 2000 (omission ou adjonction de voyelle, de digramme vocalique et d'accent, cat. II) voit son nombre d'occurrences terriblement baisser. Pourtant ce choix de rejet se justifie : dans un très grand nombre de cas, la chute de l'accent a effectivement lieu pour des raisons de facilité d'encodage.

Nous choisissons de présenter la typologie des variantes orthographiques inconscientes en reprenant la classification de Catach (1980) :
- I. Erreurs à dominante phonétique : *descentre* pour *descendre* ;
- II. Erreurs à dominante phonogrammique altérant la valeur phonique : *creuvée* pour *crevée* ;
- III. Erreurs à dominante phonogrammique n'altérant pas la valeur phonique : *commance* pour *commence* ;
- IV. Erreurs à dominante morphogrammique : *mes parents ont encore décidés/mes parents ont encore décidé*
- V. Erreurs concernant les homophones : *court/cours* ;
- VI. Erreurs concernant les idéogrammes [3] : *d'orénavant/dorénavant* ;
- VII. Erreurs concernant les lettres non fonctionnelles : *cauchemard/cauchemar.*

FIGURE 9.1 – *Distribution des variantes orthographiques de l'écrit sms selon leur type*



La distribution des variantes (Figure 9.1) montre une nette prédominance des variantes de nature morphogrammique (cat. IV, 54,78%, surtout les morphèmes grammaticaux) et les variantes dues à des homophones (cat. V, 33,57%). On note très peu de variantes à dominante phonétique (cat. I, 0,22%), de variantes concer-

3. Au sens de Catach (1980, p. 15).

nant des idéogrammes (cat. VI, 0,11%) et concernant des lettres non fonctionnelles (cat. VII, 0,43%).

La typologie détaillée est présentée dans les figures A.86 et A.87 (en annexe) et 9.2, p. 219. Chaque catégorie est suivie de son nombre d'occurrences, entre parenthèses, et est illustrée d'exemples non limitatifs. Quelques remarques générales concernant la constitution de cette typologie sont nécessaires.

Le tableau originel proposé par Catach (1980) n'a pas pu être conservé tel quel, notamment parce qu'il nous parait quelquefois y avoir confusion entre l'intention du locuteur (ex. code 313 : ignorance du maintien ou non du radical, cat. IV) et le résultat graphique (ex. code 210 : omission de voyelle, cat. III). Ce genre de confusion a rendu l'attribution des catégories difficile. Ensuite, même si le tableau était uniquement basé sur les intentions des locuteurs, il serait très difficile d'interpréter ces intentions. Prenons un de nos exemples : le substantif *chien* écrit *chient*. Le locuteur a-t-il souffert d'une ignorance des lettres finales justifiables (code 314, cat. IV) ou a-t-il simplement confondu ponctuellement les digrammes vocaliques (code 2110, cat. III) ? Bien sûr, Catach (1980) propose de croiser les critères et d'apposer plusieurs étiquettes à une même variante. Mais cela n'aurait-il pas été plus simple de créer une grille d'analyse uniquement basée sur le résultat graphique, de sorte que chaque variante n'appartienne qu'à une seule catégorie et ne soit pas sujette à une interprétation subjective ? Dans le cas de l'analyse des sms, cette approche nous paraitrait en tout cas plus appropriée.

Certaines catégories proposées par Catach (1980) ne sont pas représentées dans notre corpus d'écrit sms. Plusieurs explications se profilent :
- certaines catégories semblent redondantes, par ex. les codes 3004 et 3002 (cat. IV) pour l'accord du participe passé et du nom sans auxiliaire. Pourquoi le code 3004 ne représente-t-il pas la catégorie des participes passés en général ? Face à ce questionnement, nous n'avons rien étiqueté à l'aide de ce code ;
- quelques catégories ont été supprimées pour contourner la subjectivité. Ainsi le code 312 (cat. IV) : ignorance des pré- et suffixes, qui pouvait à chaque fois être interprété comme un code 2110 (confusion digramme vocalique, cat. III) ;
- certaines catégories concernent clairement des variations abrégeant/simplifiant la forme standard (ex. code 610 : omission des consonnes grecques, cat. VII) et ont donc été éliminées de l'étude avant même d'entamer l'analyse selon la grille de Catach (1980) ;
- certaines catégories participent du mouvement plus général de chute massive de l'accentuation et de mutation fonctionnelle de la casse (ex. code 424 : *ou/où*, cat. V) et ont donc été, comme au point précédent, éloignées de l'analyse dès ses prémices ;
- nous ne traitons pas des variations de ponctuation (503 et variantes) ni de casse (500 et variantes) car l'eSMS ne répond pas à l'usage standard de la ponctuation et de la casse.

FIGURE 9.2 – *Typologie des variantes orthographiques 2 (inspirée par Catach, 1980)*

IV. *Erreurs à dominante morphogrammique* (1519)		
Morphèmes grammaticaux (1470)		
Relation mal établie entre :		
3000 les catégories grammaticales (4) *bienvenue/bienvenues* (22) *Bienvenues à vous 3*	**3002 adjectif et nom** (32) *beau/beaux* (23) *Merci Lol il est trop beaux*	
3001 les formes du pluriel (7) *bleus/bleux* (24) *tu as les cheveux bleux*	**3003 sujet-verbe** (865) *aimerais/aimerait* (25) *jaimerait justement savoir!*	**30040 avec *être*** (24) *né/née* (26) *Le petit {Nom, masc} est née ce 22.09.04*
3010 nom et déterminant (29) *classe/classes* (27) *c toute des bombes dan÷ ta classes*	**30120 infinitif/participe** (204) *assumé/assumer* (28) *Heyyy catiiiiin! T'as pas assumer toi aujourd'hui!* **30121 infinitif/indicatif** (40) *embêtez/embêter* (29) *Si vous vous embeter vous pouvez nous rejoindre à la doud*	**30041 avec *avoir* sans C.O.D** (27) *décidé/décidés* (31) *Mes parents ont encore décidés de m'em...*
	30122 indicatif/conditionnel (112) *demanderai/demanderais* (30) *ok ma mimi que j'aime je lui demanderais*	**3013 Groupe-désinence** (46) *agir/agire* (32) *Je peux agire légalement*
3011 nom, adj + cplmt (2) *bonheur/bonheurs* (33) *je vs souhaite bcp d bonheurs*	**30123 autres** (77) **indicatif/impératif** (34) *amuses-toi bien ce soir* **indicatif/subjonctif** (35) *c ce que je t ai dis hier je n aie po les notes* **indicatif/participe** (36) *je suis degoutai*	**3014 Temps** (1) *dirai/disais* (37) *il voudrait que vous soyez là... Je te disais quel week end*
Morphèmes lexicaux (49)		
	313 Ignorance du maintien ou non du radical (1) *fatigant/ fatiguant* (38) *tu crois que je vais monter pour te dire bonjour? Trop fatiguant ;o)*	**314 Ignorance des lettres finales justifiables** (48) *fois/foix* (39) *on les regarderais une foix a 2 d ak ?*

9.2 Variantes à dominantes morphogrammiques

Ce sont les variantes morphogrammiques touchant aux morphèmes grammaticaux, la catégorie la plus récurrente (1470 occurrences), que nous allons détailler dans cette partie (cf. Figure 9.2). Certains types de variantes apparaissent peu dans les sms ; c'est le cas de la confusion des catégories grammaticales, de la mauvaise maitrise des formes du pluriel et de l'accord nom/adjectif et son complément. D'autres variantes orthographiques sont plus fréquentes : par exemple l'accord du nom avec son déterminant, l'accord du nom et de son adjectif, l'accord du participe passé avec *être* et avec *avoir* (sans complément d'objet direct), ainsi que la confusion dans les désinences propres aux groupes de verbes. Enfin, on note un nombre d'occurrences beaucoup plus important de deux types de variantes : l'accord sujet-verbe (865 occurrences) et la confusion des modes verbaux [4] (433 occurrences).

4. Il est important de préciser ce que nous entendons par *confusion des modes verbaux*. En effet, nous reprenons dans cet ensemble les sous-catégories de Catach (1980), ceci incluant la confusion

Le marquage de l'accord grammatical dans les sms francophones fait l'objet de peu d'études, à l'exception de Stark (2011) qui admet elle-même : « Dans le contexte de la recherche relative à la langue ou la graphie dans les nouveaux média, [ce sujet] s'avère [...] négligé [...], et cela en dépit du fait que le grand public se montre souvent préoccupé par la menace potentielle que les nouveaux média pourraient représenter pour la maitrise de la 'bonne écriture' ». Certains travaux abordent des questions liées à la morphosyntaxe dans les sms mais sans que celles-ci en constituent la thématique centrale (Fairon *et al.*, 2006b ; Sourdot, 2004). Stark (2011) se base sur un échantillon de 400 messages du corpus suisse identique au nôtre pour classer, manuellement, les variations dans l'accord sujet-verbe pour 6 classes de marquage graphique différent. Bien que nous nous intéressions également à l'accord sujet-verbe, nous n'allons pas suivre cette méthodologie car nous ne nous focalisons pas sur les mêmes manifestations de ce phénomène [5].

Le verbe le plus concerné par les variations dans l'accord sujet-verbe est le verbe *faire*, suivi de *dire*, de *être* et de *venir* (Annexe A.88). Le groupe verbal le plus touché par ces variantes est le troisième groupe (78,77% des cas, cf. Annexe A.89). Si l'on analyse de plus près les dix types de variation les plus fréquents dans l'accord sujet-verbe, sans contexte [6] (Figure 9.3), on remarque qu'elles concernent les verbes *faire*, *dire*, *devoir*, *dormir* et *avoir*. Il est intéressant de constater que cette liste n'est pas identique à celle des verbes les plus fréquents du corpus (cf. liste élaborée dans la partie 8.1.1 où *devoir* et *dormir* n'apparaissent pas). La Figure 9.3 présente la répartition, à parts égales, des confusions $1^{er}/2^{e}$ – 3^{e} p. sg et 3^{e} – $1^{er}/2^{e}$ p. sg.

FIGURE 9.3 – *Les 10 variantes les plus courantes dans l'accord sujet-verbe conjugué (corpus de sms belge, par ordre décroissant)*

Occurrence	Forme rectifiée	Type d'erreur (personne)
fais	fait	1/2 – 3 sg
fait	fais	3 – 1/2 sg
dit	dis	3 – 1/2 sg
dis	dit	1/2 – 3 sg
dois	doit	1/2 – 3 sg
dors	dort	1/2 – 3 sg
avais	avait	1/2 – 3 sg
doit	dois	3 – 1/2 sg
avait	avais	3 – 1/2 sg
ferait	ferais	3 – 1/2 sg

infinitif/participe, infinitif/indicatif, indicatif/impératif, indicatif/subjonctif, indicatif/participe et indicatif/conditionnel. Nous sommes pourtant tout à fait consciente que cette dernière confusion est en réalité une confusion de temps (deux temps de l'indicatif) et non de modes.

5. Pour rappel, nous ne travaillons pas sur les variantes abrégeantes

6. C'est-à-dire sans préciser la personne exacte (1ère ou 2e) de la confusion, que le contexte pourrait désambigüiser.

Nous nous penchons à présent sur la variation pour un accord plus spécifique : la confusion des finales *e/es* (1e/3e - 2e sg). Si l'on observe ce phénomène en contexte, on remarque que c'est dans l'emploi de l'expression *t'inquiète*, style familier et abrégé pour *ne t'inquiète pas*, que l'écart survient quasi systématiquement. De manière générale, ce sont **les verbes du 1er groupe** (88,33%) qui sont les plus touchés par cette variation. On retiendra encore que la méconnaissance de la forme survient plus fréquemment lors de l'usage de **l'impératif** deuxième personne du singulier (56%). Dans ces cas d'usage à l'impératif, il ne s'agit donc pas d'une variation inconsciente de l'accord sujet-verbe mais d'une confusion des modes verbaux (voir ci-après). Dans 44% des cas restants, on constate que la variation de l'accord à l'indicatif survient le plus souvent avec des sujets pronominaux clitiques. Pour 68% des formes attestées à l'indicatif, la variation se produit dans une construction sur base d'un sujet pronom personnel à la 1^e personne, omis (cf. 79) ou attesté (cf. 80), avec un ou plusieurs pronom(s) objet(s) insérés entre le sujet et le verbe et un accord du verbe à la 2^e personne.

(79) Cmt vas tu ? Te **sonnes**$_{2.sg}$ des ke j ai du temps pr papoter. Et le boulot cmt ca s arrange ? Biz tendre gaelle (smsFB)

(80) Coucou ptit gars !vu qu'on s'envoie tjs des messages qd on à bu un verre,**je**$_{1.sg}$ t'en **envoies**$_{2.sg}$ un maintenant..comment ça se passe pr toi ?pq tu ne donnes + de nvelle ! ? (smsFB)

On relève quelques exemples, plus rares, de variation dans l'accord avec un pronom démonstratif ou indéfini sujet (3^e p. sg) :

(81) L.O. [NPERS] ! C'est assez chouet l'ID ke tu viennes 1 lundi. [NPERS] trouve ca sympa... Donc, si t'as 1 lundi en vue, et si **ca**$_{3.sg}$ **te tentes**$_{2.sg}$, fais moi signe, sympa de te montrer la ville et boire 1 pot ! BISOUS ! (smsFB)

Pas moins de 16% des cas de variation dans l'accord sujet-verbe à l'indicatif présentent un sujet lexical [7] :

(82) Pq **la directrice**$_{3.sg}$ te **donnes**$_{2.sg}$ pas l autorisation ? Julien a pas de piano chez lui pr le moment... Tu peux, peut-etre faire intervenir ta mère pdt qu elle est là... (smsFB)

La confusion des modes verbaux retient également toute notre attention, en raison de la fréquence de ce phénomène, mais également de sa disparité : on remarque la confusion, comme nous venons de le voir avec la forme *t'inquiète*, de l'impératif et de l'indicatif à la 2^e personne du singulier, mais aussi de l'infinitif et du participe, de l'indicatif et du participe et de l'infinitif et de l'indicatif, entre autres. Ce sont les confusions infinitif/participe passé et indicatif/participe qui sont les plus fréquentes.

7. Notons que ce qui semble expliquer la variation, dans l'exemple présent, c'est la présence d'un complément objet sous la forme *te* qui précède la forme verbale et qui engendre ainsi une forme d'accord de proximité avec une forme 2^e sg qui n'est pas le sujet.

Nous approfondissons la question de la confusion des modes infinitif/participe passé (sous-catégorie la plus importante de cet ensemble). Tout d'abord, la confusion du participe passé en *-é* et de l'infinitif en *-er* : les verbes les plus touchés par cette confusion sont *oublier*, *passer*, *parler*, *trouver*, *demander* et *rentrer*. Nous constatons qu'à l'exception de *passer*, il ne s'agit pas des verbes les plus fréquents du corpus (cf. liste de la partie 8.1.1, p. 136). Un comptage permet de montrer que 6,5% des sujets concernés sont implicites. Parmi les sujets explicites, 13% sont des sujets lexicaux (exemple 83). L'emploi des auxiliaires dans une construction (avec variation) du participe passé se fait à 75% avec *avoir* et 25% avec *être*.

(83) Ce soir apres 8h paske la jsui$_{Vaux}$ aller$_{Vinf}$ voir 1 ex au foot mai le modem **est**$_{Vaux}$ **reparer**$_{Vinf}$ kiss jtad (smsFB)

On cherche à savoir si l'emploi d'un auxiliaire lié à un type de sujet particulier engendre davantage de variation. On réalise un test khi-carré pour évaluer le lien entre les deux variables qualitatives *sujet* et *auxiliaire*. Le test montre une valeur p = 0,001 ; la valeur p étant inférieure à 0,05, on rejette l'hypothèse nulle qui voudrait qu'il n'y ait pas de lien entre les variables. En se penchant sur le détail des résidus standardisés ajustés (Figure 9.4), on conclut que le taux de pronoms démonstratifs *ce/c'/ça* apparaissant avec l'auxiliaire *être* est supérieur à celui auquel on aurait pu s'attendre. Ceci se justifie probablement par le suremploi des constructions *c'est* et *ce n'est* + V$_{pp}$ comme dans l'exemple 84.

(84) Non **c**$_{VauxPhon}$ **réparer**$_{Vinf}$ bisous partout (smsFB)

FIGURE 9.4 – *Résidus standardisés ajustés liés au test khi-carré pour les variables sujet et auxiliaire dans une construction au participe passé engendrant de la variation*

Tableau croisé sujet * auxiliaire

			aux		
			avoir	être	Total
sujet	autre	Effectif	8	7	15
		Résidu ajusté	-2,0	2,0	
	ce	Effectif	2	6	8
		Résidu ajusté	-3,3	3,3	
	je	Effectif	73	13	86
		Résidu ajusté	2,9	-2,9	
	lexical	Effectif	16	7	23
		Résidu ajusté	-,6	,6	
	tu	Effectif	38	13	51
		Résidu ajusté	-,1	,1	
Total		Effectif	137	46	183

On remarque également que **la proportion du pronom *je* apparaissant avec l'auxiliaire *avoir* est supérieure à celle à laquelle on aurait pu s'attendre** dans une situation de confusion participe passé/infinitif.

Nous nous intéressons à présent à une catégorie importante en termes d'occurrences : il s'agit de la confusion de temps, et en particulier la confusion futur/conditionnel à la première personne du singulier. Cette confusion concerne majoritairement les verbes *savoir*, *être*, *aller*, *avoir* et *passer*. On remarque qu'il s'agit bien des verbes les plus fréquents du corpus ; on pourra donc se demander si cette fréquence supérieure ne peut pas être expliquée par la fort taux d'occurrences de ces verbes dans le corpus.

Du point de vue des thématiques, c'est l'arrangement de rendez-vous qui est le plus marqué par cette confusion de temps :
- (ne pas) être + lieu ou temps : *je serais chez toi - serais sans doute pas là ce we* ;
- (ne pas) passer + lieu ou temps : *OK ben je passerais dans 20 min cas maintenant je termine mes math.*
- (ne pas) aller + lieu ou temps : *Je suis au cours d [NPERS] je m irais pas au cortège - Coucou ce soir j irais certainement vers 21h30 22h. Et toi ?* ;
- (ne pas) pouvoir + événement : *Je ne pourrais etre chez vous demain à 18H - Je crois que j'ai de la fièvre mon ange. Je pourrais pas te rejoindre pour manger* ;
- (ne pas) venir + lieu ou temps : *Je viendrais a l'anniversaire de [NPERS], Je viendrais pas à l'école aujourd'hui... Panne d'oreiller* ;
- (ne pas) savoir + événement : *Je saurais pas y allé aujourd hui - Dsl jsaurais pas vnir sur msn pcq g oublié 2 tdire ke jfaisais du baby-sitting* ;
- autres constructions liées aux rendez-vous : *rentrer tard*, *tenir au courant*, *amener*, *apporter* + temps ou lieu : *Je rentrerais plus tar ke prevu, je ne vois la prof ka 3h30 - Salut non pas Pot ce soir jai exam demain...probablement la semaine prochaine mais je te tiendrais au courant ! - Est-ce que tu peux garder [NPERS] mercredi après-midi, je l'amènerais vers 13h30.*

Le tableau présenté dans la Figure 9.5 synthétise les données de cette partie.

FIGURE 9.5 – *Analyse détaillée de 4 types de variantes morphosyntaxiques dans les sms*

Confusion	Verbes principaux concernés	Nb erreurs	Nb corrects	% erreurs
futur/conditionnel 1 sg	savoir, être, aller, avoir, passer	116	2009	5,50%
participe passé/infinitif	oublier, passer, parler, trouver, demander, rentre	247	4249	5,50%
impératif/indicatif présent 2 sg	inquiéter, passer	60	± 2139	2,70%
indicatif présent 1-3/2 sg	-	21	± 7965	0,26%

Cette Figure reprend les informations principales concernant les variantes que nous avons détaillées : les verbes les plus souvent concernés et le nombre d'occurrences pour les variantes rencontrées. Nous avons également calculé le nombre d'occurrences pour chaque forme correctement orthographiée dans le corpus [8] ; en d'autres

8. Ce travail de longue haleine n'aurait pu être mené à bien sans la précieuse collaboration

termes, la colonne *Nb corrects* reprend le nombre de verbes à l'indicatif futur de la première personne du singulier, au participe passé, à l'impératif première personne du singulier et à l'indicatif présent première personne du singulier correctement orthographiés dans le corpus de sms belge.

Le faible pourcentage des variantes, allant de 0,26 à 5,5% du total des formes concernées, nous amène à penser que **les variantes orthographiques** (accords sujet-verbe, temps et modes non abrégeant) **ne constituent pas un phénomène aussi alarmant - en termes de fréquence - qu'on aurait pu le croire à la simple observation des messages**. Le très faible taux de variantes dans l'accord sujet-verbe des 1^e et 2^e p. du singulier de l'indicatif présent n'étonne pas grandement ; la confusion de mode participe passé/infinitif parait, elle, instinctivement plus commune. Les chiffres obtenus pour le corpus permettent de relativiser ces idées reçues. Ces résultats corroborent donc également les conclusions émises par ailleurs au sujet du marquage morphosyntaxique dans les sms par Stark (2011, p. 46) : « le noyau dur de ce que l'on appelle *syntaxe* est inaccessible à la conscience des auteurs/locuteurs et reste stable, même dans son marquage graphique ».

9.3 Variation diastratique

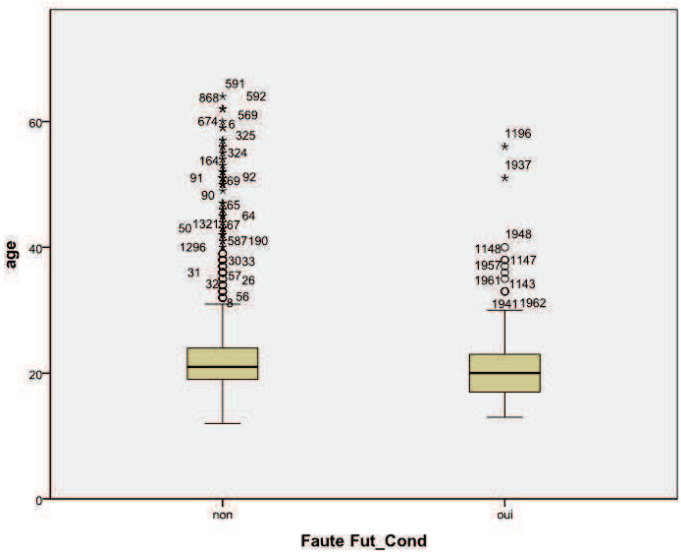
Étant donné les limites de notre méthode d'évaluation du taux de variantes orthographiques (cf. note 8), nous proposons ici une approche diastratique pour les confusions futur/conditionnel et participe passé/infinitif uniquement. Les variables étudiées pour les confusions de temps et de modes sont les suivantes : âge, sexe, niveau d'études et utilisation des différents types de Cémo.

Nous comparons les profils des auteurs ayant réalisé la confusion des temps indicatif futur/conditionnel à ceux des auteurs n'ayant pas réalisé cette confusion. La Figure 9.6 présente la répartition des occurrences de variantes en fonction de l'âge ; cette répartition semble montrer une moins grande dispersion des auteurs de messages réalisant la confusion des temps que ceux ne la réalisant pas.
Le test t appliqué sur les catégories d'âge[9] indique qu'il semble y avoir une dépendance (la valeur p est inférieure à 0,05) entre les variables *âge* et *confusion des temps indicatif futur et conditionnel* (Annexe A.92). On peut donc dire qu'il existe une répartition des âges plus étendue pour les messages des auteurs ne confondant pas les temps que pour ceux les confondant (les valeurs sont plus ramassées). Les

de Sophie Roekhaut (pour le traitement automatique) et Bernadette Dehottay (pour le travail manuel de validation des formes) que nous remercions sincèrement. Notons que seules les non-confusions (Nb corrects) futur/conditionnel et participe passé/infinitif présentent des chiffres précis pour les formes correctes ; les valeurs (Nb corrects) des deux autres catégories (impératif/indicatif et indicatif présent 1-3/2 sg) ont été calculées en triplant les chiffres obtenus après validation d'un tiers des résultats sur formes correctes.

9. Pour rappel, ces catégories sont : 1 (10 à 15), 2 (16 à 17), 3 (18 à 19), 4 (20 à 22), 5 (23 à 27), 6 (28 à 40) et 7 (41 à 70). L'histogramme des valeurs regroupées se trouve en Annexe A.91.

FIGURE 9.6 – *Box-plot présentant la distribution des confusions (ou non-confusions) des temps futur et conditionnel en fonction des classes d'âge*



médianes présentées dans les box-plots semblent également indiquer davantage de messages (avec confusion des temps) provenant d'auteurs plus jeunes (moins de 19 ans) [10].

Nous voulons maintenant savoir s'il existe une influence du sexe des auteurs sur la confusion des temps indicatif futur et conditionnel dans notre corpus. Pour ce faire, nous exécutons un test khi-carré sur les deux variables. Les résultats de ce test (Annexe A.93, la valeur p est égale à 0,152) indiquent qu'**il ne semble pas y avoir d'influence du sexe d'un auteur sur la confusion de ces deux temps**. Pour tester la dépendance des variables *niveaux d'études* et *confusion des temps indicatif futur et conditionnel*, nous regroupons les niveaux d'études en 3 catégories (pour répondre aux conditions d'effectifs minimum des tests statistiques) : 2 (primaire et secondaire inférieur), 3 (secondaire supérieur) et 4 (post-secondaire). Le test khi-carré (Annexe A.94) indique qu'il semble y avoir dépendance entre les deux variables. L'analyse des résidus standardisés met en avant la tendance suivante : **le taux de diplômés ou étudiants du post-secondaire confondant les deux temps est inférieur à ce à quoi on aurait pu s'attendre** si la distribution était équilibrée.

Enfin, nous cherchons à connaitre le possible impact de la pratique du chat et du courrier électronique, ainsi que la fréquentation des forums de discussion sur la confusion des temps. Le détail des tests statistiques se trouvent en Annexes A.95, A.96 et A.97. Le résultat de ces tests est résumé dans la Figure 9.7. Cette Figure

10. En raison du biais dans la répartition des âges de notre échantillon, nous pourrions également penser que cette tendance est seulement le résultat de ce biais.

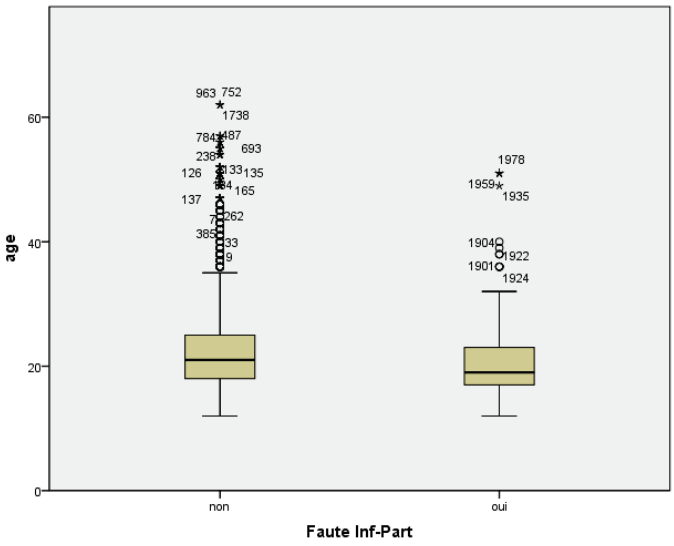
met en avant l'unique dépendance entre les variables *confusion des temps indicatif futur et conditionnel* et *pratique du chat*. Le sens de cette confusion est toutefois intéressant puisque, selon le test statistique, le taux d'utilisateurs du chat confondant les temps est inférieur à ce que l'on aurait escompté si la distribution était harmonieuse. Nous ne concluons pas à l'influence positive de la pratique du chat sur la confusion des temps, puisque la seule conclusion possible est d'affirmer la dépendance des variables ; cependant, **nous concluons** potentiellement au moins, selon les données du corpus de sms belge, **à une absence d'influence négative de la pratique du chat sur la confusion des temps**.

FIGURE 9.7 – *Résultats des test khi-carré pour le calcul de dépendance des variables confusion des temps indicatif futur et conditionnel et pratiques de la Cémo*

	valeur p du test khi-carré	dépendance variables	sens de la dépendance
chat	0,006	oui	utilisateurs chat = sous-emploi confusion
courriel	0,759	non	-
forum	0,935	non	-

Nous comparons à présent les profils des auteurs ayant réalisé la confusion participe passé/infinitif à ceux des auteurs n'ayant pas fait cette confusion lors de l'emploi du participe. La Figure 9.8 illustre la répartition des occurrences de variantes en fonction de l'âge ; cette répartition semble montrer une moins grande dispersion des auteurs de messages réalisant la confusion des modes que ceux ne la réalisant pas.

FIGURE 9.8 – *Box-plot présentant la distribution des confusions (ou non-confusions) de modes participe passé et infinitif en fonction des classes d'âge*



Le test t appliqué sur les catégories d'âge[11] indique qu'il semble y avoir une dépendance (la valeur p est égale à 0) entre les variables *âge* et *confusion des modes participe/infinitif* (Annexe A.99). On peut donc dire qu'il existe une répartition des âges plus étendue pour les messages des auteurs ne confondant pas les modes que pour ceux les confondant (les valeurs sont plus ramassées). Les médianes présentées dans les box-plots semblent (cf. note de bas de page 10, p. 225) également indiquer davantage de messages avec confusion des modes provenant d'auteurs plus jeunes (moins de 21 ans).

À présent, nous désirons calculer la dépendance des variables *sexe* et *confusion des modes* (participe et infinitif) dans le corpus. Pour ce faire, nous exécutons un test khi-carré sur les deux variables. Les résultats de ce test (Annexe A.100, la valeur p est égale à 0,003) montrent qu'**il existe une dépendance entre la variable *sexe* (d'un auteur de sms) et la confusion de ces deux modes**. En observant le détail des résidus standardisés (Annexe A.100), il nous est possible de conclure à la **sous-représentativité des hommes ne commettant pas la confusion des modes** et à une **sous-représentativité des femmes employant la variante orthographique**.

Pour tester la dépendance des variables *niveau d'études* et *confusion des modes indicatif futur et conditionnel*, nous regroupons une fois de plus les niveaux d'études en 3 catégories. Le test khi-carré (Annexe A.101) indique qu'il semble y avoir dépendance entre les deux variables (la valeur p est égale à 0). L'analyse des résidus standardisés met en avant la tendance suivante : **le taux de diplômés ou étudiants du post-secondaire confondant les modes participe passé et infinitif est inférieur à ce à quoi on aurait pu s'attendre** si la distribution était équilibrée.

Enfin, nous cherchons à connaitre le possible impact de la pratique du chat et du courrier électronique, ainsi que la fréquentation des forums de discussion sur la confusion des modes. Le détail des tests statistiques se trouve en Annexes A.102, A.103 et A.104. Le résultat de ces tests indique qu'il ne semble pas y avoir de dépendance entre ces variables.

9.4 Conclusions

Cette partie concernant les écarts orthographiques a mis en avant plusieurs tendances qui viennent confirmer et infirmer certaines idées reçues au sujet de l'orthographe dans la Cémo et qui sont valables pour le corpus belge (le seul à être transcrit au caractère près).

Nous avons tout d'abord présenté un panorama des types de variantes non

11. L'histogramme des valeurs regroupées se trouve en Annexe A.98.

abrégeantes que nous avons repérées semi-automatiquement. La typologie générale des écarts, inspirée de Catach (1980), montre une nette prédominance des écarts de nature morphogrammique (54,78%), principalement les morphèmes grammaticaux, et des écarts dus à des homophones (33,57%). Au contraire, on note un faible taux d'écarts à dominante phonétique (0,22%), d'écarts concernant des idéogrammes (0,11%) et d'écarts concernant des lettres non fonctionnelles (0,43%).

La classe la plus importante en termes de fréquence (écarts dans les morphèmes grammaticaux) a été détaillée : les variations dans l'accord sujet-verbe et la confusion des modes constituent les deux sous-ensembles les plus importants. Les verbes les plus concernés par les variations dans l'accord sujet-verbe sont les verbes *faire*, *dire*, *être* et *venir*. Dans la large majorité des formes attestées à l'indicatif, la variante se produit dans une construction sur base d'un sujet pronom personnel à la 1^e personne, omis ou attesté, un pronom objet entre le sujet et le verbe et un accord du verbe à la 2^e personne. Les confusions de modes les plus courantes sont la confusion indicatif présent/impératif (2^e p. sg) et participe passé/infinitif. La confusion de temps la plus récurrente est la confusion indicatif futur/conditionnel (1^e p. sg).

Nous avons ensuite entrepris de répertorier, dans notre corpus, pour l'ensemble des quatre confusions détaillées, les formes qui apparaissent correctement orthographiées dans le corpus brut. Ce comptage a permis de mettre en avant la nette infériorité (allant de 0,26% à 5,5%), dans le corpus de sms, des variantes par rapport aux formes correctement orthographiées.

Enfin, l'étude de la variation diastratique sur base de deux confusions orthographiques particulières a également abouti à des conclusions précises. Premièrement, la confusion des temps indicatif futur/conditionnel a montré une dépendance aux variables *âge* et *niveau d'études* de l'auteur ainsi qu'à la pratique du chat, qui ne semble pas, contrairement aux idées reçues, influer négativement sur la confusion des modes indicatif futur et conditionnel. La confusion participe passé/infinitif présente, quant à elle, des tendances différentes. Les variables *âge* et *niveau d'études* sont liées à la variable *confusion des modes participe passé/infinitif*. C'est également le cas de la variable *sexe*, alors qu'elle n'était pas liée à la variable *confusion des temps indicatif futur/conditionnel*. Enfin, il n'existe pas de dépendance entre la confusion de ces modes et la Cémo. Ces résultats nous amènent à penser que la pratique des nouveaux média de communication et la réalisation d'écarts orthographiques ne sont pas liées ou, si elles le sont, que ce lien est plutôt positif.

Pour terminer, notons que si les conclusions de cette partie sont basées sur le corpus belge uniquement, elles peuvent toutefois éclairer des tendances générales quant aux pratiques réelles dans les sms, qu'il faudra confronter à de nouvelles données à l'aide de nouveaux corpus, issues d'autres régions, des données plus récentes, alignées au niveau du caractère.

Chapitre 10

Variation syntaxique

- Retraite mon cul, dit Zazie. Moi c'est pas pour la retraite que je veux être institutrice.
- Non bien sûr, dit Gabriel, on s'en doute.
- Alors c'est pourquoi ? demanda Zazie.
- Tu vas nous expliquer ça.[...]
- Alors pourquoi que tu veux l'être institutrice ?
- Pour faire chier les mômes.
Queneau (1959), pp. 23-24.

Sommaire

10.1 Interrogatives 232
10.1.1 Introduction 232
10.1.2 Variation interne 234
10.1.3 Variation externe 240
10.1.4 Conclusions 247
10.2 Négation 249
10.2.1 Introduction 249
10.2.2 Variation interne 249
10.2.3 Variation externe 259
10.2.4 Conclusions 265
10.3 Conclusions 267

Introduction

Dans la section *Préliminaires* de la partie analytique, nous avons mis en avant notre double approche du corpus : guidée par corpus (*corpus-driven*) et fondée sur corpus (*corpus-based*). Lors des analyses lexicales, principalement guidées par corpus, les questions de recherche ont jailli de l'observation d'une liste de mots connus et une liste de mots inconnus issus du corpus. L'approche fondée sur corpus consiste, au contraire, à confirmer ou infirmer des hypothèses, souvent présentes dans la littérature ou ailleurs, en les confrontant aux données issues des corpus. C'est cette dernière approche qui guide l'analyse syntaxique de cette partie. Nous allons confronter nos données aux postulats réalisés par ailleurs dans littérature : l'écrit SMS représente-t-il une pratique systématiquement abrégeante et ses traits

sont-ils comparables à ceux du français ordinaire ?

L'approche syntaxique de cette dissertation comporte un nombre moins important de sujets d'analyse (2 sujets) que la partie lexicale (5 sujets) ; ceci s'explique par le travail d'exploitation des données qui, dans le cas de la syntaxe, est considérablement plus conséquent et nécessite plus d'attention : le repérage des structures (proposition, phrase, etc.) est beaucoup plus complexe que le repérage d'unités lexicales. Leur annotation et la désambigüisation d'un certain nombre de cas rend la tâche encore plus délicate et chronophage.

Nous avons montré dans la partie 5.2.3, de façon tout à fait théorique et conceptuelle, que nous rapprochions l'écrit sms d'une langue de la proximité, selon le modèle de Koch et Oesterreicher (1985). Nous désirons voir si la syntaxe de l'écrit sms confirme cette théorie. Pour ce faire, nous travaillons donc sur deux plans d'analyse importants : l'abrègement ou non de la syntaxe et le rapprochement ou non de cette syntaxe des traits de l'oral ordinaire.

L'étude des variantes syntaxiques qui abrègent le message (en nombre de caractères) est particulièrement importante puisque, dans certains cas, l'abrègement ne coïncide pas avec la langue de la proximité, l'immédiat communicatif : la langue de « la distance, [...] vise une formulation bien formée, élaborée, explicite et compacte, planifiée. L'immédiat favorise au contraire une formulation décompactée, agrégative et provisoire » (Gadet, 2005, p. 4). Ce conflit entre nécessité de faire court et élan naturel d'une langue de proximité constitue la base de notre interrogation sur la syntaxe dans l'eSMS puisque, comme nous le verrons dans le deuxième phénomène choisi, les interrogatives, ce conflit est tout à fait de rigueur : une interrogative introduite à l'aide du syntagme *est-ce que* (moins formelle) est plus couteuse en nombre de caractères qu'une interrogative par inversion sujet-verbe (davantage formelle), par exemple.

La partie 5.2.3 a présenté les traits de l'oral ordinaire, selon Blanche-Benveniste (1997). Nous avons choisi d'étudier en profondeur deux de ces traits [1]. La sélection de ces deux objets d'analyse s'est réalisée selon plusieurs critères :

- nous avons éliminé les phénomènes qui ne tiennent pas de la syntaxe (absence de schwas, assimilation consonantique, syncopes et apocopes) ;
- nous avons éliminé les phénomènes qui sont non pertinents vu la nature de notre corpus (absence de ponctuation) ;
- enfin, nous avons mis de côté les phénomènes dont nous avons déjà traité précédemment dans la partie lexicale (prépondérance du présent de l'indicatif, des pronoms *ça* et *on* sur *cela* et *nous*).

1. Cette méthode est tout à fait critiquable, notamment parce qu'elle se base sur une liste pré-établie et non exhaustive de traits. Néanmoins, nous avons choisi de la suivre car elle répond à certaines questions de notre recherche générale (l'eSMS est-il assimilable à de l'oral ordinaire, comme le prétendent certains ?).

Conséquemment, les deux traits de l'oral ordinaire que nous avons gardés comme base de l'analyse sont les suivants :
– l'absence du clitique de négation *ne* ;
– l'évitement de l'interversion du sujet et du verbe par l'introduction des questions à l'aide de *est-ce que* et par le point d'interrogation en unique marqueur (entre autres).

De plus, l'étude de ces deux traits de l'oral ordinaire semble tout à fait pertinente dans le cadre de la conception de l'acte de communication qui est la nôtre et qui suit le modèle de Koch et Oesterreicher (1985). Comme le met en avant Gadet (2008, p. 520), l'absence du *ne* de négation et les questions avec éléments interrogatifs *in situ* forment deux des trois « faits syntaxiques attendus dans la proximité » [2]. L'auteure souligne d'ailleurs que l'on observe des résultats différents pour ces phénomènes entre le chat et le courriel ; notre étude propose des conclusions pour les sms.

Nous tenterons donc, dans cette partie, de répondre aux questions suivantes :

1. Les traits syntaxiques que l'on attribue généralement à l'oral ordinaire se manifestent-ils également dans l'eSMS ?
2. Comment se manifestent les interrogatives dans les sms ? Quelles répartitions des formes observe-t-on ? Les formes plus formelles sont-elles le fait d'une population en particulier ?
3. Quelles structures de négation courantes peut-on relever dans les sms ? Quelles sont les variables (internes et externes) de leur variation ?
4. Les auteurs des messages préfèrent-ils faire court (stratégie d'abrègement) ou utiliser des formulations décompactées (typiques de la communication immédiate) ?

2. Le troisième fait correspondant aux réponses au moyen de syntagmes et non d'énoncés complets.

10.1 Interrogatives

10.1.1 Introduction

La langue française met à disposition différentes constructions pour former une proposition interrogative directe ou indirecte. L'interrogative directe peut être totale (c'est l'ensemble du contenu propositionnel qui est mis en question : *mangerez-vous ce soir ?*) ou partielle (qui porte au contraire sur l'un des constituants de la phrase : *avec qui mangerez-vous ?*).

Nous déterminons deux grands groupes de constructions interrogatives en français : celles que nous nommerons *plus formelles* sont en réalité liées à des situations de langue soutenue, alors que celles que nous nommerons *moins formelles* apparaissent lors des situations de langue moins surveillées. Il existe trois stratégies de marquage morphosyntaxique des interrogatives en contexte *formel* :

- inversion de l'ordre des mots (interrogative totale) : le sujet y est postposé au verbe. L'interrogative est simple si le sujet est pronominal : *danserez-vous avec nous ?* ; elle est complexe si le sujet est lexical : *Luc et toi danserez-vous avec nous ?* ;
- emploi du syntagme interrogatif *est-ce que* qui comporte l'inversion en son sein et permet ainsi de conserver l'ordre des mots de la phrase assertive (interrogative totale[3]) : *est-ce que vous danserez avec nous ?* ;
- emploi des morphèmes interrogatifs (dorénavant *élément qu-*) avec inversion sujet-verbe : typiques de l'interrogative partielle, ils sont des déterminants (*quel*, *combien de*), des pronoms (*qui*, *que*, *quoi*, *lequel*) et des adverbes interrogatifs (*où*, *quand*, *comment*, *pourquoi*, *combien*) : *lequel désires-tu acheter ?*.

Dans l'usage moins formel, typique d'un registre moins soutenu, l'interrogative se manifeste principalement :

- introduite par un élément qu- sans inversion sujet-verbe (interrogative partielle) : *pourquoi il est parti ?* ;
- sous la forme d'une déclarative ou d'une impérative. À l'oral, ces interrogatives vont de pair avec une intonation particulière, montante (Gadet, 1989, p. 135). À l'écrit, c'est le point d'interrogation qui prend la place de l'intonation pour le marquage du questionnement : *c'est Luc qui part.* versus *c'est Luc qui part ?*. Ces interrogatives peuvent être totales ou partielles (*je vois pas qui tu es ?*[4]).

Sélectionner les structures concernées par une interrogation n'est pas une tâche évidente. C'est pourquoi, afin d'étudier la modalité interrogative dans le corpus de sms,

3. Excepté lorsqu'elle est combinée avec un élément qu- : *où est-ce qu'on se voit demain ?*

4. Notons que ce style d'interrogative peut tout à fait être interprété comme une faute de frappe : *je ne vois pas qui tu es !*. Néanmoins, les pratiques actuelles montrent de plus en plus d'emploi du point d'interrogation en fin de phrase assertive servant à demander un supplément d'information au destinataire.

nous entreprenons d'analyser **les interrogatives marquées d'un point d'interrogation**[5] dans le corpus de messages transcrit. Pour y parvenir, nous recourons à des grammaires de repérage de structures interrogatives en utilisant le logiciel de traitement de corpus Unitex. Nous avons segmenté la grammaire en deux parties :
– le repérage des interrogatives *formelles* (inversion sujet-verbe, syntagme *est-ce que*, élément qu-) ;
– le repérage des interrogatives *moins formelles* (**toute autre construction d'interrogative marquée d'un point d'interrogation**).

L'évaluation de la validité de ces grammaires de repérage des structures syntaxiques se fonde sur des mesures d'évaluation communes en extraction d'information (traitement automatique du langage) : le rappel et la précision. La linguistique y a également recours lorsqu'elle analyse de manière automatique des corpus (Rebeyrolle et Tanguy, 2000). Les deux mesures qui, combinées, donnent la F-mesure, permettent d'évaluer la pertinence du repérage de structures. Les formules de la Figure 10.1 détaillent le calcul de ces mesures.

FIGURE 10.1 – *Mesures d'évaluation des grammaires en vue de l'identification d'expressions dans un corpus*

$$Rappel = \frac{\text{Nombre de sms contenant une interrogative bien repérés}}{\text{Nombre de sms contenant une interrogative à repérer}}$$

$$Précision = \frac{\text{Nombre de sms contenant une interrogative bien repérés}}{\text{Nombre de sms contenant une interrogative repérés}}$$

$$\text{F} - \text{mesure} = \frac{2 \times \text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}}$$

Dans notre cas, nous voulons estimer la validité de deux types de repérage : le repérage des messages comprenant au moins une interrogative et le repérage des sms comprenant au moins une interrogative formelle par rapport à ceux ne comprenant qu'une ou plusieurs interrogatives moins formelles.

Les deux évaluations se basent sur un échantillon de 1000 messages. Pour la première évaluation, le rappel montre que 97 % des formes intéressantes (messages comprenant une interrogative au moins) sont repérées[6]. La précision prouve que 99% des messages repérés contenaient une interrogative[7]. La F-mesure obtenue est

5. Même si cela concerne en majorité des interrogatives directes, certaines formes d'interrogatives indirectes, qui se construisent, dans notre cas, avec une complétive introduite par un élément qu- suivie d'un point d'interrogation (*Je me demandais pourquoi tu voudrais venir ce soir ?*), sera donc incluse dans cette étude.

6. Les erreurs de rappel peuvent notamment être expliquées par la mauvaise reconnaissance des interrogatives pour lesquelles le point d'interrogation était accolé au mot le précédant.

7. Les erreurs de reconnaissance portaient sur des structures de type *Je t'aime à donf (?)* où le point d'interrogation apparait clairement comme une faute de frappe ou forme un signe que nous n'avons pas reconnu mais qui ne participe pas à une interrogative.

de 98 %.

Pour le deuxième repérage, le rappel prouve que 97% des messages comprenant au moins une interrogative formelle sont repérés [8]. Le calcul de la précision montre que 97 % des messages repérés étaient correctement étiquetés (interrogative formelle-moins formelle). La F-mesure obtenue est de 97 %.

10.1.2 Variation interne

Ces résultats permettent de créditer les typologies et les statistiques basées sur les graphes de reconnaissance. Nous avons identifié 11.659 messages contenant des interrogatives formelles. L'hypothèse envisagée dans Gadet (1989) pour un corpus téléphonique est la suivante : on téléphone pour demander quelque chose, ou pour mettre au point quelque chose qui demande un ajustement. Le corpus étudié montrait qu'on emploie en moyenne 3 interrogatives par minute lors d'un appel téléphonique. L'objectif est-il le même pour les échanges de sms ? Sur un total de 51.280 messages, 11.659 comportent - au moins - une interrogative (c'est-à-dire 22,7%) ; ces interrogatives sont émises par 1.776 auteurs sur les 2.563 qui ont participé à l'enquête. Ces chiffres sont relativement bas et ne permettent pas de conclure à un emploi du sms dont l'objectif principal serait similaire à celui d'une conversation téléphonique [9].

Parmi les 11.659 messages identifiés, 10.085 comportent au moins une interrogative moins formelle (mais aucune forme formelle) et 1.574 contiennent au moins une forme formelle. Au niveau des interrogatives repérées elles-mêmes, on recense 13.789 interrogatives moins formelles pour 1.721 interrogatives formelles.

La Figure 10.2 présente la répartition des formes formelles. La structure basée sur *est-ce que* est la plus fréquente (43%). Les interrogatives basées sur un élément qu- ou une inversion sujet – verbe apparaissent avec une fréquence identique, plus faible que celle du syntagme interrogatif (28%). Les structures basées sur la forme *est-ce que* et celles basées sur l'inversion sujet – verbe montrent une plus grande diversité que celles basées sur un élément qu- qui sont plus répétitives (par ex. *comment vas-tu ?*).
Les structures des interrogatives basées sur *est-ce que* ou un élément qu- (associé à une inversion) et attestées dans le corpus de sms sont détaillées en Annexe A.105

8. Les erreurs de rappel peuvent notamment être expliquées par la confusion des interrogatives moins formelles contenant la forme *rendez-vous*, qui est repérée comme une inversion sujet – verbe et est donc étiquetée comme *formelle*.

9. Ceci peut sans doute s'expliquer en partie par le fait que l'unité que constitue un conversation téléphonique n'est pas comparable à celle que constitue un sms : l'unité du sms est une sorte de tour de parole d'une conversation alors que l'unité de la conversation téléphonique est bien la conversation dans son ensemble.

FIGURE 10.2 – *Répartition des types de structures d'interrogatives formelles*

Type de structure	occurrences %	types %
est-ce que	43%	47%
élément *qu-*	28%	16%
inversion s-v	29%	37%

et A.106 [10]. La liste des 42 verbes employés dans les interrogatives formelles avec inversion sujet – verbe est présentée en Annexe A.107.

Plusieurs commentaires viennent compléter ces trois annexes. Premièrement, l'observation du graphe des interrogatives basées sur le syntagme *est-ce que* met en lumière plusieurs phénomènes :

- le graphe montre que les éléments qu- combinés avec le syntagme *est-ce que* sont *comment*, *pourquoi*, *où*, *qu'* et *quand*, là où on aurait pu s'attendre à d'autres formes telles que *qui* (par ex. *qui est-ce que tu vois demain ?*), étant donné le type de thématiques abordées dans les sms ;
- l'élément qu- le plus fréquent dans le corpus est l'élément *qu'* ; l'expression *qu'est-ce que* s'associe majoritairement avec les verbes *faire*, *raconter* et *penser* et le pronom *tu* ;
- les boites reprenant les diverses formes de *ce* (*c'*, *ça*, *cela*) représentent les constructions les plus fréquentes du corpus ; la structure interrogative *est-ce que ça va ?* (et ses variantes : *est-ce que tout va bien ?*, *est-ce que tout va bien pour toi ?*, etc.) est la plus courante dans le corpus ;
- le chemin central, pour lequel le sujet est lexical, représente seulement 2% des constructions du corpus ;
- les verbes apparaissent à 77% au **présent** ; **le conditionnel** (12%) et le passé composé (8%) occupent majoritairement les 23 % restants.

En ce qui concerne le graphe des interrogatives basées sur un élément qu- et une inversion sujet-verbe :

- la structure *comment vas-tu* concerne 47% des formes interrogatives avec élément qu- ; les structures *comment allez-vous ?* et *que faites-vous ?* sont également plus fréquentes que la moyenne ;
- conséquemment, c'est l'élément qu- *comment* qui est le plus employé (61%), suivi de *que* (19%), de *quand* (6%) et des différentes flexions de *quel* (5,5%) ;
- *combien*, en tant qu'élément qu- d'une interrogative formelle, n'apparait qu'une seule fois dans le corpus ;
- les verbes *aller*, *faire* et *être* sont à la base de 78% des constructions avec élément qu- et inversion, avec, respectivement, 57%, 9% et 12% ;

10. Le graphe ne représente donc pas la grammaire de repérage des interrogatives concernées mais permet une meilleure visualisation des structures attestées dans le corpus.

– **l'indicatif présent** est employé dans 95% des cas ; **le passé composé et le conditionnel** occupent une large part des 5% restant.

Enfin, la liste des 42 verbes employés dans les interrogatives par inversion mérite également quelques remarques. Premièrement, on note la présence soutenue des verbes *avoir* (27%), *pouvoir* (25%), *être* (15%), *venir* (5%), *vouloir* (5%) et *savoir* (4%), qui forment 81% des verbes employés. Contrairement à ce qui est de rigueur dans les constructions avec le syntagme *est-ce que*, les verbes *faire* et *aller* ne sont pas récurrents dans les structures par inversion sujet-verbe. **L'indicatif présent** est de rigueur dans 88% des cas ; **le conditionnel et le futur** occupent une majeure partie des 12% restant.

Les structures interrogatives formelles représentent 11% du nombre total d'interrogatives. Elles laissent transparaitre des structures étonnamment complexes dans le contexte rédactionnel des sms (cf. exemples 85 et 86) et des formules d'ordinaire assez rares dans un contexte aussi informel (ex. 87).

(85) Hello the american man in London ! Ton séjour **se passe-t-il** bien ? Repères-tu de chouettes trucs ? Bisous (smsFB)

(86) ...Et paf à nouveau en larmes dans la voiture... **Y a-t-il** seulement une solution ? Je déteste les comptes d'apothicaire (trop de mauvais souvenirs) (smsFB)

(87) Alors, c'était quoi le bouquin en français ? Ça été ? T'es déjà en vacances ou t'as encore une semaine à tirer ? Ma mère m'a lâché sa caisse vendredi, grande première ! **Qu'en est-il** de ta Fiesta ? Biz (smsFB)

Dans le corpus, 89 % des interrogatives totales sont des interrogatives moins formelles qui présentent une typologie distincte des formelles. Les catégories sont exclusives [11] et apparaissent en ordre décroissant d'importance.

1. Interrogatives régionales (emploi d'une syntaxe ou d'un lexique régional) [12] : (Suisse) *ça jouerait pour toi vendredi soir ?* - (La Réunion) *Bonsoir bébé, je fais rien lol, kwa fé ta journée ?* ;
2. Interrogatives familières de type *ça roule ?*, *ça boume ?*, *ça craint ?*, *ça colle ?*, *ça roots ?*, *ça farte ?*, *ça marche ?*, *ça schtroumpf ?*, *ça vate ?* et *Ça swingue ?* : *Je peux passer demain vers 15h30. Ça colle ? - j'ai pas encore ma carte d'étudiant plastifiée et sans photo ça craint ? - Je te propose 17h15 à Payerne. Ça marche ? - Salut [NPERS], ça roule ?* ;

11. Cela signifie que dans le cas d'une interrogative appartenant à plusieurs catégories (par exemple *familière*, *conditionnel* et *par point d'interrogation*, elle entre dans la première catégorie et n'est pas reprise par les autres (dans notre cas, *familière*). Ainsi, appartiennent à la catégorie *par point d'interrogation* toutes les propositions suivies d'un point d'interrogation et qui n'appartiennent à aucune autre catégorie.

12. On notera la totale absence dans le corpus de sms de l'élément *ti*, reste de l'inversion *t-il*, typique du français du Québec, que Gadet (1989) a déjà qualifié de « tombé en désuétude » dans le cadre de l'étude du français ordinaire.

3. Déclaratives devenant interrogatives par l'adjonction des interjections *hein* ou *hein dis* : *On se voit quand même demain hein ? - c'était chouette ce week-end hein dis ?* ;
4. Interrogatives disjonctives[13] en *ou* + alternative, *ou pas*, *ou quoi* et *non* : *T'es au cours ou t'es au kot ? - Alors cette journée bien ou pas ? - Je te dis quand je sors ou quoi ? - L'élève apprend bien non ?* ;
5. Interrogatives introduites ou complétées (après transformation) par un élément qu- sans inversion sujet – verbe : *comment ça va ? - comment c'est allé ton examen de géométrie différentielle ? - Et vous avez mis combien chacun ? - c'est qui [NPERS] ? - tu as quel âge ?* ;
6. Interrogatives par le conditionnel : *t'arriverais à avoir l'adresse e-mail de [NPERS] [NPERS] qui est à l'uni de Neuchâtel en droit ? - on irait pas sur la foire tous ensemble vendredi après le cours ?* ;
7. Interrogatives avec ellipse de verbe, présentant des syntagmes nominaux seuls ou introduits par un élément qu-, *et* ou *alors*, entre autres, ou présentant des informations liées aux rendez-vous, introduites par *à*, *en*, *chez*, etc. : *Quel genre de jeu ? - quelles nouvelles ? - Ça va et toi ? - Et les amours ? - Quoi lol ? - À demain 19h chez moi ?* ;
8. Interrogatives dont la structure syntaxique est celle d'une déclarative ou d'une impérative suivie d'un point d'interrogation (désormais *interrogative par point d'interrogation*) : *je vois pas qui tu es ? - Ce week-end si t'es là, dis-le-moi ? - Au fait t'as lu ma carte qui était avec la lettre ?.*

Devant cette diversité des types, la question de leur fréquence relative n'est pas triviale. La répartition des différents types dans le corpus de sms francophone est illustrée dans la Figure 10.3. L'interrogative par point d'interrogation est la forme la plus répandue du corpus de sms. L'interrogative par élément qu- (sans inversion sujet-verbe) et l'interrogative avec ellipse de verbe apparaissent en deuxième et troisième positions. Les interrogatives par conditionnel et les disjonctives sont plus rares mais concernent tout de même un millier de messages. Les interrogatives moins formelles les moins fréquentes sont celles qui se closent par *hein* et *hein dis* et celles qui présentent des tournures populaires ou régionales.

FIGURE 10.3 – *Répartition des types d'interrogatives moins formelles*

	élément qu-	point d'interr.	hein, hein dis	conditionnel	disjonctives	averbales	populaires	régionales
pourcentage	20,31%	62,00%	0,30%	2,82%	4,75%	9,27%	0,22%	0,33%
occurrences	2800	8549	42	389	655	1278	30	46

Au vu de la quantité de messages concernés par les interrogatives moins formelles, une analyse semi-automatique de ceux-ci est peu envisageable (important travail d'annotation et de désambigüisation). Notre intérêt va donc porter sur deux points spécifiques : les interrogatives avec ellipse de verbe et les disjonctives (1.933

13. Interrogatives proposant une alternative au sein de la question.

interrogatives sont concernées).

Les interrogatives avec ellipse de verbe peuvent être présentées en deux grands groupes : (1) les syntagmes nominaux seuls (14,45% : *je ne sais pas. Toi ?*) et (2) les syntagmes introduits par un élément (qu- et autres) (85,55%). Parmi les éléments introducteurs, on recense :

- *et* : il s'agit de la structure la plus courante[14] (52,6%). On retrouve majoritairement la forme *et toi ?* suivant une assertion (*Salut, ça va ? Demain moi j'arriverai vers 8h15. Et toi ?*). En dehors des structures associant *et* à un pronom, on recense également de nombreux sujets/objets lexicaux (*Comment vas ? Alors le stage ? Cool ? Et l'étude ?*) ;
- *quel* : l'élément apparait dans 11% des cas, le plus souvent dans la structure *quelle heure ?* ; il est accompagné d'un sujet/objet lexical simple (*quelles mains ?*) ou composé (*quel genre de question ?*) ;
- *quoi* (7%) : l'élément apparait principalement dans la structure *quoi de neuf ?* et sa variante *quoi de new ?* ;
- *pourquoi* (4,6%) : on recense principalement *pourquoi pas ?* et un complément optionnel (*Pourquoi pas un midi ou une après-midi ?*) et *pourquoi donc ?*
- *comment* (4%) : on trouve uniquement l'expression *comment ça ?*, quelquefois accompagnée d'un complément (*Je regarde. Comment ça dans les buts ?*).
- *quand* (2,9%) : il n'est jamais associé, dans le corpus, à un véritable syntagme ; il fait partie d'une expression (*quand ça ?*), il est accouplé à un deuxième élément qu- (*quand où ?*) ou à un autre élément introducteur des interrogatives avec ellipse du verbe (par ex. avec *alors*, dans *quand alors ?*) ;
- *alors* (1,7%) : dans le corpus, il est assemblé à un sujet animé (*alors [NPERS] ?*) ou inanimé (*alors les examens ?*) ;
- *à*, *avec* et *combien* forment à eux trois 1,75% des formes introductives rencontrées. La forme *à* introduit un lieu ou une information temporelle, *avec* est attesté uniquement dans la structure *avec qui* et *combien* dans la structure *combien de* + complément.

Les interrogatives disjonctives se regroupent de la façon suivante : la disjonction en *ou* concerne 77% des formes attestées, celle en *ou quoi* arrive en deuxième position avec 6,6%, suivie de *non* (5,5%), de *ou pas* (5,5%) et de *ou bien* (2,2%). Les autres formes (*ou non*, *ou...ou* et *ou aussi*) sont plus négligeables (elles occupent chacune 1% des observations). La disjonction se réalise donc majoritairement en milieu de phrase et non en clôture.

La disjonction en *ou*, la plus fréquente, regroupe des formes assez diversifiées. On segmentera les formes où la disjonction se fait entre deux propositions (ex. 88), dont certaines se font avec changement de sujet (ex. 89), et celles où la disjonction se fait entre deux syntagmes (ex. 90 et 91).

14. Elle est étonnamment plus fréquente que la structure *syntagme nominal* seule.

(88) Encore moi ! ;) J'ai vu les chaussures chez Heine mais elles ont 8 cm de talon ! C'est un peu beaucoup pour moi... Je$_{S1sg}$ te ramène le catalogue **ou** je$_{S1sg}$ le jette ? (smsCH)

(89) Eh quoi là, tu fais quoi ! ? On$_{S3sg}$ t'attend **ou** tu$_{S2sg}$ nous rejoins au Mac Do ? (smsBE)

(90) Tu pourrais demander si tu peux emprunter le DVD$_{OLex}$ **ou** vidéo de petu$_{OLex}$? (smsCH)

(91) Oki, au fait, tu parles plus créole$_{OLex}$ **ou** français$_{OLex}$? (smsLR)

En ce qui concerne la disjonction en *ou quoi*, elle clôture donc une proposition. Elle n'apparait jamais apposée et concerne toujours, dans le corpus de sms, un verbe dont le sujet est le pronom personnel de la deuxième personne du singulier (ex. 92 à 96).

(92) Pourquoi tu ne décroches plus pour ta petite maman adorée ? ? T'es fâché **ou quoi** ? (smsBE)

(93) Tu t'es fait attaquée chaussée de Gand **ou quoi** ? (smsBE)

(94) T'as re du crédit **ou quoi** ? (smsLR)

(95) Kwa fé t'es bien arrivé vivant à ta case **ou quoi** ? (smsLR)

(96) Ouh la la. A cette heure-ci y a pas grand chose à faire ! Tu bosses pas demain **ou quoi** ? (smsLR)

À l'examen des 5 exemples précédents, il apparait clairement que l'élément disjonctif *ou quoi* évoque la plupart du temps une demande d'information plus grande que la vérification d'une information sous la forme d'un choix entre plusieurs alternatives (comme c'est le cas pour *ou* et *non*, par exemple). En ce sens, elle est une interrogative partielle. La question est d'ailleurs la plupart du temps dirigée (vers l'une des deux alternatives) : dans des cas tels que les exemples 92 et 94, c'est la première partie de la disjonction qui est envisagée (*t'es fâché*, *T'as re du crédit*), le pronom invite à être rassuré par rapport à la première partie et à livrer plus d'informations. Au contraire, dans des exemples tels que 93 et 95, c'est sur la deuxième alternative, matérialisée par le pronom *quoi*, que porte uniquement la question. Enfin, on recense quelques rares cas où la disjonction en *ou quoi* envisage réellement deux alternatives (ex. 96).

La disjonction en *non* appartient à la classe des interrogatives totales, la véracité de la première alternative étant remise en question ; elle n'apparait qu'une seule fois apposée, séparée par une virgule. Pour les autres occurrences, *non* est intégré à la fin de la proposition (ex. 97) sans apposition.

(97) Ah ok lol. Oui c'est vrai 'Ben c'est cool alors tu pourras m'apprendre à cuisiner mdr. **Mais ta copine elle cuisine des fois non ?** (smsLR)

10.1.3 Variation externe

Afin d'approfondir la question de l'interrogative dans les sms, nous avons choisi d'en étudier la variation selon deux axes : intra-individuel et diastratique.

Pour entamer la première approche de la variation, au sein de l'ensemble des messages d'un même individu, nous prenons le parti de reprendre une assertion issue de Gadet (1989, pp. 142 et 146) :

> « l'interrogation constitue [...] l'un des rares domaines de langue où une situation peut directement influer sur la fréquence des formes : certaines situations mettant en jeu des demandes de renseignements entrainent une accumulation de questions. [...] les locuteurs ont la capacité, au moins passive, de faire usage de toutes les formes offertes par la langue, et tout montre que le recours à l'une ou l'autre forme n'est pas indifférent »

Nous avons donc entamé une analyse de la variation intra-individuelle pour deux locuteurs donnés[15].

Le premier locuteur (LOC1) est une jeune femme belge de 16 ans, de langue première française et étudiante dans le secondaire inférieur. Nous recensons pour cette locutrice 191 interrogatives contenues dans 107 messages. La typologie de ces interrogatives est présentée dans la Figure 10.4.

FIGURE 10.4 – *Répartition des types d'interrogatives pour LOC1*

structure	tokens	%
inversion	1	0,52%
est-ce que	7	3,66%
disjonction	9	4,71%
averbal	24	12,57%
point d'interrogation	73	38,22%
élément qu-	77	40,31%
total	191	100%

LOC1 présente un sous-emploi des interrogatives formelles (8 occurrences). Les interrogatives comprenant un élément qu- constituent la forme la plus répandue chez ce locuteur ; l'inversion n'y est jamais attestée. L'élement qu- apparait rarement en début d'interrogative (*comment ça se fait que t'as fini tôt ?*) ; c'est la forme *tu fais quoi (toi/de beau) ?* qui est la plus fréquente. Les interrogatives par point d'interrogation sont quasiment aussi courantes (73 occurrences) et concernent principalement la structure *ça va ?* et ses variantes *ça va toi ?* et *et toi ça va ?*. Il s'agit d'interrogatives totales dans leur forme mais qui appellent une réponse plus complexe que *oui/non*. Au sein de cette classe, nous observons deux structures plus

15. Notre choix a porté sur deux participants dont le nombre de messages comportant des interrogatives était important.

populaires par dédoublement de l'élément qu- : *et comment qu'on doit s'habiller demain ?* et *pourquoi qu'on se ferait du mal ?*.

Notons que les deux interrogatives les plus récurrentes du corpus, que sont les formes *tu fais quoi ?* et *ça va ?* (et leurs variantes), apparaissent presque toujours comme tremplin en vue d'introduire une information concernant l'auteur lui-même (par ex. *Tu fais quoi ? Moi me prépare à bouffer !* ou *Ça va bien ? ? Moi bof j'arrive pas à être bourrée !*[16]).

Les interrogatives avec ellipse de verbe (24 occurrences) concernent surtout des structures amorcées par *et*, comme c'était le cas dans la typologie générale, ainsi que des formes *ok ?* seules. Enfin, les disjonctives recensées chez LOC1 reprennent certaines des formes énumérées précédemment : *ou*, *ou pas*, *ou quoi* et *non*.

Le second locuteur (LOC2) est une jeune femme réunionnaise de 16 ans, de langue première française et étudiante dans le secondaire supérieur. Nous recensons pour cette locutrice 177 interrogatives contenues dans 139 messages. La typologie de ces interrogatives est présentée dans la Figure 10.5.

FIGURE 10.5 – *Répartition des types d'interrogatives pour LOC2*

structure	tokens	%
qu- + inversion	3	1,69%
est-ce que	7	3,95%
disjonction	9	5,08%
averbal	37	20,90%
élément qu-	41	23,16%
point d'interrogation	80	45,20%
total	177	100%

LOC2 présente, comme LOC1, un sous-emploi des interrogatives formelles (10 occurrences). Toutefois, contrairement à LOC1, elle emploie des interrogatives avec élément qu- et inversion sujet-verbe (*Et comment pourrais-je te croire ?*). Les interrogatives par point d'interrogation constituent la classe la plus importante de ce sous-corpus (80 occurrences). Elles concernent majoritairement la structure *ça va ?*, mais, contrairement aux interrogatives de LOC1, elles ne servent pas de tremplin en vue d'introduire une information concernant l'auteur lui-même (par ex. *Hé mon cœur ça va ? T'inquiète j'arrête de te demander des trucs lol*). Les interrogatives comprenant un élément qu- arrivent en seconde position mais sont moins nombreuses (41 occurrences). L'élément qu- y apparait souvent en début d'interrogative (*Quel coin triste ?*) ; une forme originale apparait dans cette classe, à deux reprises : *quoi* en début d'interrogative (*Quoi c'est une blague ?* et *Quoi t'es*

16. On note ici un cas de multiplication du signe de ponctuation servant à souligner l'importance de la question.

en terminale ? ! Ah j'ai dû rater un épisode alors lol[17]).

Les interrogatives avec ellipse de verbe (37 occurrences) concernent surtout des structures amorcées par *et*, des formes d'interjections telles que *ah bon ?* et des expressions telles que *quoi de neuf ?*. Enfin, les disjonctives recensées chez LOC2 reprennent certaines des formes énumérées précédemment : *ou*, *ou pas*, *ou bien* et *non* ; notons encore que ce dernier type de disjonctive, à la différence de ce que l'on trouve chez LOC1, est souvent concurrencé par sa variante *nan* (*Sinon on peut aller en ville nan ?*).

Les interrogatives des locutrices LOC1 et LOC2 présentent quelques points intéressants. Premièrement, les locutrices n'emploient pas uniquement un ou deux types d'interrogatives en particulier mais exploitent effectivement **la diversité des formes à leur disposition**. Ensuite, on remarque que les constructions formelles (inversion sujet-verbe, forme en *est-ce que* et élément qu- avec inversion) sont minoritaires chez les deux locutrices[18] mais sont tout de même attestées, signe que les locutrices les possèdent dans leur répertoire. Les interrogatives moins formelles par conditionnel sont absentes de ces deux corpus. Enfin, les deux locutrices, bien que du même sexe et du même âge, présentent une distribution des types non identique (primauté de l'interrogative par point d'interrogation chez LOC2, absence de l'interrogative avec élément qu- et inversion chez LOC1, etc.). Cela tend à démontrer que **la préférence des formes n'est pas homogène au sein de la population étudiée**.

La deuxième approche de la variation externe, diastratique, a pour objectif d'explorer les pratiques écrites des interrogatives dans les sms en fonction de l'âge, du sexe, de la langue première, du nombre de sms envoyés par semaine et du niveau d'études des auteurs.

Pour ce faire, nous allons, comme dans la partie lexicale, avoir recours à des tests statistiques adaptés aux types de données. Pour examiner l'influence de l'âge sur le type d'interrogative, nous effectuons un test t après avoir transformé la variable en regroupant les données en catégories[19]. Le résultat du test de Levene (Annexe A.108) semble indiquer qu'il y a une inégalité des variances entre les groupes (la valeur p est inférieure à 0,05). Le test t indique qu'il semble y avoir dépendance (la valeur p est inférieure à 0,05) entre les variables *âge* et *emploi d'une interrogative formelle ou moins formelle*. Au vu du box-plot de la Figure 10.6, on peut dire

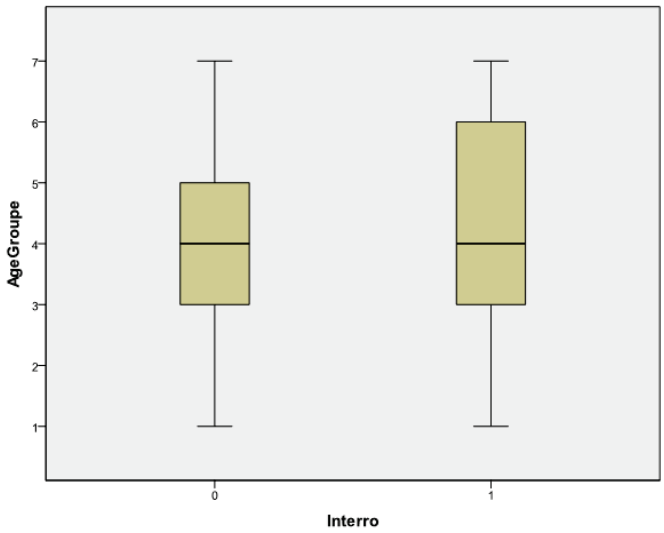
17. On remarque dans cet exemple un cas non isolé de combinaison de signes de ponctuation : la modalité exclamative vient s'ajouter à l'interrogation. Il y a à la fois interrogation et étonnement.

18. Ce phénomène est peut-être explicable par leur âge (LOC1 et LOC2 ont toutes les deux 16 ans) ou par leur sexe (F).

19. Pour rappel, les catégories d'âge sont les suivantes : 1 (10 à 15), 2 (16 à 17), 3 (18 à 19), 4 (20 à 22), 5 (23 à 27), 6 (28 à 40) et 7 (41 à 70). Ces catégories permettent une distribution dite « normale » (cf. Annexe A.9).

qu'il existe une répartition des âges plus étendue pour les sms comprenant une interrogative formelle (1) que pour les messages comprenant une interrogative moins formelle (0). Nous pouvons donc conclure que **la distribution des sms comprenant des interrogatives moins formelles concerne une classe d'âge plus restreinte, plus jeune (les valeurs sont plus ramassées) que celle des sms comprenant une interrogative formelle (la distribution est plus variée)**.

FIGURE 10.6 – *Box-plot représentant les types d'interrogatives en fonction des classes d'âge*



Pour évaluer le lien entre le sexe des auteurs de messages et l'emploi des types d'interrogatives (deux variables qualitatives), nous effectuons un test khi-carré. Le résultat du test (la valeur p est égale à 0,19, cf. Annexe A.109) indique qu'il ne semble pas y avoir de dépendance entre les deux variables. **Il n'y a donc pas de lien entre le sexe des auteurs de message du corpus de sms et l'emploi d'une interrogative formelle ou moins formelle**.

La dépendance entre les variables *langue première* et *types de propositions interrogatives* se calcule en regroupant les langues comme suit : (0) langue première française uniquement et (1) bilinguisme ou langue première autre que le français. Le résultat du test khi-carré (la valeur p est inférieure à 0,05, cf. Annexe A.110) indique qu'il semble y avoir dépendance entre les variables. Les résidus ajustés de la Figure 10.7 permettent de détailler ce résultat puisque ceux-ci montrent dans quelles cellules se trouvent des écarts significatifs par rapport à la situation d'indépendance des deux variables. Les chiffres mis en couleur indiquent à l'intérieur de quelles cellules les résidus sont « significativement négatifs », c'est-à-dire là où il est presque certain que les sous-représentations ou sur-représentations ne sont

pas dues à une erreur d'échantillonnage. Ainsi, **on note une sous-représentation des locuteurs de langue première française (uniquement) employant des interrogatives moins formelles et une sous-représentation des locuteurs bilingues natifs ou non francophones employant des constructions interrogatives formelles**. On peut sans doute conclure que l'emploi des interrogatives formelles, formes plus complexes de la langue, est dépendant de la maitrise de cette langue.

FIGURE 10.7 – *Analyse des résidus standardisés relatifs au test khi-carré (interrogatives-langue première)*

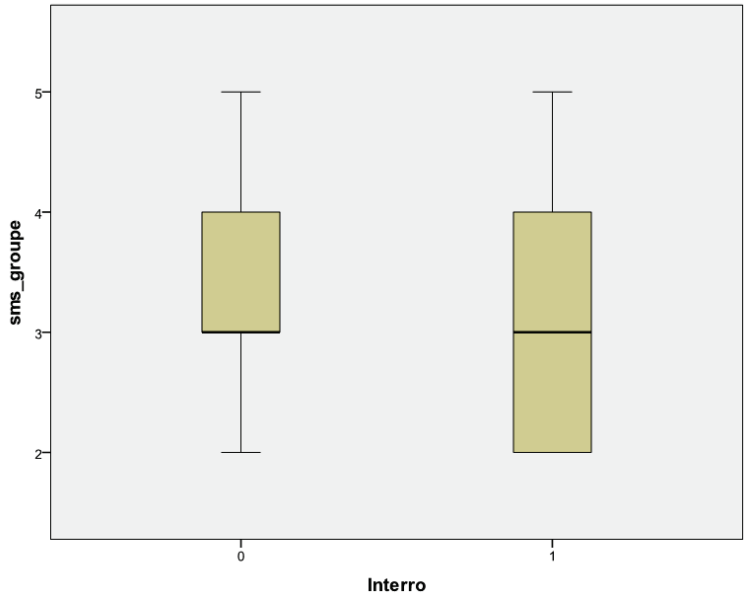
Tableau croisé Interro * Langue Maternelle : Groupe

			Langue Maternelle		
			0	1	Total
Interro	0	Effectif	8767	1318	10085
		Résidu ajusté	-6,4	6,4	
	1	Effectif	1458	116	1574
		Résidu ajusté	6,4	-6,4	
Total		Effectif	10225	1434	11659

L'évaluation de la dépendance des variables nombre de sms envoyés par semaine et emploi d'un type spécifique d'interrogative se base sur un test t dont le détail se trouve en Annexe A.111. Pour rappel, les groupes de fréquence d'envoi des sms sont les suivants : (2) moins de 10 sms (3) de 10 à 20 sms (4) de 20 à 50 sms et (5) plus de 50 sms envoyés par semaine. Le test t indique qu'il semble y avoir dépendance (la valeur p est inférieure à 0,05) entre les variables nombre de sms envoyés par semaine et emploi d'une interrogative formelle ou moins formelle. Au vu du box-plot de la Figure 10.8, on peut dire qu'il existe une répartition des interrogatives moins formelle (0) plus ramassée que celle des interrogatives formelles (1). Nous pouvons donc conclure que **la distribution des sms comprenant des interrogatives moins formelles concerne une classe plus restreinte que celle des sms comprenant une interrogative formelle (la distribution est plus variée)** : le sous-groupe de population envoyant moins de 10 sms par semaine (peu de sms) est sous-représenté dans l'emploi des interrogatives moins formelle.

Enfin, nous étudions le lien de dépendance des variables types d'interrogatives et niveau d'études. Pour rappel, les catégories de niveaux d'études sont les suivantes : (2) diplôme de primaire ou de secondaire inférieur obtenu ou en cours (3) diplôme de secondaire supérieur obtenu ou en cours et (4) diplôme d'études post-secondaires obtenu ou en cours. Le résultat du test khi-carré (la valeur p est inférieure à 0,05, cf. Annexe A.112) indique qu'il semble y avoir dépendance entre les variables. Les résidus ajustés de la Figure 10.9 permettent de détailler ce résultat puisque ceux-ci montrent dans quelles cellules se trouvent des écarts significatifs par rapport à la situation d'indépendance des deux variables. Les chiffres mis en couleur indiquent là où il est certain que les sous-représentations ne sont pas dues à une erreur d'échantillonnage. Ainsi, **on note une sous-représentation des locuteurs diplômés**

FIGURE 10.8 – *Box-plot représentant les types d'interrogatives en fonction de la fréquence d'envoi de messages*



ou en cours d'études du secondaire supérieur employant des interrogatives formelles (1) et une sous-représentation des locuteurs diplômés ou en cours d'études post-secondaires employant des constructions interrogatives moins formelles[20]. Il semblerait donc, selon le corpus sms4science, que plus le niveau d'étude est élevé, plus on utilise des formes d'interrogatives formelles.

FIGURE 10.9 – *Analyse des résidus standardisés relatifs au test khi-carré (interrogatives-niveau d'études)*

Tableau croisé Interro * éducation

			éducation			
			2	3	4	Total
Interro	0	Effectif	1195	3142	5748	10085
		Résidu ajusté	-1,3	5,4	-4,2	
	1	Effectif	204	385	985	1574
		Résidu ajusté	1,3	-5,4	4,2	
Total		Effectif	1399	3527	6733	11659

Pour clore cette étude des facteurs externes de la variation, nous investiguons les différences de pratiques entre régions de la francophonie. Les effectifs concernant chaque région se trouvent en Annexe A.113. Le résultat du test khi-carré (la valeur p est inférieure à 0,05, cf. Annexe A.114) indique qu'il semble y avoir dépendance entre les variables. Les résidus ajustés de la Figure 10.10 permettent de détailler ce résultat, puisque ceux-ci montrent dans quelles cellules se trouvent des écarts significatifs par rapport à la situation d'indépendance des deux variables. Les chiffres mis

20. Aucune conclusion n'est rendue possible pour la classe d'étude (2).

en couleur indiquent la position où il est certain que les sous-représentations ne sont pas dues à une erreur d'échantillonnage. Ainsi, note-t-on **une sous-représentation des locuteurs belges et québécois employant des interrogatives moins formelles (0) et une sous-représentation des locuteurs réunionnais employant des constructions interrogatives formelles** (1) (une tendance probablement induite par la présence soutenue du créole dans les sms). En ce qui concerne les tendances suisses, les conclusions sont moins tranchées.

FIGURE 10.10 – *Analyse des résidus standardisés relatifs au test khi-carré (interrogatives-région)*

Tableau croisé Région * Interro

			Interro		
			0	1	Total
Région	BE	Effectif	7016	1210	8226
		Résidu ajusté	-5,9	5,9	
	CH	Effectif	1001	119	1120
		Résidu ajusté	3,0	-3,0	
	LR	Effectif	1386	64	1450
		Résidu ajusté	10,8	-10,8	
	QU	Effectif	682	181	863
		Résidu ajusté	-6,7	6,7	
Total		Effectif	10085	1574	11659

10.1.4 Conclusions

Les grammaires de repérage des structures interrogatives ont permis d'identifier, dans le corpus de 51.280 sms, un ensemble de 10.085 messages contenant uniquement des interrogatives moins formelles et un ensemble de 1.574 messages comprenant (au moins) une interrogative formelle. **Au vu de la proportion de sms contenant des interrogatives** (un cinquième des messages seulement), **il est difficile de conclure que la communication par sms sert principalement, comme la communication téléphonique, à demander quelque chose ou à mettre au point quelque chose qui demande un éclaircissement** (Gadet, 1989). Cet ensemble de 11.659 messages contient 15.510 formes d'interrogatives dont seuls 11% sont des interrogatives formelles. **Il existe donc une très faible propension à utiliser les interrogatives formelles dans les sms du corpus**.

La répartition des interrogatives formelles repérées met en avant la supériorité numérique de la construction basée sur le syntagme interrogatif *est-ce que* et la faible diversité des constructions sur base d'un élément qu- accompagné d'une inversion sujet – verbe, qui se résument, la plupart du temps à la question *comment vas-tu ?*. La réalisation la plus courante du syntagme *est-ce que* dans le corpus est *est-ce que ça va ?* ; les interrogatives formelles construites avec le syntagme et un sujet lexical représentent seulement 2% du corpus total. Les interrogatives par inversion sujet – verbe se construisent surtout avec *avoir*, *pouvoir*, *être*, *venir*, *vouloir* et *savoir*, et beaucoup plus rarement avec *faire* et *aller*. **L'indicatif présent est le temps le plus employé dans les interrogatives formelles** (77% avec *est-ce que*, 95% avec élément qu- et inversion et 88% avec inversion seule).

Les interrogatives moins formelles, supérieures en nombre, présentent une diversité importante. C'est l'interrogative par point d'interrogation qui est la forme la plus courante de cet ensemble (62%), suivie par l'interrogative basée sur un élément qu- (20,31%) et l'interrogative par ellipse du verbe (9,27%). Au sein de l'ensemble des interrogatives moins formelles par ellipse du verbe, c'est la construction introduite par *et* qui est la plus fréquente (52,6%). Les interrogatives disjonctives concernent 4,75% des interrogatives moins formelles et c'est la forme *ou* entre deux propositions complètes qui est la plus fréquente.

Le classement total des formes interrogatives du corpus de sms est donc le suivant (par ordre de fréquence décroissante) :

1. Interrogatives par point d'interrogation (8.549 occurrences) : 55,12% des formes attestées ;
2. Interrogatives par élément qu- avec (481 occurrences) et sans (2.800 occurrences) inversion : 21,15% des formes attestées ;
3. Interrogatives avec ellipse du verbe (1.278 occurrences) : 8,24% des formes attestées ;

4. Autres interrogatives moins formelles (disjonction, régionales, etc.) : 7,5% des formes attestées (1.162 occurrences) ;
5. Interrogatives par syntagme interrogatif *est-ce que* (740 occurrences) : 4,77% des formes attestées ;
6. Interrogatives par inversion (500 occurrences) : 3,22% des formes attestées.

Si l'on compare ce classement avec, par exemple, des conversations non soutenues [21], on remarque une nette différence ; là où nous trouvons les interrogatives par point d'interrogation en tête, suivies des interrogatives par morphèmes interrogatifs, Vinay (1971) relève une majeure partie d'interrogatives par intonation seule (qui sont en fait celles marquées d'un point d'interrogation), suivies des interrogatives par syntagmes interrogatifs (*est-ce que*) qui se positionne chez nous en bas de classement. Cette différence de pratique dans le contexte des sms peut sans doute s'expliquer par le nombre important de caractères que le syntagme *est-ce que* occupe dans un message.

L'approche intra-individuelle des interrogatives a permis de montrer qu'**un seul et même locuteur profite de la diversité des formes interrogatives qui lui sont accessibles** et que **les préférences d'emploi de ces formes peuvent varier d'un locuteur à l'autre, même lorsqu'ils sont du même sexe et du même âge** (LOC1 préférant les constructions avec élément qu- aux interrogatives par point d'interrogation).

L'approche diastratique, elle, permet de dégager des tendances générales pour une population donnée (les participants à l'enquête). Ainsi avons-nous montré qu'**il n'y a pas d'incidence du sexe d'un locuteur sur l'emploi des interrogatives formelles plutôt que moins formelles**. Par contre, les tests statistiques ont mis en avant **la sous-représentation des locuteurs non francophones purs (c'est-à-dire les non-francophones natifs et les bilingues natifs) et des locuteurs diplômés (ou en cours d'études) du secondaire supérieur qui emploient des interrogatives formelles**. Enfin, au niveau des régions, nous avons relevé **des taux de locuteurs belges et québécois employant des interrogatives moins formelles et de locuteurs réunionnais employant des interrogatives formelles plus bas que ce à quoi on aurait pu s'attendre** si la répartition avait été équilibrée.

21. Conversations issues d'un roman et censées imiter le français parlé (Vinay, 1971).

10.2 Négation

10.2.1 Introduction

Cette partie a pour objectif de décrire la possible absence du *ne* de négation dans les messages du corpus francophone. Le phénomène d'absence du *ne* fait l'objet de très nombreux travaux en français, principalement axés sur le français parlé (Armstrong, 2002 ; Armstrong et Smith, 2002 ; Ashby, 2001 ; Dufter et Stark, 2007 ; Gadet, 1989). Gadet (2000, p. 156) précise : « Un renouvellement décisif de ces études intervient avec la possibilité de référence à de grands corpus oraux : alors qu'auparavant on ne fait que signaler l'opposition entre présence et absence, les corpus vont désormais permettre dénombrement, corrélations, et tentatives d'explications ». Les années 1970 ont vu la diffusion des études sur la variation du *ne* selon des critères sociolinguistiques (par ex. Pohl (1972) pour l'étude du *ne* en fonction de l'âge ou Sankoff et Vincent (1977) pour la variation diatopique France-Québec).

Les nouveaux médias de communication ont élargi les recherches en (socio)linguistique au monde de l'écrit ; ceux-ci, comme nous l'avons expliqué précédemment, ont offert un espace de communication moins formel où l'étude de la présence du *ne* a pris tout son sens (Panckhurst, 1998 ; Van Compernolle, 2007). Dans Panckhurst (1998, pp. 57-58), l'auteure conclut à une résistance du *ne* (- de 10% d'absence) dans le courrier électronique en 1997 et 1998. Van Compernolle (2007), pour sa part, a étudié des conversations de chat en 2005 ; selon lui, il existe un pourcentage de 85% d'absence du *ne*, mais son emploi est plus fréquent avec les forclusifs (Damourette et Pichon, 1927) *jamais*, *aucun*, les forclusifs multiples (*plus rien*, *plus personne*) et l'impératif. Les types différents de Cémo (le courrier électronique et le chat) influent-ils sur une pratique différente de la négation ou la variation est-elle due aux différentes dates de collectes (8 années séparent les deux enquêtes) ?

Plus précisément, nous avons pour objectif de répondre aux questions suivantes :

1. La présence du clitique *ne* dans les sms est-elle plutôt comparable à celle du courrier électronique (faible absence du *ne*) ou à celle du chat (forte absence du *ne*) ?
2. Comment les phrases négatives se manifestent-elles dans les sms, notamment en fonction des forclusifs, des modes verbaux et de la nature du sujet ?
3. L'emploi du *ne* de négation varie-t-il en fonction du profil des auteurs et de la région enquêtée ?

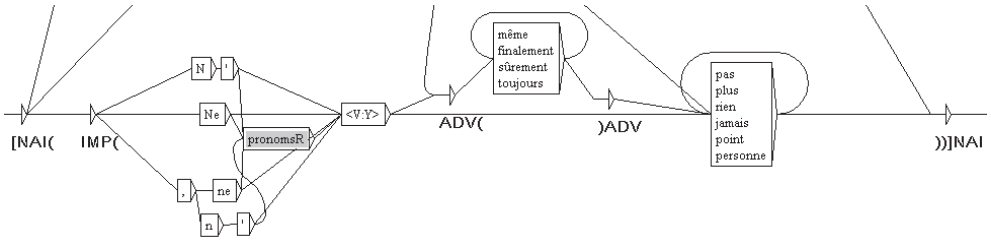
10.2.2 Variation interne

La présente étude porte sur trois des quatre corpus francophones à notre disposition étant donné, comme nous l'avons mentionné dans la partie 6.1.4.2, que l'équipe

québécoise a ajouté manuellement dans la transcription le *ne* de négation là où il faisait défaut. Il n'est dès lors pas possible de rechercher l'absence du *ne* dans le corpus transcrit.

Afin d'évaluer la structure des phrases négatives dans le corpus de sms, nous avons employé, comme dans la partie interrogative, des règles de repérage que nous avons élaborées avec le logiciel Unitex. Ces règles lexico-syntaxiques permettent d'extraire dans le corpus des structures négatives définies. La figure 10.11 présente un exemple de règles élaborées pour la reconnaissance d'une structure avec *ne* de négation et un verbe à l'impératif.

FIGURE 10.11 – *Extrait des règles lexico-syntaxiques d'identification des négatives*



Étant donné la méthode d'identification des négatives, nous reconnaissons tout aussi bien les emplois positifs de *rien*, *jamais*, etc. comme dans *as-tu rien compris ?*. Vu la complexité et l'emploi assez rare d'une telle structure, il est peu crédible d'en trouver un assez grand nombre pour que cela engendre un biais dans les données.

Ces règles ont été conçues pour distinguer les facteurs morphosyntaxiques suivants :
– la nature du forclusif[22] ;
– la nature et le mode du verbe ;
– la nature du sujet ;
– la présence d'un adverbe de renforcement ou le figement (formules stéréotypées).
Nous avons évalué les unités identifiées (présence/absence du *ne*) par les règles en observant un échantillon de 1.000 messages. La première évaluation concerne la reconnaissance des structures négatives. Le rappel montre que 90,5 % des formes intéressantes (messages comprenant une négative au moins) sont repérées[23]. La précision prouve que 98 % des messages repérés contiennent une négative[24]. La

22. Notons que nous avons mis de côté l'adverbe restrictif *que* qui entraine un trop grand biais dans la reconnaissance automatique. Le biais engendré lorsque l'on inclut cet adverbe étant plus important que celui provoqué par son élimination, nous avons pris le parti de supprimer *que* des règles d'extraction.

23. Les formes non reconnues sont en grande majorité des formes *pa* du corpus réunionnais, dont on ne sait si elles appartiennent au français de La Réunion (et auraient dû être transcrites *pas*), ou au créole. Elles ne font donc pas l'objet de cette étude.

24. Les formes mal reconnues sont surtout dues à l'ambigüité de certaines structures pour lesquelles un verbe est confondu avec un substantif et le syntagme avec une négative à l'impératif tel que dans l'exemple suivant : *Restaurant pizzeria ou* ***cuisine plus*** *gastro (resto a l'air top, mais prix plus élevé) ?*. Les erreurs peuvent également provenir de la confusion du forclusif avec le

F-mesure obtenue est de 94 %.

La seconde évaluation porte sur l'identification correcte de la distinction entre présence du *ne* et absence du *ne*. Le rappel prouve que 99 % des messages comprenant au moins une phrase négative avec *ne* sont repérés. Le calcul de la précision montre que 97 % des messages repérés sont correctement étiquetés (présence du *ne*). La F-mesure obtenue est de 98 %.

Ces résultats permettent de créditer les typologies et les statistiques basées sur les graphes de reconnaissance. Nous avons identifié 13.357 messages contenant des phrases négatives. Parmi les 13.357 messages identifiés, 3.296 comportent au moins une négative formée avec *ne* et 10.061 sms comportent une ou plusieurs négatives formées sans *ne*. Ces 13.357 messages contiennent 15.855 négatives dont 3.669 comportent le clitique *ne* (23,14%) [25] et 12.186 montrent une absence du *ne* (76,86%).

Le tableau 10.12 expose les chiffres relatifs aux variables principales que nous voulons analyser (nature du sujet, type d'adverbe de renforcement et type de forclusif) selon la construction de la négative. Il apparait distinctement que la proportion de sujets pronominaux est plus importante dans les négatives avec absence du *ne* (92% contre 71% avec *ne*) ; les autres types de sujet sont soit implicites (omis ou verbe à l'impératif) soit lexicaux. Le chiffre probablement le plus intéressant du tableau (la différence est fortement marquée) concerne la proportion des **sujets lexicaux** dans les deux structures : alors que cette proportion est de **3% dans les constructions sans la particule *ne***, elle atteint **26% dans les constructions en présence du *ne*** (ex. 98). Ces résultats suivent les conclusions émises par Meisner (2010) au sujet du français parlé [26].

(98) Bon **quelques bananes flambées ne** me feront **pas** de mal ! (smsFR)

Nous observons également que la proportion d'adverbes insérés dans la négative est plus importante dans les phrases négatives construites sans *ne* (2,49% contre 1,61% avec le clitique).
Le classement des adverbes est identique dans les deux types de construction : nous notons une prévalence de l'adverbe *même*, suivi de *toujours* (ex. 99). Les adverbes *surement* et *finalement* closent le classement. Leurs proportions ne sont toutefois pas semblables : si *toujours* est employé en proportion identique dans les deux constructions, *même* est proportionnellement sur-employé dans la construction

comparatif *plus* (*j'ai passé un excellent week-end probablement moins festif que toi mais chouette. Expliquerai plus dans mail*) ou du manque de ponctuation et de mots de liaison typiques de ce mode de communication (*on en a **envie rien** n'est sûr.*)

25. Ce pourcentage est plus élevé que la moyenne dans les études (relativement) récentes, telles que Ashby (2001) et sa moyenne de 16% de rétention chez les hommes comme chez les femmes.

26. « Clitic subjects hinder ne-realization while lexical [seem] to favor it » (Meisner, 2010, p. 1951).

FIGURE 10.12 – *Distribution de la négation en fonction des variables étudiées*

	Omission Ne		Présence Ne		Rétention Ne
	tokens	% total	tokens	% total	% total
Sujets					
pronom	11194	91,86%	2612	71,19%	18,92%
omis + imp	605	4,96%	88	2,40%	12,70%
lexical	387	3,18%	969	26,41%	71,46%
Adverbes total	303		59		16,30%
finalement	3	0,99%	2	3,39%	40,00%
même	184	60,73%	31	52,54%	14,42%
sûrement	9	2,97%	3	5,08%	25,00%
toujours	107	35,31%	23	38,98%	17,69%
Forclusifs					
pas	9968	81,80%	3092	84,27%	23,68%
plus	1398	11,47%	297	8,09%	17,52%
rien	532	4,37%	177	4,82%	24,96%
jamais	211	1,73%	81	2,21%	27,74%
personne	54	0,44%	10	0,27%	15,63%
point	1	0,01%	3	0,08%	75,00%
multiple	22	0,18%	9	0,25%	29,03%
Total négatives	12186		3669		23,14%

sans clitique (61% contre 53% avec clitique) alors que *finalement* (ex. 100) et *surement* apparaissent proportionnellement plus fréquemment en présence du *ne* (respectivement 3% et 5%, contre 1% et 3% sans clitique).

(99) Son natel **va toujours pas** ! Si jamais, dis lui que pour ce soir ça va être trop compliqué, et que j'essaierai de l'appeler ce soir ! (smsFS)

(100) Bonjour, **je ne prendrai finalement pas** l'appartement. J'ai trop besoin d'un balcon ! :-) Bonne chance et bon déménagement ! (smsFS)

Le classement des types de forclusifs est également similaire dans les deux constructions : *pas* et *plus* sont les plus fréquents, que la négative soit construite avec ou sans *ne*, suivis de *rien*, *jamais* et *personne*. Avec ou sans la particule *ne*, on remarque aussi que les forclusifs multiples (ex. 101) sont plus fréquents que les négatives avec *point*. Les proportions ne sont toutefois pas identiques : les résultats mettent en valeur un emploi plus important des forclusifs *pas* et *jamais* (proportionnellement aux autres forclusifs) dans les constructions avec *ne*[27]

27. Notons que ce pourcentage élevé du forclusif *pas* dans les constructions avec *ne* (23,68% de

et une utilisation plus importante de *plus* et *personne* dans les constructions sans *ne*.

(101) Tu m'as dit ça chaque année et ça n'a **jamais rien** changé. Enfin,je suis contente d'avoir compté à tes yeux, mais sache qu'il ne se passera **plus jamais rien** entre nous. J'aime mon copain plus que tout au monde ! (smsFB)

Suite à la présentation des chiffres concernant les variables *nature du sujet*, *type d'adverbe* et *type de foclusif*, nous détaillons la variation de la structure négative en fonction des formules dites « stéréotypées » ou « préformées ». Nous nous intéressons spécifiquement à la distribution des séquences proposées par Moreau (1986) [28] : *c'est pas*, *(il) faut pas*, *(je/tu) crois pas*, *(il) y a pas* et *(je/tu) sais pas*. Nous avons recensé 2.661 séquences préformées dans le corpus de sms, dont 309 sont construites avec le clitique *ne* et 2.352 sans. Le détail de ces séquences se trouve en Figure 10.13.

FIGURE 10.13 – *Distribution des séquences préformées en fonction du type de négatives*

	Omission (*ne*)	Présence *ne*	% présence *ne*
c'est pas	823	77	8,56%
faut pas	95	16	14,41%
(je/tu) crois pas	71	18	20,22%
(je/tu) sais pas	1111	170	13,27%
(il) y a pas	252	28	10,00%

On constate que **la proportion de séquences préformées construites avec *ne* (moyenne de 11,6%) est inférieure à la proportion générale de négatives construites sans le clitique**. La séquence *(je/tu) crois pas* est celle dont la différence de proportion entre les deux types de négatives est la plus faible. À l'analyse des 89 formes rencontrées, on remarque qu'il existe une large majorité de constructions de cette séquence avec une complétive introduite par *qu-* (ex. 102) dans le cas de la structure avec *ne* (66% des formes *(je/ne) crois pas*), alors que dans le cas des absences de *ne*, c'est la construction sans complétive (ex. 103) qui est la plus fréquente (55%).

(102) **je ne crois pas** que j'arriverai à aimer quelqu'un d'autre aussi fort, bisous (smsFB)

(103) Ben le dernier cours c'était mardi. Et si je redescends Saint Denis ça sera pour vider l'appart ou alors pour aller voir les résultats... mais pour rester même non **j'crois pas**. (smsFR)

La dernière variable intralinguistique que nous avons décidé d'étudier dans les phrases négatives concerne un ensemble de propriétés relatives au verbe : nature et

rétention) va à l'encontre des conclusions généralement émises à ce sujet (Dufter et Stark, 2007 ; Gadet, 1989).

28. Citée dans Dufter et Stark (2007, p. 7).

mode.

Nous commençons par une analyse manuelle des types et modes de verbes, ainsi que de l'insertion d'un pronom entre le sujet et le verbe, pour les constructions avec le clitique *ne* (3.669 négatives). Le résultat de cet étiquetage se trouve dans la Figure 10.14.

FIGURE 10.14 – *Distribution des types de verbes dans les constructions négatives avec clitique ne*

Verbe	tokens
Nombre savoir	323
Nombre oublier	289
Nombre aller	168
Nombre pouvoir	149
Nombre venir	128
Nombre vouloir	97
Nombre faire	78
Nombre voir	62
Nombre dire	57
Nombre hésiter	52
Nombre répondre	47
Nombre arriver	42
Nombre aimer	37
Nombre devoir	33
Nombre trouver	30
Nombre déranger	29
Nombre inquiéter	29
Nombre prendre	28
Nombre comprendre	25
Nombre falloir	25

Tokens Pronoms objets	659
Nombre t'	115
Nombre te	113
Nombre me	110
Nombre m'	93
Nombre y	93
Nombre l'	38
Nombre en	30
Nombre le	25
Nombre les	14
Nombre lui	8
Nombre nous	8
Nombre la	6
Nombre vous	6
% total Pronoms objets	18%

Modes	% total
Nombre impératif	14,00%
Nombre auxiliaires	12,00%
Nombre autres	74,00%

La Figure illustre trois ensembles d'informations. Tout d'abord, nous avons relevé les types de verbes différents ; nous avons répertorié 312 types. Nous avons choisi d'exposer les 20 verbes les plus fréquents en laissant de côté les verbes être (578 tokens) et avoir (424 tokens). Deuxièmement, nous avons repéré les structures négatives dans lesquelles un ou plusieurs pronoms objets se sont insérés (ex. 104) ; le deuxième tableau précise la nature et le nombre d'occurrences de ces pronoms

dans le corpus. Le taux de négatives comprenant un pronom s'élève à 18%. Enfin, nous avons relevé les modes principaux dans l'étude de la négation : impératif (14%), participe avec auxiliaire (12%) et autres (74%).

(104) Fais de beaux rêves de nous, on aura aussi un jour notre maison ! J'espère dans pas trop longtemps... Je t'aime même si tu ne me le dis plus ! (smsFB)

Étant donné l'étendue du sous-corpus de négatives construites sans le clitique *ne*, nous n'avons pas envisagé un étiquetage manuel de celui-ci. Nous approfondissons les constructions des 10 verbes les plus fréquents du sous-corpus de négatives construites avec *ne* en comparant leurs occurrences avec celles présentes dans le corpus de négatives construites sans *ne*. Le résultat de cet étiquetage se trouve dans la Figure 10.15.

La colonne intitulée *% Rétention ne* reprend le pourcentage des variables étudiées pour chaque verbe construit avec le clitique, par rapport au nombre total de constructions sur base de ce verbe (tokens) et de cette variable (*pronom*, *auxiliaire* et *impératif*). La sous-colonne *tokens* illustre la répartition d'un verbe. Ainsi, il nous est possible de constater que les verbes *oublier* et *hésiter* sont majoritairement construits avec le clitique (51,89% et 74,29%) comme l'illustrent les exemples 105 et 106. Nous interprétons ce phénomène comme étant probablement influencé par les initiales vocaliques de ces verbes.

(105) Hey ! J'ai été voir les Indestructibles ! C'est tout simplement GÉNIAL ! Bonne fin de semaine et **n'oublie pas** ta bougie demain... Bisous ! (smsFS)

(106) Bonne soirée à toi aussi, si tu veux te confier ou avoir quelques conseils **n'hésite pas**. Bisous (smsFR)

Les sous-colonnes *pronom*, *auxiliaire* et *impératif* montrent, pour un verbe donné, si la variable apparait plutôt dans une construction avec ou sans *ne*. La sous-colonne *pronom* ne présente aucune prépondérance de la construction avec *ne*. La sous-colonne *auxiliaire* illustre la part importante de négatives construites avec *ne* et *avoir* pour les verbes *vouloir* et *dire*. L'influence de l'auxiliaire *avoir* en initiale vocalique peut sans doute expliquer cette tendance ; ces résultats viennent corroborer des conclusions émises par ailleurs au sujet des auxiliaires (Gadet, 1989).

Enfin, la sous-colonne *impératif* met en avant les pourcentages plus élevés des constructions négatives à l'impératif avec l'élément *ne* pour les verbes *oublier*, *venir*, *dire* et *hésiter* (exemples 105, 106, 107 et 108).

(107) Ben écoute **je ne suis pas venue** ce matin parce que j'ai pas dormi du tout de la nuit à cause d'un dossier. (smsFR)

(108) Dans ton SMS 12h **tu n'as pas dit** t'étais libre ce soir ! ! Demain soir va pas ? (smsFB)

Enfin, l'analyse des types de forclusifs pour les 10 verbes étudiés (Figure 10.16) ne permet pas de soutenir les thèses émises par ailleurs (Gadet, 1989 ; Van Compernolle, 2007). En effet, le forclusif *pas* apparait à 25,33% dans les négatives avec le clitique *ne*, ce qui constitue un taux plus important, pour ces verbes, que celui des constructions avec *rien* (18,32%), *plus* (21,43%) et *jamais* (21,84%), par exemple.

FIGURE 10.15 – *Distribution des variables étudiées (pronom, auxiliaire et impératif) pour l'étude de la négation sur base des 10 verbes les plus fréquents dans les négatives en ne*

	Omission *ne*				Rétention *ne*				% Rétention *ne*			
Verbe	tokens	pronom	auxiliaire	impératif	tokens	pronom	auxiliaire	impératif	tokens	pronom	auxiliaire	impératif
Nombre savoir	1456	34	38	0	323	6	13	0	18,16%	15,00%	25,49%	0,00%
Nombre oublier	268	77	30	101	289	36	26	220	**51,89%**	31,86%	46,43%	**68,54%**
Nombre aller	551	28	9	3	168	10	1	2	23,37%	26,32%	10,00%	40,00%
Nombre pouvoir	574	55	59	0	149	5	12	0	20,61%	8,33%	16,90%	0,00%
Nombre venir	242	18	23	3	128	2	10	8	34,59%	10,00%	30,30%	**72,73%**
Nombre vouloir	402	67	4	0	97	2	8	0	19,44%	2,90%	**66,67%**	0,00%
Nombre faire	323	110	45	42	78	26	20	14	19,45%	19,12%	30,77%	25,00%
Nombre voir	225	90	60	0	62	24	24	0	21,60%	21,05%	28,57%	0,00%
Nombre dire	208	106	15	10	57	28	18	13	21,51%	20,90%	**54,55%**	**56,52%**
Nombre hésiter	18	0	0	7	52	0	0	43	**74,29%**	0,00%	0,00%	**86,00%**

FIGURE 10.16 – *Distribution de la variable forclusif pour l'étude de la négation sur base des 10 verbes les plus fréquents dans les négatives en ne*

	Omission *ne*						Rétention *ne*						% Rétention *ne*					
	pas	rien	plus	jamais	personne	composé	pas	rien	plus	jamais	personne	composé	pas	rien	plus	jamais	personne	composé
Nombre savoir	1305	21	102	27	0	1	281	9	23	10	0	0	17,72%	30,00%	18,40%	27,03%	0,00%	0,00%
Nombre oublier	258	3	0	7	0	0	284	2	1	2	0	0	**52,40%**	**30,00%**	**100,00%**	22,22%	0,00%	0,00%
Nombre aller	496	6	39	10	0	0	155	0	11	2	0	0	23,81%	0,00%	22,00%	16,67%	0,00%	0,00%
Nombre pouvoir	464	24	82	4	0	0	124	3	21	1	0	0	21,09%	11,11%	20,39%	20,00%	0,00%	0,00%
Nombre venir	236	0	5	1	0	0	125	0	2	1	0	0	34,63%	0,00%	28,57%	50,00%	0,00%	0,00%
Nombre vouloir	340	12	41	5	0	4	80	0	15	1	0	1	19,05%	0,00%	26,79%	16,67%	0,00%	20,00%
Nombre faire	246	55	12	9	0	1	66	6	5	1	0	0	21,15%	9,84%	29,41%	10,00%	0,00%	0,00%
Nombre voir	187	4	25	3	2	4	51	3	6	0	1	1	21,43%	**42,86%**	19,35%	0,00%	33,33%	20,00%
Nombre dire	167	31	2	2	0	6	**44**	12	0	0	0	1	20,85%	27,91%	0,00%	0,00%	0,00%	**14,29%**
Nombre hésiter	18	0	0	0	0	0	51	0	0	1	0	0	**73,91%**	0,00%	0,00%	**100,00%**	0,00%	0,00%
Total	3717	156	308	68	2	16	1261	35	84	19	1	3	**25,33%**	18,32%	21,43%	21,84%	33,33%	**15,79%**

10.2.3 Variation externe

Les **messages du corpus de sms contenant des négatives et liés à un profil sociodémographique** sont au nombre de 10.414. Parmi ceux-ci, 2.731 présentent la rétention du *ne* (au moins une occurrence), le reste des messages incluent uniquement des messages contenant l'absence de *ne*.

Partant de l'hypothèse énoncée par Pohl (1972, p. 107) selon laquelle « les taux d'omission [du *ne*] varient de zéro à cent selon les locuteurs et les discours, soumis à d'innombrables facteurs qui peuvent se renforcer ou se neutraliser », nous allons nous pencher sur la question de la variation des phrases négatives en fonction de la région et du profil des auteurs de sms. Cette question parait coïncider avec notre corpus et notre méthodologie puisque, comme l'ont expliqué Armstrong et Smith (2002, p. 28), « une large proportion de la variation dans les niveaux de rétention du *ne* exposée [dans la plupart des recherches européennes sur le français] est explicable par les caractéristiques sociales des locuteurs enquêtés ainsi qu'à la situation de communication dans laquelle le corpus a été enregistré » [29]. Ce n'est pas le cas pour cette enquête, qui a tenté d'interroger des populations semblables dans chaque région. C'est pourquoi nous entamons d'emblée une étude de la variation à travers les régions, afin de répondre, notamment, à des questions restées en suspens dans la littérature, telles que le fait que le taux d'absence paraisse « très élevé à Paris et au Canada, [il] est moins haut dans le Pas-de-Calais, moyen en Lorraine rurale et faible ou très faible en Wallonie » (Pohl, 1972, p. 108). Si la tendance découverte dans les années 1970 est toujours d'actualité dans les années 2000, nous observerons une tendance significative dans le test appliqué.

Le test khi-carré porte sur le lien de dépendance entre la région de francophonie observée [30] et la rétention du *ne* de négation. Le résultat du test (la valeur p est égale à 0, détaillé en Annexe A.115) indique qu'il semble y avoir dépendance entre les deux variables. Les résidus standardisés ajustés de la Figure 10.17 permettent de détailler cette conclusion. Nous trouvons effectivement **une large sous-représentativité (-9,5) des auteurs belges de messages utilisant la négative avec absence du *ne*** (NAI). De la même façon, **le taux de messages réunionnais montrant la rétention du *ne* (AI) est inférieur à ce à quoi on aurait pu s'attendre** si la distribution avait été équilibrée. Les chiffres concernant la Suisse sont, eux, moins significatifs, et ne permettent pas de conclusion.

Étudier les liens de dépendance entre l'âge et la rétention du *ne* de négation ne constitue pas une tâche anodine étant donné les assertions que l'on peut trouver

29. « the rather large degrees of variation in the levels of *ne* retention reported [in most European French findings] are due to the social characteristics of the speakers sampled as well as the speech situations in which the corpora were recorded ».

30. À ce stade, nous regrettons fortement la non accessibilité aux données québécoises pour le même phénomène, étant donné les précédents résultats (Sankoff et Vincent, 1977) qui ont montré le très haut taux d'absence de *ne* dans cette région.

FIGURE 10.17 – *Analyse des résidus standardisés relatifs au khi-carré (types de négatives - région)*

Tableau croisé Neg * Région

			Région			Total
			BE	CH	LR	
Neg	AI	Effectif	5938	699	1046	7683
		Résidu ajusté	-9,5	,9	11,2	
	NAI	Effectif	2343	233	155	2731
		Résidu ajusté	9,5	-,9	-11,2	
Total		Effectif	8281	932	1201	10414

dans la littérature à ce propos et qui considèrent « que l'âge est de loin le paramètre le plus important » [31] (Coveney, 1989, p. 124) à prendre en compte dans l'étude de la rétention du clitique *ne*. Avant de tester le rapport de dépendance entre les variables *âge* et *type de négation* sur l'ensemble de la population étudiée, nous allons d'abord nous focaliser sur l'hypothèse suivante : « Dans tous les milieux, semble-t-il, le français des enfants se caractérise par un taux d'absence particulièrement haut. On a même des raisons de penser que ce taux est toujours de cent [...] jusqu'à un certain âge que je situerai très approximativement aux environs de trois ans ou même un peu plus » (Pohl, 1972, p. 108). Nous ne disposons pas de message d'auteurs aussi jeunes mais nous pensons qu'un aperçu des effectifs pour les auteurs les plus jeunes de notre corpus peut nous aider à évaluer une possible évolution de l'absence à travers les âges. Nous analysons donc la distribution du type de négative pour la population âgée de 11 à 17 ans (Figure 10.18). La Figure montre effectivement une évolution constante dans le nombre de sms contenant des phrases négatives avec *ne* à travers les âges malgré une distribution de messages à travers les âges non constante (voir par exemple le pic de messages d'auteurs de 16 ans employant des négatives).

En ce qui concerne l'ensemble des auteurs, nous appliquons le test t aux valeurs regroupées par catégories [32]. Le résultat du test (dont le détail se trouve en Annexe A.117) montre une valeur p = 0, ce qui signifie qu'il y a dépendance entre les deux variables. Le box-plot présenté à la Figure 10.19 montre qu'il existe une répartition des âges plus étendue pour les sms comprenant une négative avec *ne* (NAI) que pour les messages comprenant une négative avec absence de *ne* (AI). Nous pouvons également conclure, au vu de la dispersion autour de la médiane pour les négatives avec absence de *ne* que **la distribution des sms comprenant des négatives avec absence de *ne* concerne une classe d'âge plus restreinte (les valeurs sont plus ramassées) et qui se situe surtout en deçà de 22 ans.**

31. « That age is by far the most important parameter ».

32. Pour rappel, les catégories d'âge sont les suivantes : 1 (10 à 15), 2 (16 à 17), 3 (18 à 19), 4 (20 à 22), 5 (23 à 27), 6 (28 à 40) et 7 (41 à 70). Ces catégories permettent une distribution dite « normale » (Annexe A.116).

FIGURE 10.18 – *Distribution des types de négatives en fonction des âges (11 à 17)*

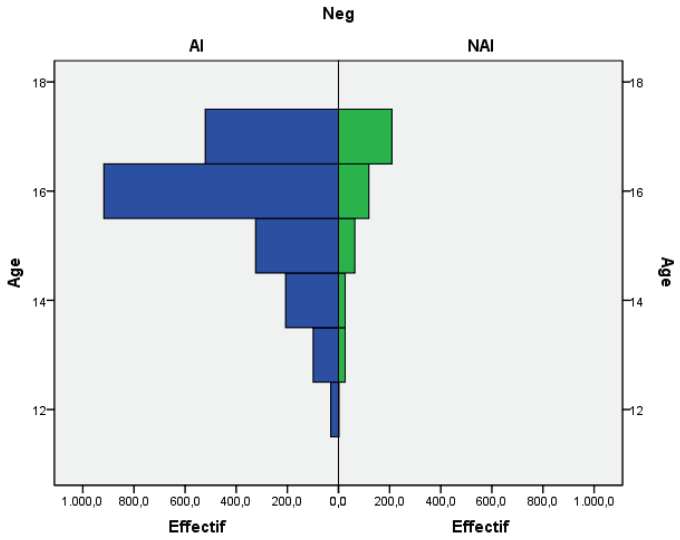
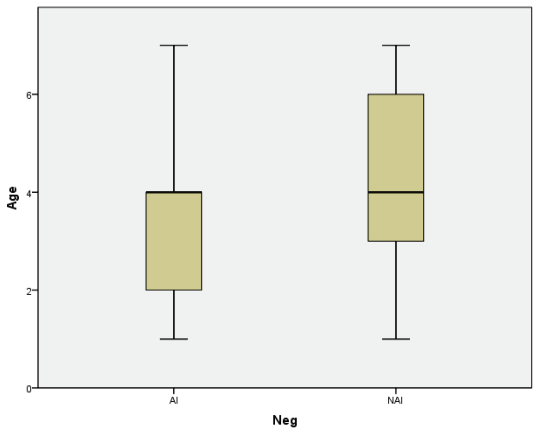


FIGURE 10.19 – *Box-plot présentant la répartition des types de négatives en fonction des catégories d'âge*



Le tableau des effectifs croisés pour les variables *âge* et *type de négatives* (Figure 10.20) illustre la constante progression du pourcentage de messages comprenant une négative construite avec *ne* par rapport au nombre total de messages comprenant des négatives. Ce tableau est d'autant plus intéressant que l'on remarque, pour la catégorie des plus de 40 ans, un plus grand nombre de messages comprenant une négative avec clitique que sans (51,8%).

La variation de l'emploi des négatives en fonction du *sexe* est également importante. En effet, selon Coveney (1989, p. 124) : « il ne semble pas que le sexe du locuteur soit un facteur qui permette une large différentiation de la variation » [33].

33. « It does not seem that the sex of the speaker is a very significant differentiating factor in

FIGURE 10.20 – *Effectifs croisés pour les variables âge et type de négative (AI, sans ne ; NAI, avec ne)*

Tableau croisé Neg * Age

			Age							Total
			1	2	3	4	5	6	7	
Neg	AI	Effectif	660	1439	1667	2198	1047	416	256	7683
		% compris dans Age	84,8%	81,4%	81,0%	76,3%	66,9%	49,8%	48,2%	73,8%
	NAI	Effectif	118	328	391	683	517	419	275	2731
		% compris dans Age	15,2%	18,6%	19,0%	23,7%	33,1%	50,2%	51,8%	26,2%
Total		Effectif	778	1767	2058	2881	1564	835	531	10414
		% compris dans Age	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

L'application du test khi-carré sur l'ensemble du corpus (dont le détail se trouve en Annexe A.118) montre pourtant que dans le contexte des sms, le sexe de l'auteur semble être un facteur de variation. Le détail des résidus standardisés (Figure 10.21) permet d'interpréter cette dépendance des variables : **les hommes auteurs de sms du corpus montrent une sous-représentation de l'emploi de négatives sans le clitique *ne*** (AI) **alors que les femmes, au contraire, montrent un sous-emploi des constructions de phrases négatives avec *ne*** (NAI).

FIGURE 10.21 – *Analyse des résidus standardisés relatifs au khi-carré (types de négatives - sexe)*

Tableau croisé Neg * Sexe

			Sexe		Total
			F	M	
Neg	AI	Effectif	5079	2604	7683
		Résidu ajusté	5,5	-5,5	
	NAI	Effectif	1644	1087	2731
		Résidu ajusté	-5,5	5,5	
Total		Effectif	6723	3691	10414

La quatrième variable que nous désirons approfondir est la langue première. En effet, doit-on s'attendre à ce que les francophones purs montrent une plus grande rétention que des allophones ou bilingues qui font preuve de variation linguistique ? Ou, au contraire, peut-on s'imaginer que les allophones et bilingues souffrent d'hypercorrectisme [34] et qu'ils présentent un plus haut taux de rétention du *ne* que les francophones natifs purs, qui se laissent aller à la spontanéité de la communication efficace (rapport cout - réception) ?

Le résultat du test khi-carré (Annexe A.119) présente une valeur p inférieure à 0,05, ce qui signifie que les variables *rétention du ne* et *langue première* sont liées. L'observation des résidus standardisés ajustés de la Figure 10.22 permet d'affiner le

the variability ».

34. Il s'agit d'une « pratique qui va au-delà de ce que recommande la norme évaluative » (Klinkenberg, 1999, p. 54) où la norme évaluative du sms consisterait en un laisser-aller à l'absence du *ne* dans un contexte où cette absence permet d'abréger le message.

résultat du test statistique en montrant à quel endroit et dans quel sens a lieu cette dépendance. On remarque que **le taux de messages provenant de locuteurs de langue première française uniquement (0) et employant des négatives avec absence du *ne* (AI) est inférieur à celui auquel on aurait pu s'attendre** si la distribution avait été équilibrée.

FIGURE 10.22 – *Analyse des résidus standardisés relatifs au khi-carré (types de négatives - langue première)*

Tableau croisé Neg * Langue_Maternelle

			Langue_Maternelle		
			0	1	Total
Neg	AI	Effectif	6735	948	7683
		Résidu ajusté	-6,5	6,5	
	NAI	Effectif	2519	212	2731
		Résidu ajusté	6,5	-6,5	
Total		Effectif	9254	1160	10414

C'est donc la première hypothèse qui est validée : les francophones purs montrent une plus grande rétention du *ne* que les allophones et bilingues qui s'écartent généralement et plus naturellement de la norme standard, d'une langue spontanément plus surveillée (chez les francophones).

Avant de clôturer ce tour d'horizon diastratique, nous décidons de porter notre attention sur les liens qui unissent *absence du ne de négation* et *type d'occupation professionnelle*. Pour rappel, les types d'occupation ont été regroupés en 4 catégories : 1. sans emploi, élèves et étudiants 2. prestation de services 3. industrie 4. artisanat.

Le résultat du test khi-carré (la valeur p est égale à 0) se trouve en Annexe A.120 et montre une dépendance entre l'occupation professionnelle d'un auteur de message et son type de négation (absence ou non du *ne*). Les résidus standardisés (Figure 10.23), pour leur part, précisent la nature de cette dépendance : le nombre de messages envoyés par des auteurs de sms sans emploi, élèves et étudiants employant le clitique *ne* est très inférieur (-12,5) à ce que l'on aurait pu escompter si la distribution avait été équilibrée. Au contraire, les messages émis par la population appartenant à 2. prestation de service et 3. industrie montrent un taux d'absence du clitique supérieur à celui auquel on aurait pu s'attendre (respectivement -10,9 et -6,6). Curieusement, on ne peut rien conclure à propos de la catégorie 4. artisanat.

FIGURE 10.23 – *Analyse des résidus standardisés relatifs au khi-carré (types de négatives - occupation professionnelle)*

Tableau croisé Neg * occupation

			occupation				Total
			1	2	3	4	
Neg	AI	Effectif	7158	380	63	82	7683
		Résidu ajusté	12,5	-10,9	-6,6	-1,2	
	NAI	Effectif	2328	299	67	37	2731
		Résidu ajusté	-12,5	10,9	6,6	1,2	
Total		Effectif	9486	679	130	119	10414

10.2.4 Conclusions

L'étude syntaxique de la négation que nous avons menée à terme a permis de déceler dans le corpus de sms 15.855 phrases négatives dont 23,14% comportent le clitique *ne* et 76,86% montrent une absence du *ne*. Ce taux de 23,14% de rétention peut être comparé à celui de Panckhurst (1998) (90%), pour la fin des années 1990 dans le courrier électronique et celui de Van Compernolle (2007), pour le chat en 2005 (15%) ; le corpus de sms se situe donc entre les deux tendances, bien qu'il soit plus proche de la tendance du chat. Le taux est probablement davantage plus élevé que celui du courrier électronique en raison de la date de collecte (1997) ; en comparaison au corpus de chat, nous pensons que la différence tient plutôt du médium de communication.

Selon l'analyse, la nature du sujet apparait comme une variable significative ; en effet, 3% seulement des sujets dans les constructions sans clitique sont lexicaux, alors que ce taux atteint 26% dans les constructions en présence du *ne*. Cette conclusion est comparable à celle émise au sujet du français parlé (Meisner, 2010). Les adverbes, quant à eux, montrent un classement fréquentiel identique dans les deux constructions (*même*, *toujours*, *surement*, *finalement*) mais ne semblent pas jouer en faveur de la rétention du *ne* de négation (2,49% sans clitique contre 1,61% avec clitique).

Contrairement à ce qui est souvent indiqué dans la littérature, le forclusif *pas*, bien que montrant un nombre d'occurrences plus élevé dans les négatives sans clitique, n'est proportionnellement pas moins présent dans les négatives avec *ne* que les autres forclusifs (84,27% de *pas* dans les négatives avec clitique contre 81,80% dans les négatives sans clitique).

Nous avons accordé une attention particulière aux séquences préformées ; celles-ci apparaissent en moyenne dans 88% des cas, en construction sans *ne*. Cette tendance tend à rapprocher le corpus sms de l'immédiat communicatif (Koch et Oesterreicher, 2001) [35].

Nous nous sommes penchée sur l'environnement verbal des négatives. Nous avons décortiqué les 10 verbes les plus fréquents dans la construction avec *ne*. Ceci a permis d'identifier des verbes qui sont majoritairement employés dans une construction avec le clitique *ne* (*oublier* et *hésiter*) et ce, malgré la disproportion des fréquences des deux types de structures de négatives (23% avec rétention du *ne* contre 77% sans le clitique) dans le corpus. Ce phénomène, comme celui de l'influence de l'auxiliaire *avoir* sur l'emploi du clitique, pourrait s'expliquer par l'initiale vocalique qui motiverait l'emploi du *ne*.

35. « Les séquences 'toutes faites' font partie des traits caractéristiques de l'immédiat communicatif » (Koch et Oesterreicher, 2001, p. 599), cité dans Dufter et Stark (2007, p. 7).

Enfin, le mode du verbe semble également influencer l'usage du *ne* de négation puisque nous avons noté des pourcentages plus élevés des constructions négatives à l'impératif avec l'élément *ne* pour des verbes tels que *oublier*, *venir*, *dire* et *hésiter*.

Selon l'analyse extralinguistique que nous avons menée, la rétention du *ne* est un phénomène qui semble dépendant de nombreux facteurs externes : sexe, âge, langue première, région, etc. Nous avons prouvé la variation des structures négatives en fonction de l'âge, y compris une évolution marquée au niveau de l'adolescence.

Les résultats relatifs au sexe du locuteur sont éloquents si l'on considère les recherches antérieures à ce sujet : la distribution des sexes pour la variable *absence du ne de négation* montre une disproportion en faveur des femmes. Dans notre cas, ces résultats semblent logiques. En effet, nous avons montré, pour la majorité des phénomènes étudiés (lexicaux et syntaxiques), que les femmes se permettaient plus volontiers des variations par rapport à un standard, une plus grande liberté par rapport à ce qui répond, dans la langue, à une norme sociale. Les résultats concernant la non-rétention du *ne* de négation chez les femmes dans le corpus de sms viennent donc confirmer cette tendance générale.

Les tests statistiques appliqués aux variables langue première et situation professionnelle ont également apporté des résultats significatifs : les locuteurs de langue première française (uniquement) semblent plus habitués à moins utiliser le *ne* de négation que les bilingues et allophones. De la même manière, les messages émis par la population appartenant à la classe des services et de l'industrie montrent un taux d'absence du clitique supérieur à celui auquel on aurait pu s'attendre, alors que les étudiants et sans emploi font plutôt preuve de l'absence.

Enfin, d'après l'étude diatopique de la négation dans le corpus de sms, il existerait une forte tendance des Belges à utiliser le clitique, tandis que les Réunionnais, pour leur part, l'utilisent moins volontiers.

10.3 Conclusions

Lorsque nous avons abordé cette partie consacrée à la syntaxe dans les sms, nous avons énuméré une série de questions issues de la littérature que nous voulions confronter aux données sms. Nous reprenons ici chacune de ces questions ainsi que les réponses que nous leur avons apportées.

– Comment se manifestent les interrogatives dans les sms ? Quelle répartition des formes observe-t-on ?

Les interrogatives suivies d'un point d'interrogation apparaissent dans un cinquième des messages du corpus. Ce chiffre est relativement bas et on peut donc par exemple éloigner la fonction de la communication par sms de celle de l'appel téléphonique. Nous avons distingué 15.510 structures interrogatives dont 11% sont des interrogatives que nous avons nommées « formelles ». Grâce à notre analyse semi-automatique, il nous a été possible de dresser un classement fréquentiel des structures interrogatives rencontrées :

1. 55% : interrogatives par point d'interrogation (le pendant de l'intonation à l'oral) ;
2. 21% : interrogatives par morphème interrogatif avec et sans inversion ;
3. 8% : interrogatives avec ellipse du verbe ;
4. 7,5% : interrogatives moins formelles diverses : disjonctives, régionales, etc. ;
5. 5% : interrogatives par syntagme interrogatif *est-ce que* ;
6. 3% : interrogatives par inversion.

Nous avons montré que ce classement ne suit pas les tendances évoquées pour les conversations non soutenues. L'indicatif présent est le temps majoritairement employé dans les interrogatives formelles. Les interrogatives par inversion se construisent souvent dans les sms avec les verbes *avoir*, *pouvoir*, *être*, *venir*, *vouloir* et *savoir*. En outre, nous avons constaté la supériorité numérique dans le corpus de la construction basée sur le syntagme interrogatif *est-ce que* sous la forme *est-ce que ça va ?*.

– Les formes interrogatives moins formelles sont-elles le fait d'une population en particulier ?

Les tests statistiques ont mis en avant le lien inexistant entre le sexe d'un auteur et sa pratique de l'interrogative. Par contre, nous avons conclu que les interrogatives moins formelles étaient préférées par les locuteurs bilingues ou allophones ainsi que par les élèves du secondaire supérieur. D'un point de vue diatopique, le corpus a prouvé la propension de locuteurs belges et québécois à user de formes interrogatives formelles.

– Quelles structures de négation courante peut-on relever dans les sms ?

23% des structures négatives que nous avons identifiées présentent une rétention du *ne* de négation. Ce résultat est plus élevé que celui qu'on peut trouver dans le chat, par exemple.

– Quelles sont les variables internes et externes de la variation de cette négation ? Parmi les variables intralinguistiques influant sur la pratique de la négation, nous recensons le type de sujet (26% de constructions négatives avec sujet lexical pour la rétention du *ne* contre 3% pour l'absence du *ne*), le figement (les séquences préformées apparaissent dans 88% des cas sans le clitique, ce qui rapproche les critères définitoires de notre corpus de ceux de l'immédiat communicatif), le type de verbe et l'auxiliaire (les verbes à initiale vocalique semblent influer en faveur d'une négative avec clitique) et le mode du verbe (par ex. l'impératif influe sur l'emploi du clitique avec les verbes *oublier*, *venir* et *hésiter*). *A contrario*, le type de forclusif ne semble pas constituer un critère de variation dans le corpus de sms.

Enfin, l'étude des variables extralinguistiques a révélé des dépendances entre le type de négation (avec ou sans clitique) et toutes les variables étudiées (sexe, âge, langue première, région). L'analyse de l'âge, en particulier, présente une évolution marquée, y compris dans un laps de temps court au niveau de l'adolescence. Les résultats concernant le sexe sont plus marquants étant donné qu'ils contredisent les tendances évoquées dans la littérature. Il semblerait que, dans le corpus de sms, les femmes se permettent, presque pour tous les phénomènes étudiés, une plus grande liberté par rapport aux formes soutenues.

Quatrième partie

Conclusions

Chapitre 11

Conclusions générales de la recherche

L'avenir de la communication est inscrit dans le développement accéléré des technologies de l'information et de la communication. Cette évolution altère le discours de l'Homme ainsi que son comportement. Il est donc naturel que devant cette révolution en marche, chacun – citoyens, médias, parents, politiques – s'interroge sur les conséquences linguistiques et sociales de cette révolution. C'est au monde scientifique qu'il revient de répondre à ces interrogations, notamment au travers d'enquêtes.

Au terme de notre travail de recherche, de constitution de corpus et d'analyse des données, nous pensons nous inscrire pleinement dans ce que Calvet et Dumont (1999, page 11) nommaient « l'enquête sociolinguistique » et dont les questions de base sont les suivantes :

1. Qu'est-ce que les gens parlent ?
2. Comment les gens parlent-ils ?
3. Quelles sont les différences (géographiques, sociales, etc.) entre leurs façons de parler ?

En effet, à la première question *qu'est-ce que les gens parlent ?*, nous pouvons répondre qu'il existe une forme de communication que l'on nomme désormais *écrit sms* et qui ne coïncide pas avec un nouveau *langage* (un « langage sms »), ni avec une nouvelle langue, mais avec une nouvelle pratique écrite, une nouvelle transcription de la langue. Cette pratique se distingue des modes de communication plus traditionnels et plus anciens tels que la lettre, le billet et le télégramme parce que, notamment, elle n'est pas le fait d'une élite privilégiée. Contrairement à la carte postale et à la lettre, elle s'inscrit dans un acte de l'instant, de rapidité. Néanmoins, la pratique de l'eSMS partage également un certain nombre de traits avec les écrits plus traditionnels : des formules d'adresse (avec la lettre), une exigence de brièveté imposée par le support de communication (avec la carte postale), des tournures ludiques (avec le billet), une syntaxe réduite à l'essentiel, avec absence de sujet ou de mots-liens (avec le télégramme). La pratique de l'eSMS s'inscrit dans un ensemble plus vaste, que nous avons nommé la *communication écrite médiée par ordinateur*, un ensemble de formes de communication écrites et dont le véhicule est l'*ordinateur* (compris dans son acception la plus large). Les différents types de la Cémo partagent avec l'eSMS de nombreux traits, parmi lesquels le but de

la communication (discussion interpersonnelle, chat et MI), l'asynchronicité des échanges (le courriel) et des espaces de rédaction restreints (profils de SRS). Les traits qui distinguent l'eSMS des autres types de la Cémo sont notamment le système d'encodage (par rapport à tous les autres types qui emploient un clavier plus large, aucun système de pression multiple et peu de systèmes de prédiction), l'existence d'un péritexte important et l'absence de norme d'inscription préalable (avec le courriel), la fixité d'une thématique (avec le forum) et l'objectif de vitrine pour un collectif (avec le SRS).

L'écrit sms est caractérisé par une série de procédés graphiques combinables : des abréviations graphiques avec incidence phonique (syncope, apocope, aphérèse et spécialisation de caractères) ou sans incidence phonique (phonétisation par la lettre, le chiffre, le signe et des graphies plus proche d'un phonétisme réel ou d'un style), les émoticônes, la ponctuation expressive (redoublement et combinaison de ponctuants) et les blancs graphiques (réduction de la ponctuation et suppression des blancs). Nous refusons de considérer cette pratique comme de l'oral dans l'écrit, du français ordinaire ou de la production simplement spontanée. En termes de paramètres communicatifs, l'eSMS se situe sur le continuum allant de la *Nähesprache* (la proximité communicative) à la *Distanzsprache* (distance communicative) ; selon la nature des données sms observées, on considère que cette production se situe davantage vers le pôle de la Nähesprache. Enfin, ce genre, que Biber et Conrad (2009, p. 177) nomment *registre*, présente des propriétés qui vont à l'encontre de quelques idées reçues : il s'agit bien d'une communication privée, mais dont l'interlocuteur n'est pas forcément intime et l'émotionnalité n'est pas systématiquement forte.

À la deuxième question posée par Calvet et Dumont (1999), *Comment les gens parlent-ils ?*, nous répondons à l'aide des résultats obtenus dans les parties analytiques de ce travail. L'analyse lexicométrique a par exemple prouvé que la tendance à l'abréviation chez les auteurs des messages n'était pas particulièrement marquée : les moyennes du nombre de caractères dans le corpus brut et dans le corpus transcrit ne sont pas très différentes (3,62 sur corpus brut pour 3,82 sur corpus transcrit). Cette analyse a également montré que le lexique employé dans le corpus de sms était moins diversifié (rapport hapax/token de 1,5%, rapport type/token de 3,2%) que celui employé dans un corpus journalistique de la même ampleur pour la même période (rapport hapax/token de 2,6%, rapport type/token de 5,9%). Le décompte grammatical des 1.000 unités les plus fréquentes du corpus parait tout à fait différent de celui que l'on peut trouver pour d'autres types de corpus (par exemple pour le français parlé) : c'est le suremploi des verbes (notamment par rapport aux substantifs) qui a particulièrement marqué notre lecture des résultats.

L'étude néologique que nous avons menée a mis en avant un phénomène courant de créativité lexicale chez les auteurs de sms (3,2% des messages sont concernés). Pourtant, les deux tiers de ces néologismes sont des hapax ; ceci est probablement l'indice

du contexte ludique qu'offrent les sms, un contexte qui incite (ou du moins n'inhibe pas) la création (on se souviendra de *soupercommuter* et *bisounuit*). Le procédé de création le plus répandu est, sans surprise, la dérivation, suivie de la conversion. Plus de la moitié de ces créations sont des substantifs. Les unités néologiques récurrentes concernent principalement les nouvelles technologies (*smser*, *chatter*, *e-mailiser*, etc.) et les traits humains (*footeux*, *grunge*, *loveuse*, etc.). La néologie est également marquée par la variation régionale : on ne trouvera pas les unités *sms* et *texto* en concurrence dans toutes les régions (pas de *texto* en Suisse) et la forme *courriel* est attestée, comme on aurait pu s'y attendre, uniquement au Québec.

Les auteurs des messages du corpus emploient des éléments allogènes et des régionalismes dans un nombre de message équivalent (approx. 10% des messages pour chaque phénomène). Ce résultat va à l'encontre des idées reçues au sujet de la langue employée dans les nouveaux médias, selon lesquelles on y retrouverait des mélanges de langues anarchiques et abondants. De plus, l'analyse discursive a montré que peu de ces mélanges s'identifiaient à du code-mixing, c'est-à-dire à un mélange qui ne respecte pas les règles grammaticales des langues en présence. Par contre, l'anglo-américain est bien la langue préférée des auteurs de sms : 90% des éléments allogènes identifiés appartiennent à cette langue et remplissent plutôt des fonctions stylistiques (*faire cool*). L'emploi des autres langues, particulièrement les langues romanes, semble satisfaire un besoin d'expression de sentiments et d'émotions.

L'étude de la variation orthographique que nous avons menée sur le corpus belge a permis de déceler une prépondérance des variantes de nature morphogrammique (55%) et des écarts dus à des homophones (34%). Au sein de cette classe de variantes morphogrammiques, les variantes portant sur les morphèmes grammaticaux prévalent : les fautes d'accord sujet-verbe et la confusion des modes et des temps constituent les deux ensembles les plus importants. Nous avons identifié 1.773 variantes orthographiques non stimulées par des motivations d'abréviation ; ce chiffre peut paraitre élevé. Toutefois, lorsque nous relevons, pour quatre catégories de variantes en particulier, toutes les formes correctement orthographiées, nous concluons que le taux de formes erronées est faible : 0,26% pour l'accord sujet-verbe à l'indicatif présent (1-2^{e} p. sg), 2,70% pour la confusion des modes impératif - indicatif (2^{e} p. sg), seulement 5,5% pour la confusion des modes participe passé - infinitif et également 5,5% pour la confusion des temps futur - conditionnel. En outre, l'analyse qualitative et quantitative a mis en avant la prééminence de certains verbes spécifiques concernés par les fautes d'accord (*faire*, *dire*, *être* et *venir*), des verbes qui font partie de l'ensemble des verbes récurrents du corpus total.

C'est l'analyse sociolinguistique qui permet de répondre à la troisième question de Calvet : *Quelles sont les différences (géographiques, sociales, etc.) entre leurs façons de parler ?* Quelques grandes tendances sociolinguistiques se sont dessinées à l'issue de notre analyse. Premièrement, nous notons des pratiques particulièrement

variables d'une classe d'âge à l'autre. Ainsi, l'analyse lexicométrique a montré une nette différence chez les moins de 25 ans : leur nombre de tokens par message est plus important que celui de leurs ainés (ils exploitent de façon maximale l'espace de communication qui leur est offert). Le nombre d'unités uniques, ou hapax, est largement plus important chez les plus de 35 ans que chez les plus jeunes, leur lexique semble donc plus diversifié que ne l'est celui des autres classes. En ce qui concerne la néologie, la créativité lexicale semble concerner davantage la classe des 15-25 ans que les autres classes. D'autres variables se sont révélées intimement liées à l'âge : ainsi, les locuteurs les plus jeunes semblent plus portés sur l'emploi d'éléments allogènes, de structures interrogatives moins formelles et sur la non rétention du *ne* de négation. Grâce aux tests statistiques, nous avons également conclu à l'influence de l'âge (classes d'âge les plus jeunes) sur la propension à réaliser des écarts orthographiques.

Le genre est une variable significative dans la plupart des phénomènes étudiés : les femmes auteures de messages semblent plus propices à s'écarter d'une production formelle que leurs homologues masculins. Tel est le cas du moins pour les phénomènes d'emploi de néologismes, d'éléments allogènes, de régionalismes et de négatives avec rétention du *ne*.

Pour un plus petit nombre de phénomènes, les tests statistiques montrent que les étudiants et diplômés du post-secondaire sont plus conservateurs en matière de production linguistique. En effet, nous avons trouvé un sous-emploi de néologismes et d'éléments allogènes chez ce groupe. On pourrait ainsi mettre en corrélation le niveau croissant d'éducation et la tendance à la formalité. Nous avons également conclu à l'influence du niveau d'étude (niveaux les plus bas) sur la propension à réaliser des écarts orthographiques.

De la même manière, nous notons une plus grande inhibition des locuteurs de langue première française (uniquement) par rapport aux bilingues et aux allophones pour les phénomènes de créativité lexicale et de non rétention du *ne* de négation notamment. Nous avançons ici l'hypothèse d'une désinhibition des allophones et des bilingues face à une langue française qui leur semble peut-être moins « monolithique ».

Les pratiques en matière de chat, de forum de discussion, de courriel et du nombre de sms envoyés par semaine ne montrent pas de tendances significatives dans le corpus de sms.

Dans l'introduction de ce travail de recherche, nous avions posé un certain nombre de questions auxquelles il nous semble bon à présent de répondre. Nous venons de répondre aux questions *Quelles sont les spécificités linguistiques et extralinguistiques de l'eSMS par rapport à la Cémo en général ?*, *Quel lexique emploie-t-on dans les sms ?*, *Dans quelle mesure l'eSMS, en tant que médium de communication écrit, partage-t-il des traits du français ordinaire ?* et *L'écrit sms témoigne-t-il de varia-*

tions régionales et de variations tributaires du profil sociolinguistique de l'auteur ? dans les paragraphes précédents, lorsque nous nous penchions sur les questions de base posées par Calvet et Dumont (1999). Regardons à présent les autres questions.

1. Quels sont les enjeux et les limites d'une collecte de sms et de profils sociodémographiques ?

Les enjeux et les limites d'une collecte de sms portent sur les deux points que sont la méthodologie de collecte et l'harmonisation des données. En effet, nous avons vu que la collecte de données de la Cémo – et des sms en particulier – accompagnées de leurs profils sociolinguistiques n'était pas chose aisée puisqu'il s'agit de données *privées*. Des questions d'éthique entrent donc en jeu. Dans le cas des sms, l'obtention (l'envoi) des messages pose davantage de problèmes puisque les données sont stockées sur un téléphone et non sur un ordinateur connecté à Internet [1]. Des précautions juridiques et médiatiques ont donc été prises afin d'obtenir des données réelles, en nombre suffisant (plus de 50.000 sms et de 2.500 profils).

La procédure d'harmonisation des données a été appliquée tant au niveau des messages qu'au niveau des questionnaires sociolinguistiques. Le traitement des sms a compris une phase d'anonymisation et de transcription en graphie standard ; le traitement des profils a consisté en une phase de mise en commun des données comparables et exploitables. Ces deux phases ont différé d'une région à l'autre de l'enquête, malgré le protocole commun qui visait une harmonisation *a priori*, en raison des différentes volontés individuelles, mais également des situations sociales et légales de chaque pays. Une harmonisation *a posteriori* s'est donc révélée nécessaire.

2. Existe-t-il des constantes de l'écrit sms que l'on retrouve dans tous les sous-corpus (quels que soient la région, le profil de l'auteur, etc.) ?

Cette étude sur la variation pose évidemment la question des constantes linguistiques observables. Au niveau du lexique, nous avons mis en avant plusieurs phénomènes clairs : l'omniprésence de l'anglo-américain dans l'ensemble des corpus montre qu'il n'est pas question d'une seule influence des langues en contact au Québec, mais d'un phénomène global. L'analyse des néologismes fait transparaitre l'importance du sigle *mdr* dans toutes les régions francophones. Une enquête sur l'oral montrerait, nous n'en doutons pas, son emploi répandu. Nous nous étonnons donc qu'il ne soit pas entré dans la dernière édition du Petit Robert (2013), d'autant que son équivalent anglais *lol* vient d'y faire son apparition. Le recours à des régionalismes constitue également, selon nos résultats, un phénomène omniprésent et ce, que cet emploi soit motivé par des raisons de manifestations identitaires ou qu'il soit le résultat d'une intégration spontanée des régionalismes au vocabulaire des locuteurs. Enfin, l'analyse diastratique a montré de façon distincte que dans toutes les régions, les femmes et les locuteurs les plus jeunes avaient davantage tendance à utiliser, dans les sms, des variantes et des structures moins formelles. Nous pouvons alors nous poser la question de savoir si l'écrit sms est une variante stable ou une variante en devenir,

1. Avec les smartphones, cette difficulté tend à disparaitre.

en construction. Selon Labov (1990) : « dans une situation de développement en cours, les femmes ont tendance à devancer les hommes et à davantage utiliser des variantes innovantes, alors que dans le cas d'une variation stable, les femmes ont tendance à user de formes plus prestigieuses que les hommes » [2]. Il existerait donc dans l'eSMS des constantes, mais dans une situation de variation en cours.

3. Dans quelle mesure les assertions connues au sujet de l'écrit sms (abréviation systématique, tendance à l'informel, « destruction » du langage ou de la syntaxe) sont-elles vérifiées par l'exploration d'une large quantité de données sms ?

Nous avons vu que la tendance à l'abréviation dans les sms, commandée par l'espace restreint du médium de communication, est présente dans les sms, mais ne prévaut pas. Nous avons par exemple montré que les phénomènes graphiques engendraient tantôt l'abréviation du message, tantôt un allongement de celui-ci (étirements graphiques). Des points de vue lexical et syntaxique, il a été démontré que le recours à des éléments allogènes, à la créativité lexicale, à l'emploi de régionalismes et la préférence de tournures interrogatives ne se faisait pas toujours à l'avantage de la brièveté et de la rapidité. Le choix de formes allogènes et de régionalismes plus couteux en nombre de caractères que leur équivalent en français de référence et la préférence des formes interrogatives en *est-ce que* sur celles par inversion sujet-verbe en sont des exemples parlants. Il existe donc une balance dans l'eSMS entre nécessité de faire court (rester en deçà des 160 caractères et maintenir une fluidité dans les tours de parole) et besoin de transmettre de l'expressivité, les signes stylistiques d'une attitude, d'une appartenance ou de relations sociales.

Il existe indéniablement, dans les sms, une tendance à préférer des formules moins formelles. Nous avons conclu ceci de l'analyse syntaxique notamment. On constate par exemple que seuls 11% des interrogatives sont formés par des constructions moins formelles (interrogatives disjonctives, par point d'interrogation, avec ellipse du verbe, etc.). De la même façon, seuls 23,14% des négatives présentent une rétention du *ne* de négation. Toutefois, dans certaines constructions, ce chiffre augmente jusque 60 voire 80%. C'est notamment le cas lorsque le verbe est à l'impératif ou présente une initiale vocalique. Il semblerait donc qu'il existe des lieux de résistance de la variation syntaxique dans les sms.

Quant à la supposée destruction du langage ou de la syntaxe dans les sms, nos résultats montrent qu'il n'en est rien. Il existe une transformation du code écrit, un jeu sur ce code. On ne peut pas nier non plus la prévalence des formes moins formelles sur les formes davantage formelles. Par contre, l'analyse orthographique prouve le faible taux de variantes ; de façon identique, les mélanges de langues ne se font pratiquement jamais sans respecter les règles grammaticales des deux langues.

2. « in situation of change in progress, women tend to be ahead of men and use larger proportions of innovative variants, while in stable variation, women tend to use more prestigious forms than men »

Comme énoncé précédemment, les différences de contexte d'utilisation de l'écrit sms, les variations lexicales, morphologiques et syntaxiques qui ont lieu dans les sms par rapport à d'autres types d'écrit proposent donc une manière complètement différente de considérer l'écrit. Cette pratique, parallèlement à la pratique de l'écrit traditionnel (en milieu professionnel, scolaire, etc.), nous fait penser qu'il existe donc plutôt une **double compétence** inhérente aux usagers du sms, une digraphie.

Chapitre 12

Limites et perspectives

Parmi les limites de notre étude, certaines auraient pu être anticipées et évitées, d'autres non. La question de la méthode de collecte a notamment engendré un biais non négligeable : on soulignera en particulier la variation dans la formulation des questions dans les différents questionnaires sociolinguistiques. Le choix des ressources lexicales utilisées lors des analyses est également critiquable. Le choix du Delaf, dictionnaire du français non mis à jour depuis une décennie, ne semble pas pertinent pour des analyses portant notamment sur la néologie. Les dictionnaires de régionalismes (particulièrement en Belgique) auraient pu être mieux choisis, après une analyse approfondie de la portée de ceux-ci, afin d'éviter d'inclure un nombre important d'unités syntaxiques qui tiennent plus de la spécificité (par fréquence d'utilisation) d'une région, que de régionalismes en tant que tels.

Parmi les limites peu évitables de notre démarche, nous citerons les difficultés rencontrées lors de la mise en place du projet : constitution d'équipes dans des régions instables pour des raisons géopolitiques ou économiques (Afrique du Nord), faible intérêt de la population pour les questions scientifiques, entrainant une frilosité d'investissement de la part des sponsors (Royaume-Uni), etc. Certains comités d'éthique et autres instances de protection de la vie privée ralentissent particulièrement certaines collectes et engendrent une quantité non négligeable de perte d'information. Enfin, le système de téléphonie mobile d'un pays ne permet pas toujours la gratuité d'envoi des copies de sms (Québec), ce qui freine considérablement la participation de la population à l'enquête. Seul ce dernier point parait pouvoir être contourné en recourant, notamment, à l'envoi des copies de messages par la connexion Internet des téléphones intelligents [1].

Les perspectives qu'offrent les résultats obtenus et les pistes amorcées dans ce travail sont nombreuses. Que pensent les locuteurs de leur façon de parler ? De celle des autres ? Ces questions pourraient trouver des éléments de réponse dans notre corpus car, comme nous l'avons présenté, notre enquête sociolinguistique comprend une case de commentaire libre où chaque participant a le loisir de s'exprimer, notamment à propos de ses pratiques et de l'écrit sms. Il s'agit donc d'une première piste importante pour la recherche future. Ensuite, nous pensons qu'avec un tel corpus, il serait tout à fait intéressant d'étudier la variation des messages d'un ou de quelques locuteurs en fonction de leurs destinataires ou de la

1. Ce qui engendrerait un nouveau biais de participation : seuls les détenteurs de ce type d'appareil pourraient participer.

situation de communication. Ainsi, on pourrait regarder si, chez un même locuteur, des formes abrégées apparaissent également non abrégées, ce qui prouverait une certaine double compétence à l'écrit. Quelques thématiques plus précises nous paraissent également très prometteuses pour la recherche future. L'étude des ouvertures et fermetures du canal de communication (telle que nous l'avons amorcée pour les éléments allogènes) : une typologie de ces formules conventionnelles, mais également une étude de leur variation en fonction des profils des auteurs.

Enfin, il nous semble capital de porter les résultats obtenus dans ce travail en dehors de ce corpus de sms ; d'une part, nous pensons qu'il serait nécessaire de réorganiser une collecte dans une zone déjà couverte afin d'étudier la variation des pratiques dans le temps. D'autre part, il serait pertinent de confronter les résultats obtenus pour les sms avec des corpus de taille similaire provenant de réseaux sociaux et de chat, notamment, afin de parvenir un peu plus finement à une définition de l'écrit sms.

Bibliographie

Abbey, B. (2000). Instructional and cognitive impacts of web-based education. Hershey, PA : Idea Group. (Cité en page 22.)

Adaci, S. (2008). La néologie journalistique. Analyse des néologismes de la presse écrite francophone (Le cas du Quotidien d'Oran). Mémoire de D.E.A., Université Mentouri. Constantine. (Cité en page 143.)

Adams, D. (2000). SMS speaks to deaf needs. *IT : News from the world of information technology. Retrieved January*, 29. (Cité en page 22.)

Agresti, A. (2002). *Categorical data analysis*. Wiley. (Cité en page 120.)

Angeli, C., Valanides, N. et Bonk, C. (2003). Communication in a web-based conferencing system : The quality of computer-mediated interactions. *British Journal of Educational Technology*, 1(34):31–43. (Cité en page 22.)

Anis, J. (1998). *Texte et ordinateur : l'écriture réinventée ?* De Boeck, coll. Méthodes en Sciences humaines, Bruxelles. (Cité en pages 56 et 57.)

Anis, J. (1999). *Internet, communication et langue française*. Hermès, Paris. (Cité en pages 29, 55 et 56.)

Anis, J. (2000). L'ecrit des conversations électroniques de l'internet. *Le français aujourd'hui*, 129. (Cité en pages 51 et 58.)

Anis, J. (2001). *Parlez–vous texto ? Guide des nouveaux langages de réseau*. Le cherche midi, Paris. (Cité en page 24.)

Anis, J. (2002). Communication électronique scripturale et formes langagières : chats et SMS. *Actes des quatrièmes rencontres réseaux humains/réseaux technologiques*. (Cité en pages 18, 42, 44, 49, 50 et 51.)

Anis, J., de Fornel, M. et Franckel, B. (2004). La communication électronique : Approches linguistiques et anthropologiques. *In International colloquium, EHESS*, pages 5–6, Paris. (Cité en page 18.)

Armstrong, N. (2002). Variable deletion of french ne : a cross-stylistic perspective. *Language Sciences*, 24:153–173. (Cité en page 249.)

Armstrong, N. et Smith, A. (2002). The influence of linguistic and social factors on the recent decline of french ne. *Journal of French Language Studies*, 12:23–41. (Cité en pages 249 et 259.)

Ashby, W. (2001). Un nouveau regard sur la chute du ne en français parlé tourangeau : s'agit–il d'un changement en cours ? *Journal of French Language Studies*, 11:1–22. (Cité en pages 249 et 251.)

Atifi, H. (2007). Choix linguistiques et alternance codique dans les forums diasporiques marocains. *In La langue du cyberspace : de la diversité aux normes*, pages 31–46. L'Harmattan. (Cité en page 22.)

Auer, P. (1998). Introduction : 'Bilingual conversation' revisited. *In* Auer, P., éditeur : *Codeswitching in Conversation : Language, Interaction and Identity*, pages 1–24. Routledge, Londres. (Cité en page 156.)

Auer, P. (2000). A conversation analytic approach to code–switching and transfer. *In* Wei, L., éditeur : *The Bilingualism Reader*, pages 166–187. Routledge, New York, seconde édition. (Cité en page 160.)

Auger, J. (2006). Variation in French. *In* Brown, K., éditeur : *Encyclopedia of Language and Linguistics*, pages 354–361. Elsevier, Oxford, seconde édition. (Cité en page 128.)

Aymes, M., Hersant, J. et Massicard, E. (2005). Prière d'insérer. *Labyrinthe*, 21:13–19. (Cité en page 54.)

Bamba, F. et Barnes, S. (2007). Sms advertising, permission and the consumer : a study. *Business Process Management Journal*, 13(6):815–829. (Cité en page 20.)

Bar, F. (1973). Les néologismes chez les burlesques du XVIIe siècle. *Cahiers de l'association internationale des études françaises*, 25:45–58. (Cité en page 144.)

Baron, N. (1998). Letters by phone or speech by other means : The linguistics of email. *Language and communication*, 18:133–170. (Cité en page 31.)

Baron, N. (2008). *Always On. Language in an Online Mobile World*. Oxford University Press, New York. (Cité en page 144.)

Baron, N. et Ling, R. (2003). IM and SMS : A linguistic comparison. *In Fourth International Conference of the Association of Internet Researchers*, Toronto. (Cité en pages 30 et 31.)

Bauche, H. (1920). *Le langage populaire : grammaire, syntaxe et dictionnaire du français réel*. Payot, Paris. (Cité en page 48.)

Bautista, M. (2004). Tagalog–english code switching as a mode of discourse. *Asia Pacific Education Review*, 5(2):226–233. (Cité en page 159.)

Bauvois, C. (2002). *Nid'Ève, ni d'Adam : étude sociolinguistique de douze variables du français*. L'Harmattan, Paris. (Cité en page 128.)

Beaufort, R., Roekhaut, S., Cougnon, L.-A. et Fairon, C. (2010). A hybrid rule/model–based finite–state framework for normalizing SMS messages. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 770–779, Uppsala. Association for Computational Linguistics. (Cité en pages 24, 92, 93 et 95.)

Beaufort, R., Roekhaut, S. et Fairon, C. (2008). Définition d'un système d'alignement SMS/français standard à l'aide d'un filtre de composition. *In Proceedings of JADT*, pages 155–166. (Cité en page 24.)

Beeching, K. (2007). Social identity, salience and language change : The case of post-rhematic 'quoi'. *In* Ayres-Bennet, W. et Jones, M., éditeurs : *The French Language and Questions of Identity*, pages 140–149. Legenda, London. (Cité en page 128.)

Beniamino, M. (1996). *Le français de la Réunion : inventaire des particularités lexicales*. EDICEF, Vanves. (Cité en pages 193, 194 et 196.)

Beniamino, M. et Baggioni, D. (1993). Le français, langue réunionnaise. *In* de Robillard, D. et Beniamino, M., éditeurs : *Le français dans l'espace francophone*, pages 151–172. Champion. (Cité en page 193.)

Benveniste, E. (1966). *Problèmes de linguistique générale I*. Gallimard, Paris. (Cité en page 70.)

Berruto, G. (2005). Italiano parlato e comunicazione mediata dal computer. *In* Hölker, K. et Maass, C., éditeurs : *Aspetti dell'italiano parlato*, pages 137–156. LIT Verlag, Münster. (Cité en pages 22 et 49.)

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press, Cambridge. (Cité en page 128.)

Biber, D. (2011). Register and discourse analysis. *In* Gee, J. et Handford, M., éditeurs : *The Routledge Handbook of Discourse Analysis*. Routledge, London–New York. (Cité en page 60.)

Biber, D. et Conrad, S. (2009). *Register, Genre, and Style*. Cambridge University Press, Cambridge. (Cité en pages 60, 151 et 272.)

Bieswanger, M. (2007). 2 abbrevi8 or not 2 abbrevi8 : A contrastive analysis of different space and time–saving strategies in English and German text messages. *In Texas Linguistic Forum*, volume 50, Austin. (Cité en page 19.)

Bieswanger, M. (2010). Gendered language use in computer-mediated communication : Typography in text messaging. *In* Bieswanger, M., Motschenbacher, H. et Mühleisen, S., éditeurs : *Language in its Socio-Cultural Context : New Explorations in Gendered, Global and Media Uses*, pages 157–172. Frankfurt am Main. (Cité en page 129.)

Black, J. et Ranjan, N. (2004). Automated event extraction from email. Rapport technique, Final Report of CS224N/Ling237 Course in Stanford. (Cité en page 24.)

Blampain, D. (1997). *Le français en Belgique : une langue, une communauté*. De Boeck. (Cité en page 185.)

Blanche-Benveniste, C. (1997). *Approche de la langue parlée en français.* Ophrys, Coll. L'essentiel Français, Paris. (Cité en pages 57 et 230.)

Blanche-Benveniste, C. et Bilger, M. (1999). "Francais parlé–oral spontané". Quelques réflexions. *Revue française de linguistique appliquée*, 4(2):21–30. (Cité en page 56.)

Blom, J. et Gumperz, J. (1972). Social meaning in linguistic structures : Code–switching in norway. *Directions in sociolinguistics*, pages 407–434. (Cité en pages 156 et 160.)

Bob (2011). Bob : dictionnaire argot populaire et familier. Sur le site abc langue française : `http://www.languefrancaise.net/bob/`. (Cité en page 155.)

Bordal, G. et Ledegen, G. (2007). Le français de la réunion : lexique, morpho-syntaxe, alternance codique et prononciation. *Bulletin PFC : enjeux descriptifs, théoriques et didactiques 2007*, 7:121–134. (Cité en pages 193 et 195.)

Bouillaud, C., Chanquoy, L. et Gombert, J.-E. (2007). Cyberlangage et orthographe : quels effets sur le niveau orthographique des élèves de CM2, 5e et 3e ? *Bulletin de psychologie*, tome 60(6):553–565. (Cité en page 21.)

Boulanger, J. (1984). Quelques observations sur l'innovation lexicale spontanée et sur l'innovation lexicale planifiée. *Banque des mots*, (27):3–29. (Cité en page 143.)

Boulanger, J. (1985). À propos du concept de régionalisme. *Lexique (Presses universitaires de Lille)*, 3:125–146. (Cité en page 183.)

Bourdieu, P. (1993). *La misère du monde.* Seuil, Paris. (Cité en page 101.)

Bourguin, G., D'Halluin, C. et Hoogstoel, F. (2000). L'apprentissage coopératif et la conception de collecticiels. *In TICE 2000 : Technologies de l'Information et de la Communication dans les Enseignements d'ingénieurs et dans l'Industrie. Colloque international*, pages 361–367, Troyes. (Cité en page 22.)

Boyd, D. M. et Ellison, N. B. (2007). Social network sites : Definition, history, and scholarship. *Journal of Computer–Mediated Communication*, 13(1):210–230. (Cité en page 36.)

Bremer, J. et Rauch, P. K. (1998). Children and computers : Risks and benefits. *Journal of the American Academy of Child & Adolescent Psychiatry*, 37(5):559–560. (Cité en page 23.)

Brenner, V. (1997). Psychology of computer use : XLVII. parameters of Internet use, abuse and addiction : the first 90 days of the Internet Usage Survey. *Psychological reports*, 80:879–882. (Cité en page 23.)

Brown, P. et Levinson, S. (1987). *Politeness : Some universals in language usage.* Cambridge University Press, Cambridge. (Cité en page 53.)

Bruckman, A. (1992). Identity workshop : Emergent social and psychological phenomena in text–based virtual reality. *Unpublished manuscript*. (Cité en page 37.)

Cabré, M.-T., Bayà, M.-R., Bernal, E., Freixa, J., Solé, E. et Vallès, T. (2008). Evaluación de la vitalidad de una lengua a través de la neología : a propósito de la neología espontànea y de la neología planificada. *Lèxic i neologia*, pages 159–201. (Cité en page 143.)

Cabré, M.-T. et de Yzaguirre, L. (1995). Stratégie pour la détection semi–automatique des néologismes de presse. *TTR : Traduction, Terminologie, Rédaction : Études Sur le Texte et Ses Transformations*, 2(8):89–100. (Cité en page 143.)

Calvet, L. J. (2004). Une langue qui meurt, c'est une vision du monde qui disparait. Interview pour l'Express, propos recueillis par Marianne Payot, 22 novembre 2004. (Cité en page 101.)

Calvet, L. J. et Dumont, P. (1999). *L'enquête sociolinguistique*. L'Harmattan. (Cité en pages 63, 102, 104, 271, 272 et 275.)

Carayol, M. (1985). *Particularités lexicales du français réunionnais : propositions pédagogiques*. Nathan. (Cité en page 193.)

Carey, J. W. (1993). Everything that rises must diverge : Notes on communications, technology and the symbolic construction of the social. *Beyond agendas*, pages 171–184. (Cité en page 20.)

Carlstrom, E. L. (1992). Better living through language : The communicative implications of a text–only virtual environment. *Student Paper, Grinnell College*. (Cité en page 38.)

Caron, A. H. et Caronia, L. (2005). *Culture mobile : les nouvelles pratiques de communication*. Les Presses de l'Université de Montréal, Montréal. (Cité en pages 18, 28, 38, 42 et 50.)

Carricano, M., Poujol, F. et Bertrandias, L. (2010). *Analyse de données avec SPSS®*. Pearson Education. (Cité en pages 132 et 388.)

Catach, N. (1980). *L'orthographe française*. Nathan Université. (Cité en pages 12, 217, 218, 219 et 228.)

Cavazza, F. (2012). Panorama des médias sociaux 2012. Sur le blog MediasSociaux.fr. (Cité en pages 34 et 387.)

Charaudeau, P. et Maingueneau, D. (2002). *Dictionnaire d'analyse du discours*. Seuil, Paris. (Cité en pages 56, 64 et 65.)

Chaudenson, R. (1974). *Le lexique du parler créole de la Réunion*, volume 1. Champion. (Cité en page 193.)

Chaudenson, R., Mougeon, R. et Beniak, E. (1993). *Vers une approche panlectale de la variation du français*. Didier–Érudition, Paris. (Cité en page 193.)

Chen, T. et Kan, M.-Y. (2011). Creating a live, public short message service corpus : The nus sms corpus. Computing Research Repository abs/1112.2468. (Cité en pages 24 et 25.)

Cherny, L. (1995a). Objectifying the body in the discourse of an object–oriented mud. *Works and Days*, 13(1). (Cité en page 38.)

Cherny, L. (1995b). The modal complexity of speech events in a social MUD. *Electronic Journal of Communication*, 5(4). (Cité en page 38.)

Chung, F. et Lu, L. (2004). The average distances in random graphs with given expected degrees. *Internet Mathematics*, 1(1):91–113. (Cité en page 33.)

Clark, C. (1940). *The conditions of economic progress*. Macmillan London. (Cité en page 113.)

Cochran, M., Gunnarsson, L., Grabe, S. et Lewis, J. (1990). The social networks of coupled mothers in four cultures. pages 86–104. Cambridge University Press, London. (Cité en page 22.)

Collot, M. et Belmore, N. (1996). Electronic language : A new variety of English. *In* Herring, S., éditeur : *Computer–Mediated Communication : Linguistic, Social and Cross–Cultural perspective*, pages 13–28. John Benjamins, Philidelphie. (Cité en pages 56 et 57.)

Compagnone, M. R. (2011). *Verba volant, scripta etiam. Comunicazione « schermo a schermo » : uno scritto che cerca di avvicinarsi all'orale*. Thèse de doctorat, Paris 10. (Cité en page 22.)

Cortelazzo, M. (2000). 6 proprio 3mendo : dalla lettera ai messaggini in codice. Oralita, concisione, assenza di sintassi : Le caratteristiche di una scrittura 'allegra'. *Corriere della Sera, 19.08.2000*. (Cité en page 24.)

Cortelazzo, M. (2001). Telefonini cellulari e computer rilanciano la scrittura, ideografica. *Telèma*, 6(2–3):102–104. (Cité en page 24.)

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. Mémoire de D.E.A., Université catholique de Louvain. (Cité en pages 156 et 160.)

Cougnon, L.-A. et Beaufort, R. (2011). Néologie et sms. *Neologica*, (5):183–201. (Cité en page 144.)

Cougnon, L.-A. et Fairon, C. (2010). La néologie dans 'l"écrit spontané". Etude d'un corpus de SMS en Belgique francophone. *In Actes du Congrès International de la néologie dans les langues romanes*, pages 1139–1154, Barcelone. Sèrie Activitats. (Cité en page 144.)

Cougnon, L.-A. et François, T. (2010). Quelques contributions des statistiques à l'analyse sociolinguistique d'un corpus de SMS. *In* Bolasco, S., Chiari, I. et Giuliano, L., éditeurs : *Statistical Analysis of Textual Data. Proceedings of 10th International Conference JADT*, pages 619–630, Sapienza. Université de Rome. (Cité en pages 72, 118, 120, 132, 162, 342 et 392.)

Cougnon, L.-A. et François, T. (2011). Étudier l'écrit SMS. Un objectif du projet sms4science. *In* Stähli, A., Dürscheid, C. et Béguelin, M.-J., éditeurs : *La communication par SMS en Suisse. Usages et variétés linguistiques.* Linguistik Online. (Cité en pages 114 et 132.)

Cougnon, L.-A. et Ledegen, G. (2010). "c'est écrire comme je parle". Une étude comparatiste de variétés de français dans l'"écrit sms". *Les voix des Français. Modern French Identities*, 2(94):39–57. (Cité en pages 184, 185 et 193.)

Cougnon, L.-A., Roekhaut, S. et Beaufort, R. (2012). Typologies de variation graphique dans l'écrit sms. *In Actes du Colloque international "20e anniversaire des Rectification de l'orthographe de 1990 : enseignement, recherche et réforme, quelles convergences ?"*. (Cité en page 216.)

Coulombe, M. (2003). *Les dangers du clavardage.* Québecor, Canada. (Cité en page 23.)

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue française*, 87(1):11–22. (Cité en page 95.)

Courtois, B. (2004). Dictionnaires électroniques DELAF anglais et français. *In* et al. Leclère, C., éditeur : *Lexique, syntaxe et lexique–grammaire ; syntax, lexis & lexicongrammar*, pages 113–123. (Cité en page 130.)

Coveney, A. (1989). *Variability in interrogation and negation in spoken French.* Thèse de doctorat, Newcastle University. (Cité en pages 260 et 261.)

Crinon, J., Gautellier, C. et Azémard, G. (2001). *Apprendre avec le multimédia et Internet.* Retz, Paris. (Cité en page 22.)

Crystal, D. (2001). *Language and the Internet.* Cambridge University Press. (Cité en pages 18 et 144.)

Crystal, D. (2008). *Txtng : The gr8 db8.* Oxford University Press, Oxford. (Cité en page 183.)

Curtis, D. et Lawson, M. (2001). Exploring collaborative online learning. *Journal of Asynchronous Learning Networks*, 1(5):21–34. (Cité en page 22.)

Cusin-Berche, F. (1999). Courriel et genres discursifs. *Internet, communication et langue française*, pages 31–54. (Cité en pages 31, 42 et 57.)

Daille, B. (2001). Extraction de collocations à partir de textes. *In Huitième Conférence Nationale sur le Traitement Automatique des Langues Naturelles, Tours, France*. TALN. (Cité en page 141.)

Damourette, J. et Pichon, E. (1927). *Des mots à la pensée. Essai de grammaire de la langue française I*. Paris : D'Artrey. (Cité en page 249.)

Davau, M., Cohen, M. et Lallemand, M. (1972). *Dictionnaire du français vivant*. Bordas. (Cité en page 183.)

David, J. et Goncalves, H. (2007). L'écriture électronique, une menace pour la maitrise de la langue ? *Le Français aujourd'hui*, 156:39–47. (Cité en pages 4 et 21.)

Davis, J. R. (1991). Let Your Fingers do the Spelling : Implicit disambiguation of words spelled with the telephone keypad. *Journal of The American Voice I/O Society*, 9:57–66. (Cité en page 24.)

Degand, L. et Bestgen, Y. (2004). Connecteurs et analyses de corpus : de l'analyse manuelle à l'analyse automatisée. *In* Porhiel, S. et Klingler, D., éditeurs : *L'Unité texte. Actes du colloque Regards croisés sur l'Unité texte/Conjoint Perspectives on Text*, pages 49–73, Chypre. (Cité en page 131.)

Dejond, A. (2002). *La cyberlangue française*. La Renaissance du livre, Tournai. (Cité en page 18.)

Dejond, A. (2006). *Cyberlangage*. Racine, Bruxelles. (Cité en pages 50, 52, 56 et 60.)

Desquinabo, N. (2008). Dynamiques et impacts des propositions politiques dans les webforums partisans. *Réseaux*, 26(150):107–132. (Cité en page 33.)

Deumert, A. et Oscar Masinyana, S. (2008). Mobile language choices the use of english and isixhosa in text messages (sms) evidence from a bilingual south african sample. *English World–Wide*, 29(2):117–147. (Cité en page 154.)

Deumert, A. et Vold Lexander, K. (2012). Multilingual texting in africa — between convention and creativity. Colloque de la VALS–ASLA, Lausanne, 1–3 février 2012. (Cité en pages 154 et 164.)

Devisme, C. et Dussarps, C. (2010). Construction identitaire d'un apprenant à distance : du territoire aux péri–espaces de formation. (Cité en page 20.)

Dister, A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL*. Thèse de doctorat, Université catholique de Louvain. (Cité en page 72.)

Dister, A., Francard, M., Hambye, P. et Simon, A. (2009). Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de banque de données textuelles orales VALIBEL (1989–2009). *Cahiers de Linguistique*, 33(2):113–129. (Cité en pages 69, 94 et 123.)

Dister, A. et Moreau, M. (2006). "Dis–moi comment tu féminises, je te dirai pour qui tu votes". Les dénominations des candidates dans les élections européennes de 1989 et de 2004 en Belgique et en France. *Langage et société*, (116):5–46. (Cité en page 65.)

Do, T.-N.-D. (2011). *Extraction de corpus parallèle pour la traduction automatique depuis et vers une langue peu dotée*. Thèse de doctorat, Université de Grenoble. (Cité en page 92.)

Doehler, S. (2011). Hallo ! voulez vous luncher avec moi hüt ? le“ code switching” dans la communication par sms. *Linguistik online*, 48(4). (Cité en pages 154, 157, 159, 175 et 182.)

Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., Marcellesi, J. B. et Mével, J. P. (1994). *Dictionnaire de linguistique et des sciences du langage*. Larousse. (Cité en pages 65 et 83.)

Dubreil, E. (2008). Collocations : Définitions et problématiques. *Texto !*, XIII(1): 32–38. (Cité en page 141.)

Dubreil, E. et Daille, B. (2005). Analyse sémantico-discursive des collocations lexicales en corpus spécialisé : la base « connaissance–s ». (Cité en page 141.)

Duerscheid, C. et Stark, E. (2010). SMS4science : An international corpus–based texting project and the specific challenges for multilingual Switzerland. *In* Thurlow, C. et Mroczek, K., éditeurs : *Digital Discourse. Language in the New Media (= Oxford Studies in Sociolinguistics). Oxford : Oxford University Press (im Druck)*. Oxford, oxford uni édition. (Cité en page 79.)

Dufter, A. et Stark, E. (2007). La linguistique variationnelle et les changements linguistiques 'mal compris' : Le cas du ne de négation. *In* Combettes, B. et Marchello-Nizia, C., éditeurs : *Etudes sur le changement linguistique en français. Nancy, Presses Universitaires de Nancy*, pages 115–128. (Cité en pages 249, 253 et 265.)

Durham, M. (2003). Language choice on a swiss mailing list. *Journal of Computer–Mediated Communication*, 9(1). (Cité en page 154.)

Eco, U. (2002). Diminutive, but perfectly formed. *Guardian Newspaper*. (Cité en page 48.)

Eisenlohr, P. (2004). Language revitalization and new technologies : Cultures of electronic mediation and the refiguring of communities. *Annual Review of Anthropology*, 33:21–45. (Cité en page 183.)

Essono, J. (1998). *Précis de linguistique générale*. L'Harmattan. (Cité en page 69.)

Fabrot, B. (2001). *Les forums sur internet*. Marabout Informatique, Paris. (Cité en page 33.)

Fairon, C., Klein, J. R. et Paumier, S. (2006a). *Le Corpus SMS pour la science. Base de données de 30.000 SMS et logiciel de consultation*. CD-Rom, Presses universitaires de Louvain. (Cité en pages 4, 21 et 77.)

Fairon, C., Klein, J. R. et Paumier, S. (2006b). *Le langage SMS : étude d'un corpus informatisé à partir de l'enquête Faites don de vos SMS à la science*, volume 3. Les Cahiers du Cental, Presses universitaires de Louvain. (Cité en pages 4, 21, 77, 84, 85, 87, 89, 94, 144, 145, 183, 185 et 220.)

Fairon, C., Klein, J. R. et Paumier, S. (2006c). Le langage sms : révélateur d'1compétence. *"Le français m'a tuer". Actes du colloque L'orthographe française à l'épreuve du supérieur*, 1(Cahiers du Cental):33–42. (Cité en page 21.)

Fayada, M. (2007). Le langage SMS des jeunes. Approche lexicale et morpho–syntaxique. Mémoire de D.E.A., Université Paul–Valéry Montpellier 3. (Cité en page 61.)

Fiala, P. (1987). Pour une approche discursive de la phraséologie. Remarques en vrac sur la locutionnalité et quelques points de vue qui s'y rapportent, sans doute. *Langage et Société*, 42:27–44. (Cité en page 141.)

Francard, M., Geron, G., Wilmet, R. et A., W. (2010). *Dictionnaire des belgicismes*. De Boeck - Duculot. (Cité en page 186.)

Francard, M. et Latin, D. (1995). *Le régionalisme lexical*. De Boeck - Duculot. (Cité en page 183.)

Frehner, C. (2008). *Email – SMS – MMS : The linguistic creativity of asynchronous discourse in the new media age*. Peter Lang. (Cité en pages 21 et 41.)

Fung, P. et Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. *In Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 414–420. Association for Computational Linguistics. (Cité en page 83.)

Gadet, F. (1989). *Le français ordinaire*. Armand Colin, Paris. (Cité en pages 57, 232, 234, 236, 240, 247, 249, 253, 255 et 256.)

Gadet, F. (2000). Des corpus pour ne... pas. *In* Bilger, M., éditeur : *Corpus : méthodologie et applications linguistiques, Champion*, pages 156–167. Paris. (Cité en page 249.)

Gadet, F. (2004). *La variation sociale en français*. Ophrys. (Cité en pages 57, 58, 124, 128 et 129.)

Gadet, F. (2005). Construire une problématique de l'oral. Séminaire doctoral de Sciences du langage. (Cité en page 230.)

Gadet, F. (2006). Le locuteur comme champ de bataille. *In Conférence IUF. De l'écologie des langues et de la dynamique du langage*, volume 12. (Cité en page 109.)

Gadet, F. (2008). Ubi scripta et volant et manent. *Romanische Syntax im Wandel*, pages 513–529. (Cité en page 231.)

Gadet, F. c. (1991). Ces langues que l'on dit simples. *LINX*, (25). (Cité en page 49.)

Gains, J. (1999). Electronic Mail–A New Style of Communication or Just a New Medium ? : An Investigation into the Text Features of E-mail. *English for Specific Purposes*, 18(1):81–101. (Cité en page 21.)

Garneau, J. (1999). La Cyberdépendance : Mythe ou réalité. *La lettre du psy*, 3(10). (Cité en page 23.)

George, E. et Totsching, M. (2001). 20 ans de CMO. Dialogue sur l'histoire d'un concept et d'un champ de recherché. *La Communication Médiatisée par Ordinateur : un carrefour de problématiques*. (Cité en pages 17 et 20.)

Giancarli, P. (1999). Fonctions de l'alternance des langues chez des enfants bilingues francophones-anglophones. *Langage et société*, 88(1):59–88. (Cité en page 161.)

Girard, F. A. et Lyche, C. (2004). Norme, variation et enseignement du FLE. Chapitre de cours. (Cité en page 129.)

Goodman, P. S. et Sproull, L. S. (1990). *Technology and organizations*. Jossey-Bass, San Francisco. (Cité en page 20.)

Goumi, A., Volckaert-Legrier, O., Bert-Erboul, A. et Bernicot, J. (2011). SMS length and function : A comparative study of 13- to 18-year-old girls and boys. *European Review of Applied Psychology*, 61(4):175–230. (Cité en pages 22, 27 et 129.)

Gross, M. (1975). *Méthodes en syntaxe*. Hermann, Paris. (Cité en page 131.)

Guidère, M. (2008). *Introduction à la traductologie. Penser la traduction : hier, aujourd'hui, demain*. De Boeck. (Cité en page 92.)

Guimier, C. (2009). Les adverbes de domaine en–wise : des marqueurs du Netspeak ? *Revue LISA/LISA e-journal. Littératures, Histoire des Idées, Images, Sociétés du Monde Anglophone–Literature, History of Ideas, Images and Societies of the English-speaking World*, 7(3):394–410. (Cité en page 22.)

Guimier De Neef, E. et Fessard, S. (2007). Évaluation d'un système de transcription de SMS. *In Acte du 26e Colloque international Lexique Grammaire*, Bonifacio. (Cité en page 24.)

Guimier de Neef, E. et Veronis, J. (2004). Tutoriel : 1 pw1 sr la kestion :-). (Cité en page 18.)

Guitteny, P. (2007). Langue des signes, communauté sourde et CMT. *In* Gerbault, J., éditeur : *La langue du cyberespace : de la diversité aux normes*, pages 67–74. L'Harmattan, Paris. (Cité en page 22.)

Gumperz, J. (1989). *Sociolinguistique interactionnelle, une approche interprétative*. L'Harmattan, Paris. (Cité en page 161.)

Hård Segerstad, Y. (2005). Language in SMS—a socio-linguistic view. *In* Harper, R., Palen, L. et Taylor, A., éditeurs : *The Inside Text : Social, cultural and design perspectives on sms*, pages 33–51. Springer, Dordrecht, kluwer aca édition. (Cité en pages 4 et 56.)

Habert, B. (1999). Un corpus clé pour le français actuel. Présentation du projet en ligne : `http://estime.spim.jussieu.fr/CLEF/`. (Cité en page 72.)

Habert, B., Nazarenko, A. et Salem, A. (1997). *Les Linguistiques de corpus*. Armand Colin. (Cité en pages 92 et 93.)

Halliday, M. et Hasan, R. (1976). *Cohesion in English*. Longman, London. (Cité en page 65.)

Hambye, P. (2005). *La prononciation du français contemporain en Belgique : variation, normes et identités*. Thèse de doctorat, Université catholique de Louvain. (Cité en page 9.)

Hanse, J., Doppagne, A. et Bourgeois-Gielen, H. (1971). *Chasse aux belgicismes*. Fondation Charles Plisnier, Office du bon langage ; compte rendu dans l'information, 28 mars 1972, Bruxelles. (Cité en page 185.)

Hanse, J., Doppagne, A. et Bourgeois-Gielen, H. (1974). *Chasse aux belgicismes*. Fondation Charles Plisnier, Office du bon langage, Bruxelles. (Cité en page 185.)

Harris, Z. S. (1965). Transformational theory. *Language*, 41:363—401. (Cité en page 131.)

Hashimoto, Y. (2002). The spread of cellular phones and their influence on young people in Japan. *In* Kim, S. D., éditeur : *The Social and Cultural Impact/Meaning of Mobile Communication*, pages 101–112. Hallym University. (Cité en page 22.)

Heaton, L. et Lafrance, J. P. (1994). Les communautés virtuelles ludiques. Réflexions sur les jeux multi-utilisateurs. *Réseaux*, 12(67):95–110. (Cité en page 37.)

Herring, S. C. (1998). Le style du courrier électronique : variabilité et changement. *Revue d'aménagement linguistique*, 84-85:9–16. (Cité en pages 21 et 31.)

Herring, S. C. (2001). Computer-mediated discourse. *In* Schiffrin, D., Tanne, D. et Hamilton, H., éditeurs : *The handbook of discourse analysis*, pages 612–634. (Cité en page 27.)

Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet*, 4. (Cité en pages 11, 18, 27, 28, 42, 45 et 62.)

Herslund, M. (2003). *Aspects linguistiques de la traduction*. Presses universitaires de Bordeaux. (Cité en page 77.)

Hinrichs, L. (2006). *Codeswitching on the Web : English and Jamaican Patois in E-Mail Communication*. Benjamins, Amsterdam/Philadelphia. (Cité en page 154.)

Höflich, J. et Gebhardt, J. (2005). Changing Cultures of Written Communication : Letter—E-mail—SMS. *In* Harper, R., Palen, L. et Taylor, A., éditeurs : *The Inside Text*, pages 9–31. Springer, Dordrecht. (Cité en pages 4, 19 et 56.)

Holmer, T. (2008). Discourse structure analysis of chat communication. *Language@Internet*, 5. (Cité en page 22.)

How, Y. et Kan, M. Y. (2005). Optimizing predictive text entry for short message service on mobile phones. *In Proceedings of HCII*, Las Vegas. (Cité en page 24.)

Huang, D. (2004). *Code switching and language use in emails*. Thèse de doctorat, The University of Melbourne. (Cité en page 154.)

Hughes, A. et Grawoig, D. (1971). *Statistics : A foundation for analysis*. Addison-Wesley Publishing Company, Don Mills, Ontario. (Cité en page 134.)

Humbley, J. (1990). Le purisme dans les dictionnaires de l'informatique grand public. *Cahiers de lexicologie*, 1–2(56):241–253. (Cité en page 144.)

Jaffré, J. P. (2010). De la variation en orthographe. *Études de linguistique appliquée*, 3(159). (Cité en page 21.)

Jaffré, J. P. (2003). La linguistique et la lecture-écriture : de la conscience phonologique à la variable 'orthographe'. *Revue des sciences de l'éducation*, 29(1):37–49. (Cité en page 43.)

Jaffré, J. P. (2006). Orthographes et litéracies : Les particularités du Français. *In " Le français m'a tuer" : actes du Colloque L'orthographe française à l'épreuve du supérieur*, volume 1, pages 55–68, Louvain-la-Neuve. Les Cahiers du Cental, Presses universitaire de Louvain. (Cité en page 47.)

Johnson, W. (1944). Studies in language behaviour : I. A program approach. *Psychological Monographs*, 56:1–15. (Cité en page 136.)

Kasesniemi, E. L. et Rautiainen, P. (2010). Life in 160 characters : The text message culture of Finnish teenagers. *In* Katz, J. E. et Aakhus, M., éditeurs : *Perpetual contact : Mobile communication, private talk, and public performance*, pages 170–192. Cambridge University Press. (Cité en pages 22 et 24.)

Kennedy, C. et Levy, M. (2008). L'italiano al telefonino : Using sms to support beginners' language learning1. *ReCALL*, 20(3):315–330. (Cité en page 23.)

Kennedy, G. (1998). *An introduction to corpus linguistics*. Longman, Londres/New York. (Cité en pages 64, 66, 68 et 73.)

Kerbrat-Orecchioni, C. (1998). L'interaction épistolaire. *In* Siess, J., éditeur : *La lettre, entre réel et fiction*, pages 15–36. L'SEDES. (Cité en page 42.)

Klein, J. R. (2006). Quelques réflexions sur la dynamique lexicale du français au début du XXIe siècle. *Revue belge de philologie et d'histoire*, 84(3):673–685. (Cité en page 143.)

Klein, J. R. (2012). Entretiens privés. mai 2012. (Cité en page 135.)

Klein, J. R. et Rossari, C. (2003). Figement et variations en français de Belgique, de France, du Québec et de Suisse. *Linguisticae Investigationes*, 26(2):203–214. (Cité en page 183.)

Kling, R. (1996). Social relationships in electronic forums : Hangouts, salons, workplaces and communities. *In Computerization and controversy : Value conflicts and social choices*, pages 426–454. Academic Press, San Diego. (Cité en page 33.)

Klinkenberg, J.-M. (1996). *Précis de sémiotique générale*. De Boek, Louvain-la-Neuve. (Cité en pages 42, 46 et 47.)

Klinkenberg, J.-M. (1999). *Des langues romanes*. Duculot, Paris, Bruxelles, 2e édition. (Cité en page 262.)

Knecht, P. (1996). La Suisse romande : aspects d'un paysage francophone conservateur. *In* Robillard, D., éditeur : *Le français dans l'espace francophone*, volume 2, pages 760–770. Champion, Paris. (Cité en page 202.)

Ko, K. K. (1996). Structural Characteristics of Computer-Mediated Language : A Comparative Analysis of InterChange Discourse. *Electronic Journal of Communication/La revue électronique de communication*, 6(3). (Cité en pages 3, 56 et 57.)

Kobus, C., Yvon, F. et Damnati, G. (2008). Normalizing SMS : are two metaphors better than one ? *In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 441–448, Manchester. Association for Computational Linguistics. (Cité en page 24.)

Koch, P. et Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36:15–43. (Cité en pages 8, 131, 230 et 231.)

Koch, P. et Oesterreicher, W. (2001). Gesprochene Sprache und geschriebene Sprache/Langage parlé et langage écrit. *In* Holtus, G., Metzeltin, M. et Schmitt, C., éditeurs : *Lexicon der Romanistischen Linguistik*, volume I.2, pages 584–627. Niemeyer, Tübingen. (Cité en pages 59 et 265.)

Labbé, H. et Marcoccia, M. (2005). Communication numérique et continuité des genres : l'exemple du courrier électronique. *Revue électronique Texto*, 10(3). (Cité en page 42.)

Labov, W. (1966). *The social stratification of English in New York city*. Washington DC. : Center for Applied Linguistics. (Cité en pages 128 et 185.)

Labov, W. (1976). *Sociolinguistique*. Les Editions de Minuit, Paris. (Cité en page 128.)

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language variation and change*, 2(2):205–254. (Cité en pages 128 et 276.)

Laks, B. (1977). Contribution empirique à l'analyse socio-différentielle de la chut du /r/ dans les groupes consonantiques finals. *Langue française*, 34:109–125. (Cité en page 128.)

Lamiroy, B., Klein, J., Labelle, J., Leclère, C., Meunier, A. et Rossari, C. (2010). *Les expressions verbales figées de la francophonie : Belgique, France, Québec et Suisse*. Ophrys. (Cité en page 209.)

Laursen, D. (2005). Please reply ! The replying norm in adolescent SMS communication. *In* Harper, R., Palen, L. A. et Taylor, A. S., éditeurs : *The Inside Text*, pages 53–73. Springer, Dordrecht, kluwer aca édition. (Cité en pages 4 et 22.)

Laviosa, S. (2002). *Corpus-based translation studies : theory, findings, applications*. Rodopi, Amsterdam. (Cité en page 67.)

Lebouc, G. (2006). *Dictionnaire de belgicismes*. Racine. (Cité en pages 185, 186 et 188.)

Ledegen, G. (2004). Transformations de la société réunionnaise, évolutions sociolinguistiques et médias légitimant les pratiques linguistiques "ordinaires". *In* Klaeger, S. et Müller, M., éditeurs : *Medien und kollektive Identitätsbildung. Ergebnisse des 3. Franko-Romanistenkongresses*, pages 112–128. Praesens, Aachen. (Cité en page 193.)

Ledegen, G. (2011). Prédicats—flottants “entre le créole acrolectal et le français à la réunion. *In* Chamoreau, C. et Goury, L., éditeurs : *Systèmes prédicatifs des langues en contact*. CNRS Editions, Coll. « Sciences du langage ». (Cité en pages 193 et 196.)

Leech, G. (1991). The state of the art in corpus linguistics. *In* Aijmer, K. et Altenberg, B., éditeurs : *English Corpus Linguistics*, pages 8–29. Longman. (Cité en page 64.)

Leung, L. (2007). Unwillingness-to-communicate and college students' motives in SMS mobile messaging. *Telematics and Informatics*, 24(2):115–129. (Cité en page 20.)

L'Homme, M. (2000). Understanding Specialized Lexical Combinations. *Terminology*, 6(1):89–110. (Cité en page 141.)

L'Homme, M. (2004). *La terminologie : principes et techniques*. Presses de l'Université de Montréal, coll. Paramètres. (Cité en page 83.)

Licoppe, C. (2002). Sociabilité et technologies de communication : Deux modalités d'entretien des liens interpersonnels dans le contexte du déploiement des dispositifs de communication mobiles : Mobiles. *Réseaux*, (112-113):172–210. (Cité en page 48.)

Ling, R., Julsrud, T. et Yttri, B. (2005). Nascent communication genres within SMS and MMS. *In* Harper, R., Palen, L. A. et Taylor, A. S., éditeurs : *The Inside Text*, pages 75–100. Springer, Dord, kluwer aca édition. (Cité en page 22.)

Liu, W. et Wang, T. (2010). Index-based Online Text Classification for SMS Spam Filtering. *Journal of computers*, 5(6):844–851. (Cité en page 24.)

Loiseau, S., Poudat, C. et Ablali, D. (2006). Exploration contrastive de trois corpus de sciences humaines. *In JADT 2006*. (Cité en page 139.)

Lòpez Alonso, C. (2003). Mode d'interaction dans le courrier électronique. *Lidil. Revue de linguistique et de didactique des langues*, 28:109–120. (Cité en pages 3 et 30.)

Luzzati, D. (1991). Oralité et interactivité dans l'écrit Minitel. *Langue française*, 89:99–109. (Cité en pages 55, 56 et 57.)

MacKenzie, I. S. et Soukoreff, R. W. (2002). Text entry for mobile computing : Models and methods, theory and practice. *Human-Computer Interaction*, 17:147–198. (Cité en page 46.)

Mackey, W. (1968). The description of bilingualism. *Reading in the Sociology of Language*, pages 554–584. (Cité en page 160.)

Mangenot, F. et Zourou, K. (2007). Pratiques tutorales correctives via Internet : le cas du français en première ligne. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, 10(1):65–99. (Cité en page 22.)

Marcoccia, M. (2000). Les smileys : une représentation iconique des émotions dans la communication médiatisée par ordinateur. *In* Plantic, C., Douvy, M. et Traverso, V., éditeurs : *Les émotions dans les interactions communicatives*, pages 249–263. PUL, Lyon. (Cité en page 52.)

Marcoccia, M. (2004a). La communication écrite médiatisée par ordinateur : faire du face à face avec de l'écrit. *Journée d'étude ATALA : Le traitement automatique des nouvelles formes de communication écrite*. (Cité en pages 31, 52, 55 et 56.)

Marcoccia, M. (2004b). L'analyse conversationnelle des forums de discussion : questionnements méthodologiques. *Les Carnets du Cediscor*, 8:23–37. (Cité en page 33.)

Marcoccia, M. (2005a). Communication électronique et rapport de places : analyse comparative de la formulation d'une requête administrative par courrier électronique et par courrier papier. *Semen. Revue de sémio-linguistique des textes et discours*, 20:51–64. (Cité en page 31.)

Marcoccia, M. (2005b). La communication médiatisée par ordinateur. Problèmes de genres et de typologie. *In Journée d'étude "Les genres de l'oral" le 18 avril 2003*, Lyon. Université Lumière Lyon 2. (Cité en pages 38 et 42.)

Marcoccia, M. (2006). Les webforums des partis politiques français : quels modèles de discussion politique ? *Mots. Les langages du politique*, 80:49–60. (Cité en page 33.)

Maurais, J. (2003). *Analyse linguistique de 4000 courriels*. Conseil supérieur de la langue française, Québec. (Cité en page 31.)

McBride, K. (2009). Social-networking sites in foreign language classes : Opportunities for re-creation. *In* Lomicka, L. et Lord, G., éditeurs : *The next generation : Social networking and online collaboration in foreign language learning*, pages 35–58. (Cité en page 36.)

McEnery, T. et Hardie, A. (2012). *Corpus Linguistics : Method, Theory and Practice*. Cambridge University Press. (Cité en pages 66, 129 et 130.)

McEnery, T. et Wilson, A. (2001). *Corpus linguistics : an introduction*. Edinburgh University Press. (Cité en page 63.)

Meisner, C. (2010). A corpus analysis of intra- and extralinguistic factors triggering ne-deletion in phonic french. *In* Neveu, F., Muni Toke, V., Durand, J., Klingler, T., Mondada, L. et Prévost, S., éditeurs : *Congrès Mondial de Linguistique française*, pages 1943–1962. Institut de Linguistique Française. (Cité en pages 251 et 265.)

Melançon, B. (1996). *Sevigne@Internet : remarques sur le courrier électronique et la lettre*. Les Editions Fides, Montréal. (Cité en page 32.)

Mel'čuk, I. (à paraitre). Tout ce que nous voulions savoir sur les phrasèmes, mais .... *Cahiers de lexicologie, revue internationale de lexicologie et de lexicographie*. (Cité en page 141.)

Merle, G., Perret, R., Vince, J. et Juilliard, C. (1987). *Néologie lexicale n°2 Français*. GRIL, Université Paris VII, Paris. (Cité en page 143.)

Meunier, J. P. et Peraya, D. (2004). *Introduction aux théories de la communication*. De Boeck, Bruxelles, nouvelle édition. (Cité en page 37.)

Meyer, C. F. (2002). *English corpus linguistics : An introduction*. Cambridge University Press. (Cité en pages 127 et 131.)

Mondada, L. (1999). Formes de séquentialité dans les courriels et les forums de discussion. Une approche conversationnelle de l'interaction sur Internet. *Alsic. Apprentissage des Langues et Systèmes d'Information et de Communication*, 2(1): 3–25. (Cité en pages 21 et 33.)

Moore, G. (1990). Structural determinants of men's and women's personal networks. *American sociological review*, 55:726–735. (Cité en page 22.)

Moreau, M. (1986). Les séquences préformées : entre les combinaisons libres et les idiomatismes. Les cas de la négation avec et sans *ne*. *Le Français moderne*, (54):137–160. (Cité en page 253.)

Moreau, M. (1997). *Sociolinguistique : les concepts de base*, volume 218. Mardaga. (Cité en page 155.)

Moyal, A. (1992). The gendered use of the telephone : an Australian case study. *Media, Culture & Society*, 14:51–72. (Cité en page 22.)

Moynihan, B., Kabadayi, S. et Kaiser, M. (2010). Consumer acceptance of sms advertising : a study of american and turkish consumers. *International Journal of Mobile Communications*, 8(4):392–410. (Cité en page 20.)

Mucchielli, A., éditeur (1996). *Dictionnaire des méthodes qualitatives en sciences humaines et sociales*. Armand Colin, Paris. (Cité en page 119.)

Myers-Scotton, C. (1993). *Social motivations for codeswitching : Evidence from Africa*. Clarendon Press, Oxford. (Cité en page 155.)

Nagarajan, R. (2003). SMS gives hearing impaired a lifeline. *Deaf Today*. (Cité en page 22.)

Ochs, E. (1979). Transcription as theory. *In* Ochs, E. et Schieffelin, B., éditeurs : *Developmental pragmatics*, pages 43–72. Academic Press, New york. (Cité en page 92.)

Panckhurst, R. (1997). La communication “médiatisée” par ordinateur ou la communication “médiée” par ordinateur ? *Terminologies nouvelles*, 17:56–58. (Cité en pages 18, 19 et 41.)

Panckhurst, R. (1998). Analyse linguistique du courrier électronique. *In* Guégen, N. et Tobin, L., éditeurs : *Communication, société et Internet*, pages 47–60. Paris. (Cité en pages 21, 22, 31, 38, 56, 58, 249 et 265.)

Panckhurst, R. (1999). Analyse linguistique assistée par ordinateur du courriel. *In* Anis, J., éditeur : *Internet, communication et langue française*, pages 55–70. Hermés, Paris. (Cité en pages 3, 31 et 58.)

Panckhurst, R. (2007). Discours électronique médié : quelle évolution depuis une décennie ? *In* Gerbault, J., éditeur : *La langue du cyberspace : de la diversité aux normes.*, pages 121–136. L'Harmattan, Paris. (Cité en pages 18 et 56.)

Panckhurst, R. (2009). Short Message Service (SMS) : typologie et problématiques futures. *In* Arnavielle, T., éditeur : *Polyphonies, pour Michelle Lanvin*, pages 33–52. Université Paul-Valéry Montpellier 3,, Montpellier. (Cité en pages 21, 61 et 144.)

Paragas, F. (2003). Dramatextism : Mobile telephony and people power in the Philippines. *In* Nyiri, K., éditeur : *Mobile democracy : Essays on society, self, and politics*, pages 259–283. Passagen Verlag, Vienne. (Cité en page 20.)

Paumier, S. (2010). UNITEX 2.0 User Manual. Electronic version. (Cité en page 130.)

Perez, C. (2009). Le web 2.0 dans l'enseignement des langues étrangères. *Le français dans le monde*, 363:27–29. (Cité en page 36.)

Petitjean, C. (2009). *Représentations linguistiques et plurilinguisme*. Thèse de doctorat, Universités de Provence et de Neuchâtel. (Cité en page 97.)

Pierozak, I. (2003). Le français tchaté : un objet à géométrie variable ? *Langage et société*, 104:123–144. (Cité en pages 29 et 30.)

Piron, M. (1955). Français paralléles de belgique. *Bulletin de la Société belge des Professeurs de français*, 4(3):3–14. (Cité en page 185.)

Piron, M. (1979). Le français de Belgique. *In* Valdman, A., éditeur : *Le français hors de France*, pages 201–221. Honoré Champion, Paris. (Cité en page 185.)

Plourde, M., Duval, H. et Georgeault, P. (2001). *Le français au Québec : 400 ans d'histoire et de vie*. Fides, Montréal. (Cité en page 198.)

Pohl, J. (1972). Ne et les enfants. *In L'homme et le signifiant*, pages 107–111. Paris : Nathan. (Cité en pages 249, 259 et 260.)

Poirier, C. (2011). Base de données lexicographiques du Québec. Sur le site de la Base de données lexicographiques panfrancophones. (Cité en page 198.)

Polin, L. (1993). Global Village as Virtual Community (On Writing, Thinking, and Teacher Education). *Writing Notebook : Visions for Learning*, 11(2):14–16. (Cité en page 37.)

Pöll, B. (2001). *Francophonies périphériques : histoire, statut et profil des principales variétés du français hors de France*. L'Harmattan. (Cité en pages 183, 185, 198 et 202.)

Pombo, E. L. (2001). La construction visuelle de la parole dans le livre médiéval. *Diogène*, 196:40–52. (Cité en page 51.)

Pooley, T. (1996). *Chtimi : The urban vernaculars of northern France*, volume 2. Multilingual Matters Ltd. (Cité en page 101.)

Poplack, S. (1993). Variation theory and language contact. *American dialect research*, pages 251–286. (Cité en page 127.)

Power, M. et Power, D. (2004). Everyone here speaks TXT : Deaf people using SMS in Australia and the rest of the world. *Journal of Deaf Studies and Deaf Education*, 9(3):333–343. (Cité en page 22.)

Prochasson, E., Viard-Gaudin, C. et Morin, E. (2007). Language models for handwritten short message services. *In Proceedings of the 9th International Conference on Document Analysis and Recognition*, pages 83–87, Curitiba. (Cité en page 49.)

Prudent, L. (2001). La reconnaissance officielle des créoles et l'aménagement d'un capes dans le système éducatif de l'outre-mer français. *Études créoles*, 24(1):80–109. (Cité en page 193.)

Queneau, R. (1959). *Zazie dans le métro*. Gallimard, Edition Folio, 2007, Paris. (Cité en page 229.)

Rebeyrolle, J. et Tanguy, L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire*, 25:153–174. (Cité en page 233.)

Reid, E. (1994). Cultural formations in text-based virtual realities. Master's thesis, Université de Melbourne. (Cité en page 37.)

Renaud (1983). Dès que le vent soufflera. Titre de l'album "Morgane de toi", sorti en 1983. (Cité en page 215.)

Renouf, A. (1993). A word in time : first findings from the investigation of dynamic text. *In* Aarts, J., De Haan, P. et Oostdijk, N., éditeurs : *English language corpora : design, analysis and exploitation*, pages 279–288. Rodopi, Amsterdam. (Cité en page 73.)

Resnyansky, L. (2002). Computer-mediated communication in higher education : Educators'agency in relation to technology. *Journal of Educational Enquiry*, 3(1): 35–59. (Cité en page 23.)

Rey, A. et Chantreau, S. (1989). *Dictionnaire des expressions et locutions*. Le Robert. (Cité en page 141.)

Rey, A., Rey-Debove, J. et Robert, P. (2009). *Le nouveau Petit Robert*. Dictionnaires Le Robert, Paris. (Cité en page 3.)

Rey, A., Rey-Debove, J. et Robert, P., éditeurs (2011). *Le Petit Robert de la langue française 2012*. Dictionnaires le Robert, Paris. (Cité en pages 99, 145, 155 et 189.)

Rheingold, H. (1993). *The virtual community : homesteading on the electronic frontier*. Addison Wesley. (Cité en page 20.)

Riandey, B. et Widmer, I. (2009). Introduction aux sondages à l'usage du plus grand nombre, en ligne : `http://statistix.fr/IMG/pdf/Riandey-Widmer_sondages_.pdf`. (Cité en page 111.)

Rice, R. E. (1989). Issues and concepts in research on computer-mediated communication systems. *Communication yearbook*, 12:436–476. (Cité en page 21.)

Rice, R. E., Hughes, D. et Love, G. (1989). Usage and outcomes of electronic messaging at an R&D organization : Situational constraints, job level, and media awareness. *Office : Technology and People*, 5(2):141–161. (Cité en page 20.)

Rivenc, P. (1971). Lexique et langue parlée. *In La grammaire du français parlé*, pages 51–70. Hachette, Paris. (Cité en page 139.)

Rivens Mompean, A. (2007). Pratiques langagières sur un forum pédagogique en anglais. *In* Gerbault, J., éditeur : *La langue du cyberespace : de la diversite aux normes*, pages 221–238. L'Harmattan, Paris. (Cité en page 22.)

Rivière, C. A. (2002). La pratique du mini-message. Une double stratégie d'extériorisation et de retrait de l'intimité dans les interactions quotidiennes. *Réseaux*, 112-113:139–168. (Cité en page 20.)

Ruan, L. (2012). Lexical Formation Rules and Chatting Language Online in English. *Theory and Practice in Language Studies*, 2(1):151–155. (Cité en page 144.)

Rubin, D. (2001). Phones keep deaf people in touch with text : Cellular feature serves as a lifeline in Europe. *The Detroit Free Press*. (Cité en page 22.)

Sablayrolles, J. (2000). *La néologie en français contemporain : examen du concept et analyse de productions néologiques récentes*, volume 4. Champion, Coll. Lexica, Paris. (Cité en pages 143, 144, 148, 149 et 152.)

Sablayrolles, J. (2008). Néologie et dictionnaire(s) comme corpus d'exclusion. *In* Sablayrolles, J., éditeur : *Néologie et terminologie dans les dictionnaires (actes de la journée des dictionnaires organisée par Jean Pruvost)*, pages 19–36. Champion. (Cité en page 144.)

Sablayrolles, J. (2009). Interview RFI par Yvan Amar, Émission : La danse des mots. (Cité en pages 143 et 146.)

Sankoff, G. et Vincent, D. (1977). L'emploi productif du ne dans le français parlé à montréal. *Le Français moderne*, 45:243–256. (Cité en pages 249 et 259.)

Sautot, J. P. (2000). *Utilisation de l'orthographe et d'autres indices dans la construction du sens en lecture*. Thèse de doctorat en sciences du langage, Université Stendhal Grenoble. (Cité en page 49.)

Schneider, D. (1997). La communication médiatisée par ordinateur (CMO) dans l'éducation. "État des lieux" et avenir. EDUCADOR 97, http://tecfa.unige.ch/tecfa/research/CMC/brazil97/psfiles/internet-cmo.fm.pdf. (Cité en pages 23 et 387.)

Schneider-Hufschmidt, M. (2005). Usability Issues of Sending Text Messages. *In* Harper, R., Palen, L. A. et Taylor, A. S., éditeurs : *The Inside Text*, pages 223–236. Springer, Dordrecht. (Cité en page 24.)

Shafie, L. A., Darus, N. A. et Osman, N. (2010). SMS Language and College Writing : The languages of the College Texters. *International Journal of Emerging Technologies in Learning (iJET)*, 1(5):26–31. (Cité en page 4.)

Siebenhaar, B. (2005). Varietätenwhal und code switching in deutschschweizer chatkanälen. quantitative und qualitative analysen. *Online (Networx 43). Retrieved June*, 12. (Cité en page 183.)

Slakta, D. (1985). Grammaire de texte : synonymie et paraphrase. *In* Fuchs, C., éditeur : *Aspects de l'ambiguité et de la paraphrase dans les langues naturelles*, pages 123–140. Peter Lang, Berne. (Cité en page 65.)

Sourdot, M. (2004). La dynamique du français actuel. *Neuphilologische Mitteilungen*, 105(1):85–94. (Cité en page 220.)

Spencer, S. (2008). Le français au québec. *Le Québec à l'aube du nouveau millénaire : entre tradition et modernité*, pages 108–116. (Cité en page 198.)

SPSS (2010). IBM SPSS Statistics. Version 19. IBM Company. (Cité en page 131.)

Stähli, A., Dürscheid, C. et Béguelin, M.-J. (2011). (en cours de publication). SMS-Kommunikation in der Schweiz : Korpusdaten, Literaturüberblick und Forschungsfragen. *In* Stähli, A., Dürscheid, C. et Béguelin, M.-J., éditeurs : *SMS-Kommunikation in der Schweiz : Sprach- und Varietätengebrauch*. Linguistik Online. (Cité en page 87.)

Stark, E. (2011). La morphosyntaxe dans les SMS suisses francophones : Le marquage de l'accord sujet – verbe conjugué. *Linguistik online*, 48(4):35–47. (Cité en pages 4, 220 et 224.)

Sutton, V. (2009). *SignWriting Basics*. SignWriting Press. (Cité en page 106.)

Tagg, C. (2009). *A corpus analysis of SMS text messaging*. Thèse de doctorat, University of Birmingham. (Cité en page 130.)

Tagliamonte, S. (2006). *Analysing sociolinguistic variation*. Cambridge Univ Pr. (Cité en page 127.)

Tatossian, A. (2008). Typologie des procédés scripturaux des salons de clavardage en français chez les adolescents et les adultes. *Actes du 1e Congrès Mondial de Linguistique Française*, pages 2337–2352. (Cité en pages 29, 30, 49, 55 et 56.)

Tatossian, A. (2011). Clavardage et orthographe. *Correspondance*, (16). (Cité en page 5.)

Thelwall, M., Wilkinson, D. et Uppal, S. (2009). Data mining emotion in social network communication : Gender differences in MySpace. *Journal of the American Society for Information Science and Technology*, 61(1):190–199. (Cité en page 36.)

Thibault, A. et Knecht, P. (2004). *Dictionnaire suisse romand : particularités lexicales du français contemporain*. Zoé. (Cité en page 203.)

Thurlow, C. (2003). Generation txt ? The sociolinguistics of young people's text-messaging, En ligne : `http://faculty.washington.edu/thurlow/papers/Thurlow%282003%29-DAOL.pdf`. *Discourse Analysis Online*, 1. (Cité en pages 3, 4, 21, 22 et 61.)

Titone, R. (1972). *Le bilinguisme précoce*. Dessart, Bruxelles. (Cité en page 116.)

Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. John Benjamins, Amsterdam. (Cité en pages 129 et 130.)

Tournier, M. (2002). Des mots en politique : Préfixes branchés de la communication. *Mots*, 68:131–138. (Cité en page 143.)

Turkle, S. (1992). Living in the MUD : Multiplicity and identity in virtual reality. *In Proceedings of the 91st Annual Meeting of the American Anthropological Association*, San Francisco. (Cité en page 37.)

Turoff, M. et Hiltz, S. R. (1977). Meeting Through Your Computer. *IEEE Spectrum*, page May. (Cité en page 20.)

Ursini, F. (2005). La lingua dei giovani e i nuovi media : gli SMS. *In* Fusco, F. et Marcato, C., éditeurs : *Forme di comunicazione giovanile*. Rome. (Cité en page 129.)

Valdman, A. (1979). *Le français hors de France*. Honoré Champion, Paris. (Cité en pages 183 et 198.)

Van Campenhoudt, M. (2002). Linguistique de corpus et étude de vocabulaires spécialisés. Séminaire donné à St-Denis, 8 janvier 2002, Université Paris 8. (Cité en page 123.)

Van Compernolle, R. (2007). The variable use of ne in public French–language synchronous chat. *In* Gerbault, J., éditeur : *La langue du cyberespace : de la diversité aux normes*, pages 251–264. L'Harmattan. (Cité en pages 249, 256 et 265.)

Van Compernolle, R. et Williams, L. (2007). De l'oral à l'électronique : la variation orthographique comme ressource stylistique et pragmatique dans le français électronique. *Glottopol*, 10:56–69. (Cité en page 10.)

Van Compernolle, R. A. (2011). Use and variation of French diacritics on an Internet dating site journal. *Journal of French Language Studies*, 21(2):131–148. (Cité en page 22.)

Veale, T. et Butnariu, C. (2010). Harvesting and Understanding On-line Neologisms. *In* Onysko, A. et Michel, S., éditeurs : *Cognitive Perspectives on Word Formation*, pages 393–416. Mouton De Gruyter. (Cité en page 144.)

Véronis, J. et Guimier De Neef, E. (2006). Le traitement des nouvelles formes de communication écrite. *In* Sabah, G., éditeur : *Compréhension automatique des langues et interaction*, pages 227–248. Hermès Science, Paris. (Cité en pages 18, 19, 48, 58 et 60.)

Vinay, J. (1971). Les cadres de la phrase. *In* Rigault, A., éditeur : *La grammaire du français parlé*, pages 120–135. Hachette, Paris. (Cité en page 248.)

Volckaert-Legrier, O., Bert-Erboul, A. et Bernicot, J. (2006). Raconter par courrier électronique : une étude de l'usage de l'orthographe chez les adolescents. *Le Langage et l'Homme*, 41(2):141–156. (Cité en page 21.)

Vold Lexander, K. (2007). Langues et SMS au Sénégal. Le cas des étudiants de Dakar. *In* Gerbault, J., éditeur : *La langue du cyberespace : de la diversité aux normes*, pages 59–67. L'Harmattan, Paris. (Cité en page 22.)

Vold Lexander, K. (2009). "La communication médiatisée par les technologies de l'information et de la communication : la porte d'accès au domaine de l'écrit pour les langues africaines ?". *In* Brock-Utne, B. et Skattum, I., éditeurs : *Languages and Education in Africa. A Comparative and Transdisciplinary Analysis.*, chapitre 17, pages 289–299. Symposium, Oxford. (Cité en pages 18 et 22.)

Von Münchow, P. (2004). Le discours rapporté dans un forum de discussion sur l'internet. *Les Carnets du Cediscor*, 8:91–111. (Cité en page 33.)

Vygotski, L. (1997). *Pensée et langage*. Terrains/Editions Sociales, Paris. (Cité en page 18.)

Watrin, P. (2007). Collocations et traitement automatique des langues. *In Actes du 26e Colloque international sur le lexique et la grammaire*, pages 1530–1536. (Cité en page 141.)

Watrin, P. et François, T. (2011). An N–gram frequency database reference to handle MWE extraction in NLP applications. *In Proceedings of the ACL 2011 Workshop on MultiWord Expres- sions : from Parsing and Generation to the Real World, Portland*. Association for Computational Linguistics. (Cité en page 141.)

Wei, R., Xiaoming, H. et Pan, J. (2010). Examining user behavioral response to sms ads : Implications for the evolution of the mobile phone as a bona-fide medium. *Telematics and Informatics*, 27(1):32–41. (Cité en page 20.)

Weininger, M. J. et Shield, L. (2004). Promoting oral production in a written channel : an investigation of learner language in moo. *Computer Assisted Language Learning*, 16(4):329—-349. (Cité en page 59.)

Weiser, S. (2010). *Repérage et typage d'expressions temporelles pour l'annotation sémantique automatique de pages Web-Application au e-tourisme*. Thèse de doctorat, Université de Paris X - Nanterre. (Cité en page 24.)

Weiser, S., Cougnon, L.-A. et Watrin, P. (2011). Temporal Expressions Extraction in SMS Messages. *In Workshop on Information Extraction and Knowledge Acquisition Collocated with RANLP'2011*, Hissar, Bulgaria. (Accepted paper). (Cité en pages 24 et 85.)

Werry, C. (1996). Linguistic and interactional features of internet relay chat. *In* Herring, S., éditeur : *Computer-Mediated Communication : Linguistic, Social and Cross-Cultural Perspectives*, pages 47–63. John Benjamins, Amsterdam. (Cité en page 29.)

Wolff, A., éditeur (2010). Nathan. (Cité en pages 7 et 8.)

Woodruff, M. et Waldorf, J. (1995). Multi-user dungeons enter a new dimension : Applying recreational practices for education goals. *Electronic Journal of Communication*, 5(4). (Cité en page 37.)

Yates, S. (1996). Oral and written linguistic aspects of computer conferencing : A corpus-based study. *In* Herring, S. C., éditeur : *Computer-mediated communication : Linguistic, social and cross-cultural perspectives*. John Benjamins, Philadelphie. (Cité en pages 56 et 57.)

Ziegler, E. (2005). Die bedeutung von interaktionsstatus und interaktionsmodus für die dialekt-standard-variation in der chatkommunikation. *In* Eggers, E., Schmidt, J. et D., S., éditeurs : *Neue Dialekte – moderne Dialektologie*, pages 719–745. Steiner. (Cité en page 183.)

Zongo, B. (1993). *Stratégies de communication et individuation sociolinguistique chez les étudiants burkinabé à Ouagadougou et à Paris*. Thèse de doctorat. (Cité en page 161.)

Annexe A

Annexes

A.1 Annexes au Chapitre 1

Figure A.1 – *Typologie des logiciels de l'ACAO (de Vries, 2001)*

Fonction pédagogique	Type de logiciel	Théorie	Tâche	Connaissances
Présenter de l'information	tutoriel	cognitiviste	lire	présentation ordonnée
Dispenser des exercices	exercices répétés	behavioriste	faire des exercices	association
Véritablement enseigner	tuteur intelligent	cognitiviste	dialoguer	représentation
Captiver l'attention et la motivation de l'élève	jeu éducatif	principalement behavioriste	jouer	
Fournir un espace d'exploration	hypermédia	cognitiviste constructiviste	explorer	présentation en accès libre
Fournir un environnement pour la découverte de lois naturelles	simulation	constructiviste cognition située	manipuler, observer	modélisation
Fournir un environnement pour la découverte de domaines abstraits	micro-monde	constructiviste	construire	matérialisation
Fournir un espace d'échange entre élèves	apprentissage collaboratif	cognition située	discuter	construction de l'élève

TABLE A.1 – *Tableau récapitulatif des facteurs caractérisant les CMO (Herring : 2007, amélioré par nous : réponses possibles)*

Critère	**Définition**	**Réponses possibles**
Anonymat	Le service permet-il l'usage de pseudo ou l'identité du destinateur est-elle révélée ?	Anonyme (pseudonyme) versus identité révélée
But	Quel est le but du groupe ou du forum ? Quel est le but de l'interaction : demander de l'information, développer des relations privées ou professionnelles, s'amuser, etc. ?	Champ libre avec les caractéristiques
Canaux de communication	Le service permet-il uniquement le texte ou aussi des graphiques, de l'audio, de la vidéo, etc.	Texte, image, audio et/ou vidéo
Caract. des participants	Caractéristiques démographiques, expériences de chacun avec la CMO, rôle et statut de chacun dans la « vie réelle » et dans le « monde virtuel », motivations, etc.	Champ libre avec les caractéristiques
Citation	Le service met-il à disposition des usagers des commandes permettant de faciliter la citation d'un message précédant ?	Oui versus non
Délai de réponse	Continuum entre l'immédiateté (Muds, messagerie instantanée) et un délai relativement long (courrier électronique, forums) en passant par le délai moyen dépendant d'un certain nombre de facteurs (sms, réseautage social).	Situer sur le continuum
Filtres	Le service met-il à disposition des usagers des commandes permettant de filtrer l'arrivée de messages ?	Oui versus non
Format	Dans quel ordre les messages apparaissent-ils ? Où sont placés les nouveaux messages ? Qu'arrivent-ils lorsqu'une fenêtre est remplie de messages ? Quelles informations apparaissent dans la fenêtre des messages ? Y a-t-il possibilité de relecture avant envoi du message ?	Champ libre avec les caractéristiques

Normes	3 types : normes d'organisation (protocoles d'admission d'un nouveau membre, potentiel modérateur, comment le groupe a été créé), normes sociales (standards de comportement, sorte de nétiquette souvent disponibles en charte mais pouvant aussi être implicites) et normes linguistiques (tout type de conventions relatives à la langue, par exemple l'interdiction d'utiliser un langage abrévié). Ces trois normes forment ce que l'on nomme aujourd'hui la « Netiquette ».	Champ libre avec les caractéristiques pour les 3 normes
Privé versus public	Le service met-il à disposition des usagers des commandes permettant des conversations privées et publiques ?	Privé, public ou privé et public
Structure de conversation	Combien de participants peuvent être inclus dans la conversation ?	Champ libre avec les caractéristiques
Sujet ou thème	Sur les forums par exemple, un sujet en particulier est traité. Sur les Muds, un thème temporel ou géographique est souvent pris en compte (par exemple un village médiéval). Dans une conversation de type sms ou courrier électronique, il peut y avoir plusieurs sujets traités en même temps.	Champ libre avec les caractéristiques
Survivance de la transcription	Combien de temps les messages sont-ils maintenus par le service après avoir été échangés ?	Persistante par défaut versus effacée par défaut
Synchronie	Le service nécessite-t-il que les multiples usagers soient connectés en même temps pour communiquer ?	Synchronie versus diachronie
Taille maximale	Nombre maximal de caractères par message	Réponse numérique versus illimité
Ton	Le ton est la manière de s'exprimer ; il peut être perçu directement dans le texte ou bien livré par les émoticons accompagnant le texte.	Degré de sérieux, degré de formalité, degré de contentieux et degré de coopération
Transmission du message	La plupart des CMO sont message par message, excepté VAX, UNIX talk et le mode séparation d'écran de ICQ.	Caractère par caractère, ligne par ligne ou message par message

TABLE A.2 – *Équivalents anglais de la terminologie relative aux phénomènes graphiques du sms*

Phénomènes en français	**Equivalent(s) anglais**	**Exemple anglais**
Apocope	Shortening, back clipping, apocopation	*darlin* pour *darling* *hav* pour *have*
Syncope	Contraction, middle clipping, syncope	*cming* pour *coming* *txt* pour *text*
Aphérèse	Fore-clipping, aphaeresis	*morrow* pour *tomorrow*
Siglaison	Initialism	*Diachronie*
Acronymie	Acronym	*aka* pour *also known as*
Phonétisation par lettre	Letter homophone	*u* pour *you*
Phonétisation par chiffre	Number homophone	*2night* pour *tonight*
Phonétisation par signe	Sign homophone	*to @end* pour *to attend*
Graphies proches de la phonétique réelle	Non-conventional spelling	*luve* pour *love*
Graphies proches d'un accent régional	Accent stylization	*Gonna com* pour *going to come*
Graphies proches d'un effet de style	Stylish talk	*da* pour *the*
Simplification syntaxique	Loss of grammatical words	*Give me ring* pour *give me a ring*
Neutralisation de la punctuation	Loss of punctuation	*I dont know* pour *I don't know*

TABLE A.3 – *Récapitulatif des facteurs caractéristiques du chat et de la MI*

Critère	Chat	MI	Critère	Chat	MI
Anonymat	Anonyme (pseudonyme)	Identité révélée	Privé versus Public	Privé et public	Privé
But	Discussion	Id.	Structure de conversation	Groupe	2 ou plus si invitation
Canaux de communication	Texte + envoi de fichiers graphiques, audios, vidéos, dépend du service	Id.	Sujet ou thème	Diversité des thèmes. Certains groupes sont formés autour d'un sujet qui devient le sujet de la conversation	Diversité des thèmes
Caract. des participants	Selon le service, quelques informations peuvent être communiquées : avatar [1], statut par rapport à l'activité de discussion [2], etc.	Id.	Survivance de la transcription	Effacée par défaut ou maintenue par option	Id.
Citation	Oui, dépend du service	Id.	Synchronie	Synchronie	Syncronie
Délai de réponse	Immédiateté	Immédiateté	Taille maximale	Oui, le nombre dépend du service employé	Id.
Filtres	Non	Non	Ton	Variable	Variable
Format	Nouveaux messages arrivent en-dessous des précédents, écran défile lorsque rempli, relecture possible. Fenêtre comprend pseudo ou adresse mail et informations temporelles	Id. sauf relecture pour anciens systèmes	Transmission du message	Message par message	Id. sauf caractère par caractère pour anciens systèmes
Normes	l'admission peut se faire via modérateur, le langage abrégé est quelquefois prohibé, une conduite éthique (racisme, sexisme, pornographie, etc.) est presque toujours prohibée	Peu de normes			

TABLE A.4 – *Tableau récapitulatif des facteurs caractérisant le courrier électronique*

Anonymat	Identité révélée (plus rare : ouverture de compte avec pseudo)	Privé versus Public	Privé
But	Communication variée	Structure de conversation	Échange basique comprend un destinateur et un destinataire. Le nombre de destinataires peut ensuite varier. Des destinataires « cachés », ou « recevant une copie » peuvent également être ajoutés
Canaux de communication	Texte + pièce jointe (image, texte, etc. : dépend du service)	Sujet ou thème	Plusieurs sujets peuvent être traités en même temps
Caract. des participants	NC	Survivance de la transcription	Persistante par défaut
Citation	Oui	Synchronie	Diachronie
Délai de réponse	Long	Taille maximale	La taille du mail et de ses pièces jointes, calculée en mégaoctets, varie entre 5 et 25 Mo, selon le service
Filtres	Oui	Ton	Variable avec degré de formalité plus élevé que dans les autres types
Format	Nouveaux messages arrivent au-dessus des précédents, deuxième page créée lorsque écran rempli, relecture possible. Fenêtre comprend adresse mail, nom, informations temporelles et objet du message	Transmission du message	Message par message
Normes	Pas de normes linguistiques ou sociales, bien que le registre est généralement plus soutenu que dans les autres type de la Cémo. Norme d'organisation : un nouveau contact peut envoyer un courriel mais son message, selon le service et le filtre, est renvoyé ou non dans le courrier indésirable		

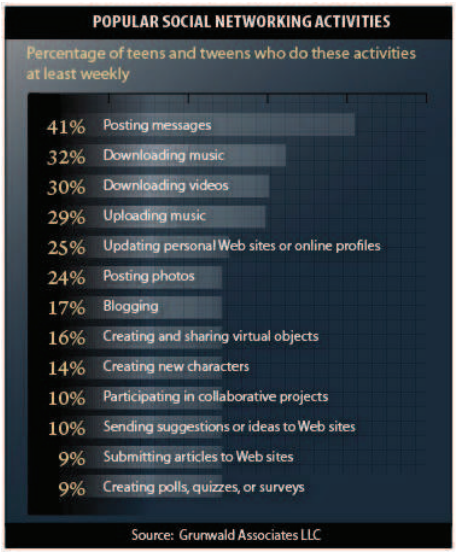


FIGURE A.2 – *Fréquence des activités sur les services de réseautage social (source : Grunwald Associates LLC, http ://www.grunwald.com/)*

TABLE A.5 – *Tableau récapitulatif des facteurs caractérisant les forums de discussion*

Anonymat	Anonyme (pseudonyme)	Privé versus Public	Public
But	Discussion	Structure de conversation	Groupe, nombre de participants illimité
Canaux de communication	Texte avec émoticônes (souvent évolués) et image, selon le service	Sujet ou thème	Un sujet en particulier
Caract. des participants	Quelques caractéristiques peuvent être communiquées : lieu, âge, avatar, nombre de contributions sur le forum, adresse électronique, etc.	Survivance de la transcription	Persistante par défaut, pour une période donnée (souvent quelques années)
Citation	Oui	Synchronie	Diachronie
Délai de réponse	Long	Taille maximale	Oui, le nombre dépend du service employé
Filtres	Oui, sous la forme d'une modération	Ton	Variable, un ton agressif est toutefois mal perçu sur la plupart des forums et peut être modéré
Format	Nouveaux messages arrivent au-dessus des précédents, écran défile lorsque rempli, relecture possible. Liste de messages comprend sujet, pseudo de l'initiateur du sujet et informations temporelles	Transmission du message	Message par message
Normes	L'admission se fait généralement via inscription, le langage abrégé est souvent prohibé, une conduite non éthique (racisme, sexisme, pornographie, etc.) est souvent prohibée		

TABLE A.6 – *Tableau récapitulatif des facteurs caractérisant les sites de réseautage social*

Anonymat	Identité révélée. L'anonymat (sous pseudonyme) est une option mais elle va à l'encontre du but premier des SRSs	Privé versus Public	Privé par défaut, mais le profil peut être (semi-) public si l'utilisateur le désire
But	Discussion, comptes rendus de vie	Structure de conversation	Groupe pour le statut, un seul destinataire pour les commentaires, bien que le groupe entier puisse ensuite participer à l'échange
Canaux de communication	Texte, photos, audios, vidéos, dépend du service	Sujet ou thème	Plusieurs sujets peuvent être traités en même temps, bien que le statut et les commentaires portent généralement sur un seul sujet
Caract. des participants	Selon le service et les options choisies par l'utilisateur : toutes les informations personnelles, sociolinguistiques, curriculum vitae, centres d'intérêt, activités, etc.	Survivance de la transcription	Maintenue par défaut
Citation	Non	Synchronie	Diachronie
Délai de réponse	Moyen, en fonction des utilisateurs et du type de réseau	Taille maximale	Oui, le nombre dépend du service employé et du type de message : statut (140 caractères) et commentaires (pouvant aller jusqu'à 1000 caractères ou plus)
Filtres	Oui, dans la plupart des services, mais pas par défaut	Ton	Variable (les émotions y sont généralement exacerbées)
Format	Nouveaux messages arrivent au-dessus des précédents, écran défile lorsque rempli, puis création d'une page supplémentaire, relecture possible. Fenêtre comprend profil entier de l'utilisateur	Transmission du message	Message par message

Normes	Admission par inscription, le langage abrégé est permis, une conduite éthique (racisme, sexisme, pornographie, etc.) est encouragée		



FIGURE A.3 – *Exemple de fenêtre Mud contenant plusieurs zones de texte (World of Warcraft, 2001)*

TABLE A.7 – *Tableau récapitulatif des facteurs caractérisant les Muds*

Anonymat	Anonymat (sous pseudonyme)	Privé versus Public	Privé par défaut, mais le profil peut être (semi-) public si l'utilisateur le désire
But	Remplir une mission (demande d'aide, etc.) et discuter	Structure de conversation	Groupe (tous les joueurs de la partie)
Canaux de communication	Texte et audio	Sujet ou thème	Ordres, demandes d'aide, présentation de la stratégie ou communication simple avec sujets variés
Caract. des participants	Selon le service et les options choisies par l'utilisateur, quelques caractéristiques (âge, sexe, etc.) peuvent être affichées. Par défaut, ce sont les caractéristiques du personnage virtuel qui seront présentées	Survivance de la transcription	Effacée par défaut
Citation	Non	Synchronie	Synchronie
Délai de réponse	Immédiateté	Taille maximale	Oui, le nombre dépend du service employé
Filtres	Non	Ton	Variable (les émotions sont liées aux activités du jeu)
Format	Nouveaux messages arrivent au-dessus ou en-dessous des précédents (dépend du service et de l'emplacement de la boîte de dialogue), écran défile lorsque rempli, relecture possible. Fenêtre comprend profil entier du personnage	Transmission du message	Message par message
Normes	Admission par inscription, le langage abrégé est permis, une conduite éthique (racisme, sexisme, pornographie, etc.) est encouragée		

TABLE A.8 – *Tableau récapitulatif des facteurs caractérisant l'eSMS*

Anonymat	Identité révélée	Privé versus Public	Privé
But	Communication variée	Structure de conversation	Échange basique comprend un destinateur et un destinataire. Le nombre de destinataires peut ensuite varier, en fonction notamment du service offert par le téléphone
Canaux de communication	Texte seul (signes de ponctuation peuvent former des émoticônes)	Sujet ou thème	Plusieurs sujets peuvent être traités en même temps
Caract. des participants	NC	Survivance de la transcription	Persistante par défaut
Citation	Dépend du service offert par le téléphone	Synchronie	Diachronie
Délai de réponse	Long	Taille maximale	Limite de 160 caractères par message (sans surcout) [3]
Filtres	Non (des numéros spécifiques peuvent être bloqués par le service)	Ton	Variable
Format	Nouveaux messages arrivent au-dessus des précédents, écran défile lorsque rempli, relecture possible. Fenêtre comprend nom et/ou numéro du destinataire et informations temporelles	Transmission du message	Message par message
Normes	Pas de normes linguistiques ou sociales. Norme d'organisation : un nouveau contact peut envoyer un sms sans être filtré, redirigé vers du courrier indésirable et sans demander d'inscription au préalable		

A.2 Annexes au Chapitre 2

A.2.1 Questionnaire belge

En Belgique, les questions de fait sont les suivantes (suivies du type de réponse) :

1. le sexe (réponse fermée : M / F) ;
2. l'âge (réponse fermée avec menu déroulant allant de 12 à 80 ans) ;
3. le code postal du domicile et celui du travail (réponse fermée avec zone libre demandant une valeur numérique respectant la structure des codes postaux belges) ;
4. le pays de domicile et celui du travail (réponse fermée comprenant une liste de propositions finie : France, Belgique, Luxembourg) ;
5. la langue maternelle (réponse ouverte : zone de commentaire limitée en nombre de caractères) ;
6. les 2e, 3e et 4e langues (réponse ouverte : zone de commentaire limitée en nombre de caractères) ;
7. le niveau d'étude (réponse fermée de 1 à 9 : primaire, secondaire inférieur général, secondaire inférieur technique, secondaire inférieur professionnel, secondaire supérieur général, secondaire supérieur technique, secondaire supérieur professionnel, supérieur non-universitaire, supérieur universitaire) ;
8. la profession et la profession 2 (réponse ouverte : zone de commentaire limitée en nombre de caractères) ;
9. le nombre d'années de pratique du sms (réponse fermée de 1 à 4 : moins d'1 an, 1 à 3 ans, 3 à 5 ans, plus de 5 ans) ;
10. le nombre de sms envoyés par semaine (réponse fermée de 1 à 6 : moins de 5, entre 5 et 10, entre 10 et 20, entre 20 et 50, entre 50 et 100, plus de 100) ;
11. la possession d'un dictionnaire d'aide à la rédaction dans le téléphone employé (réponse fermée : Oui / Non / je ne sais pas) ;
12. utilisation du dictionnaire d'aide à la rédaction (réponse fermée : Oui / Non) ;
13. le type de clavier du téléphone : nombre de touches du clavier (réponse fermée : valeur numérique) ;
14. l'envoi de sms via un service web (réponse fermée de 1 à 4 : presque tous les jours, régulièrement, rarement, jamais) ;
15. l'utilisation d'autres types de Cémo : chat, forum et courrier électronique (réponse fermée de 1 à 4 : presque tous les jours, régulièrement, rarement, jamais) ;
16. les types de destinataires des sms : famille, collègues, amis, concours/jeux, autres (réponse fermée avec zone libre demandant une valeur numérique de 1 à 5, 1 étant le plus fréquent ; autres : réponse ouverte : zone de commentaire limitée en caractères) ;

17. la compréhension de l'eSMS : « Comprenez-vous les SMS que vous recevez quand ils contiennent des codes ou des abréviations ? » (réponse fermée de 1 à 4 : oui facilement, oui mais je dois parfois décoder, oui mais difficilement, non c'est illisible) ;

En Belgique, la question d'opinion est la suivante (suivie du type de réponse) :

1. Zone de commentaire libre : « Pourriez-vous commenter la manière dont vous écrivez des SMS ? » (réponse ouverte : zone de commentaire limitée en nombre de caractères).

A.2.2 Questionnaire réunionnais

À l'île de la Réunion, les questions de fait, singulières par leur références récurrentes à la langue des signes [4] et au créole, sont les suivantes (suivies du type de réponse) :

1. le sexe (réponse fermée : M / F) ;
2. l'âge (réponse fermée avec zone libre demandant une valeur numérique) ;
3. le code postal du domicile et celui du travail (réponse fermée avec zone libre demandant une valeur numérique respectant la structure des codes postaux réunionnais) ;
4. la ou les langue(s) maternelle(s) (réponse fermée : langue des signes, créole, français et autre : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
5. le ou les autre(s) langue(s) parlée(s) couramment à la maison, entre amis et dans le monde du travail (pour chaque contexte, réponse fermée : langue des signes, créole, français et autre : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
6. le niveau d'étude (réponse fermée de 1 à 10 : école primaire, collège, lycée, bac, bac + 1, bac + 2, bac + 3, bac + 4, bac + 5, au-delà de bac + 5) ;
7. si niveau d'étude supérieur à l'école primaire, le type d'enseignement (réponse fermée : général ou professionnel) ;
8. la profession (réponse fermée de 1 à 11 : cadre, profession libérale, enseignant en secondaire, enseignant en primaire ou maternel, infirmier, commerçant/artisan, employé administratif, employé de commerce, ouvrier, sans activité professionnelle, retraité) ;
9. le nombre d'années de pratique du sms (réponse fermée de 1 à 4 : moins d'1 an, 1 à 3 ans, 3 à 5 ans, plus de 5 ans) ;
10. le nombre de sms envoyés par semaine (réponse fermée de 1 à 6 : moins de 5, entre 5 et 10, entre 10 et 20, entre 20 et 50, entre 50 et 100, plus de 100) ;
11. la possession d'un dictionnaire d'aide à la rédaction dans le téléphone employé (réponse fermée : Oui / Non / Je ne sais pas) ;

4. Ceci est dû à l'intérêt personnel de l'équipe de recherche en place.

12. l'utilisation d'autres types de Cémo : chat, forum et courrier électronique (réponse fermée, de 1 à 4 : presque tous les jours, régulièrement, rarement, jamais) ;
13. les types de destinataires des sms : famille, collègues, amis, concours/jeux, autres (réponse fermée, avec zone libre demandant une valeur numérique de 1 à 5, 1 étant le plus fréquent ; autres : réponse ouverte : zone de commentaire limitée en caractères) ;
14. la compréhension de l'eSMS : « Comprenez-vous les SMS que vous recevez quand ils contiennent des codes ou des abréviations ? » (réponse fermée de 1 à 4 : oui facilement, oui mais je dois parfois décoder, oui mais difficilement, non c'est illisible) ;
15. changement de langue : « Changez-vous de langue en fonction des destinataires ? » (réponse fermée : Oui / Non) ;
16. si changement de langues, la langue en fonction du destinataire : amis, connaissances, familles, collègues, autres (réponse fermée : langue des signes, créole, français et autre : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
17. alternance de langues : « Quand vous rédigez vos SMS, mélangez-vous les langues ? » (réponse fermée : Oui / Non) ;
18. si alternance de langues, précision des langues en question (réponse fermée : langue des signes, créole, français et autre :
    e avec zone de commentaire limitée en nombre de caractères) ;
19. pratique de l'eSMS : « Quand vous rédigez vos SMS, vous utilisez le langage SMS lorsque vous écrivez en : » (réponse fermée : aucun, créole, français et autre : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
20. temps de lecture : « Combien d'heures par jour consacrez-vous à la lecture (hormis la lecture de SMS et d'e-mails) lors de votre profession/scolarisation et de vos loisirs ? » (réponse fermée avec zone libre demandant une valeur numérique) ;
21. types de lecture (réponse fermée : choix entre articles de journaux, BD, des pages de sites Web, des textes de chansons, des livres et autres : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
22. lecture créole (réponse fermée : Oui / Non) ;
23. types de lecture créole (réponse fermée : choix entre articles de journaux, BD, des pages de sites Web, des textes de chansons, des livres et autre : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
24. pratique de l'écriture : « Lors de vos activités professionnelles/scolaires et de vos loisirs, consacrez-vous du temps à écrire sur papier, de quelque manière que ce soit ? » (réponse fermée : Oui / Non) ;

25. si pratique de l'écriture, précision du type (réponse fermée : choix entre petits mots, correspondance traditionnelle, journal intime et autre : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
26. si pratique de l'écriture, précision du but et de la fréquence (réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
27. pratique de l'écriture en créole (réponse fermée : Oui / Non) ;
28. si pratique de l'écriture en créole, précision (réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
29. question facultative en vue de récolte de dialogue : « Quelles personnes, avec qui vous correspondez par SMS, seraient susceptibles de participer à cette enquête ? Veuillez indiquer leurs numéros de téléphone séparés par des virgules.» (réponse ouverte avec zone de commentaire libre demandant des valeurs numériques).

À La Réunion, la question d'opinion est la suivante (suivie du type de réponse) :

1. Zone de commentaire libre : « Pourriez-vous commenter la manière dont vous écrivez des SMS ? » (réponse ouverte : zone de commentaire limitée en nombre de caractères).

A.2.3 Questionnaire québécois

Au Québec [5], les questions de fait sont les suivantes (suivies du type de réponse) :

1. le sexe (réponse fermée : M / F) ;
2. l'âge (réponse fermée avec zone libre demandant une valeur numérique) ;
3. le code postal du domicile (réponse fermée avec zone libre demandant une valeur numérique respectant la structure des codes postaux canadiens) ;
4. la ou les langue(s) maternelle(s) (réponse fermée : français, anglais, espagnol, italien, portugais, chinois et autre : zone de commentaire limitée en nombre de caractères) ;
5. le ou les autre(s) langue(s) parlée(s) couramment (réponse fermée : français, anglais, espagnol, italien, portugais, chinois et autre : zone de commentaire limitée en nombre de caractères) ;
6. le ou les langues parlées fréquemment (réponse fermée : français, anglais, espagnol, italien, portugais, chinois et autre : zone de commentaire limitée en nombre de caractères) ;
7. le niveau d'étude : « Niveau d'étude actuel (ou le dernier complété si vous avez complété vos études) » (réponse fermée de 1 à 6 : primaire, secondaire, CEGEP collège et autres écoles professionnelles, université 1e cycle, université cycle supérieur, autre : réponse ouvert, zone de commentaire limitée en caractères) ;

5. On notera la particularité du Québec de remplacer *sms* par *texto*.

8. la profession : « Dans quel(s) domaine(s) travaillez-vous ? » (réponse fermée de 1 à 19 : administration des affaires, agriculture, arts, sciences, commerce affaire et économie, communications, droit, éducation (enseignants), éducation (étudiants), gouvernement et service public (élus, fonctionnaires), santé, foresterie, génie, journalier, personnel de bureau, sports, technologies, vente, et autre : réponse ouvert, zone de commentaire limitée en caractères) ;
9. expérience dans la pratique du sms (réponse fermée de 1 à 7 : moins de 6 mois, entre 6 mois et 1 an, entre 1 an et 2 ans, entre 2 ans et 3 ans, entre 3 ans et 4 ans, entre 4 ans et 5 ans, plus de 5 ans) ;
10. le nombre de sms envoyés par semaine (réponse fermée de 1 à 6 : moins de 5, 6 à 10, 11 à 20, 21 à 50, 51 à 100, 101 ou plus) ;
11. le nombre de sms reçus par semaine (réponse fermée de 1 à 6 : moins de 5, 6 à 10, 11 à 20, 21 à 50, 51 à 100, 101 ou plus) ;
12. les types de destinataires des sms : « Utilisez « 1 » pour le plus fréquent, « 2 » pour le suivant, et ainsi de suite. Notez « 0 » pour les éventuelles catégories qui ne vous concernent pas. ». Famille, collègues, amis, concours/jeux, petit(e) ami(e) ou conjoint, autres (réponse fermée avec menu déroulant de 0 à 6) ;
13. la compréhension de l'eSMS : « Comprenez-vous facilement les textos que vous recevez lorsqu'ils contiennent des codes ou des abréviations ?» (réponse fermée de 1 à 3 : oui sans problème, oui mais je dois parfois décoder, non c'est illisible) ;
14. pratique de l'eSMS : « En général, quand vous rédigez vos textos, utilisez-vous le français "normal" ou le "langage texto" ? » (réponse fermée de 1 à 4 : le français normal pour les textos formels (ex. : travail, professeur, etc.) et le langage texto pour les textos informels (ex. amis, famille, proches, etc.), un mélange de français normal et de langage texto, peu importe le destinataire, le français normal uniquement, le langage texto uniquement) ;
15. mélange de langues : « Mélangez-vous parfois une ou plusieurs langue(s) au français dans vos textos ? » (réponse fermée : Oui / Non) ;
16. si mélange de langue, précision de la langue (réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
17. si mélange de langue, fréquence par jour (réponse fermée de 1 à 5 : moins de 5 fois, 5 à 10 fois, 11 à 20 fois, 21 à 50 fois, 50 fois ou plus) ;
18. pratique de l'eSMS en d'autres langues : « Rédigez-vous parfois des textos entièrement dans une autre langue que le français ? » (réponse fermée : Oui / Non).
19. si pratique de l'eSMS en d'autres langues, précision de la langue (réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
20. si pratique de l'eSMS en d'autres langues, fréquence par jour (réponse fermée de 1 à 5 : moins de 5 fois, 5 à 10 fois, 11 à 20 fois, 21 à 50 fois, 50 fois ou plus) ;

21. si pratique de l'eSMS en d'autres langues, utilisation du « langage texto » (réponse fermée : Oui / Non) ;
22. but de la pratique de l'eSMS : pour remplacer le courriel ou le téléphone, pour réduire les coûts d'utilisation de mon téléphone cellulaire, pour envoyer échanger ou demander des informations, pour prendre ou donner rendez-vous, pour garder le contact avec mes proches, pour bavarder ou discuter (réponse fermée : liste déroulante de 1 à 6, 1 étant le plus fréquent) ;
23. contexte de la pratique de l'eSMS : à la maison, au bureau ou à l'école, dans les lieux publics, dans les transports ou lors de déplacements (réponse fermée : liste déroulante de 0 à 4, 1 étant le plus fréquent) ;
24. la possession d'un dictionnaire d'aide à la rédaction dans le téléphone employé (réponse fermée de 1 à 6 : oui et j'utilise toujours les suggestions du dictionnaire, oui et j'utilise souvent les suggestions du dictionnaire, oui et j'utilise parfois les suggestions du dictionnaire, oui et je n'utilise jamais les suggestions du dictionnaire, non, je ne sais pas) ;
25. type de clavier (réponse fermée, liste déroulante de 1 à 5 : un clavier de téléphone régulier (12 touches), un clavier de téléphone qwerty (par exemple sur le BlackBerry), un écran tactile (par exemple sur le iPhone), un stylet (par exemple sur les Palm), autre : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
26. l'utilisation d'autres types de Cémo : « Quels autres modes ou systèmes de communication sur Internet utilisez-vous ? » (réponse fermée de 1 à 4 à cocher : le courrier électronique, les systèmes de clavardage (chat) ou de messagerie instantanée dans Internet (MSN Messenger, Yahoo Messenger, etc.), les réseaux sociaux (Facebook, MySpace, Twitter, etc.), les forums dans Internet) ;
27. temps de lecture : « Combien d'heures par jour consacrez-vous à la lecture pour votre profession/scolarisation et pour vos loisirs ? » (réponse fermée, liste déroulante de 1 à 23 : 0 min/jour, 30 min/jour, 1h/jour, 1h30/jour, 2h/jour jusque 20h/jour) ;
28. types de lecture (réponse fermée : choix entre des articles de journaux, des BD et/ou mangas, des pages de sites Web, des textes de chansons, des livres et autre(s) : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
29. pratique de l'écriture : « Lors de vos activités professionnelles/scolaires et de vos loisirs, consacrez-vous du temps à écrire sur papier, de quelque manière que ce soit ? » (réponse fermée : Oui / Non) ;
30. si pratique de l'écriture, précision du type (réponse fermée : choix entre petits mots, correspondance traditionnelle, journal intime et autre : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
31. si pratique de l'écriture, précision du but (réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;

32. si pratique de l'écriture, précision de la fréquence (réponse fermée, liste déroulante de 1 à 23 : 0 min/jour, 30 min/jour, 1h/jour, 1h30/jour, 2h/jour jusque 20h/jour) ;

Au Québec, la question d'opinion est la suivante (suivie du type de réponse) :

1. Zone de commentaire libre : « Si vous le souhaitez, vous pouvez faire un commentaire sur la manière dont vous écrivez des textos » (réponse ouverte : zone de commentaire limitée en nombre de caractères).

A.2.4 Questionnaire suisse

En Suisse, les questions de fait sont les suivantes (suivies du type de réponse) :

1. le sexe (réponse fermée : M / F) ;
2. l'âge (réponse fermée, avec zone libre demandant une valeur numérique) ;
3. la ou les langue(s) maternelle(s) (réponse fermée : haut allemand, français, italien, romanche, dialecte alémanique et autre : zone de commentaire limitée en nombre de caractères) ;
4. le ou les autre(s) langue(s) parlée(s) à la maison (réponse fermée : haut allemand, français, italien, romanche, dialecte alémanique et autre : zone de commentaire limitée en nombre de caractères) ;
5. le ou les autre(s) langue(s) parlée(s) en dehors de la maison (réponse fermée : haut allemand, français, italien, romanche, dialecte alémanique et autre : zone de commentaire limitée en nombre de caractères) ;
6. le niveau d'étude (réponse fermée de 1 à 5 : éducation primaire (Primarstufe, Primaire or Scuola Primaria), ISCED Level 2 (Sekundarstufe I, Secondaire I or Scuola Media), ISCED Level 3 (Sekundarstufe II or Scuola Media superiore), British GCE A levels (Mittelschulstufe, Gymnase/Lycée or Maturità), Bachelier ou supérieur (Hochschulabschluss, Diplôme de fin d'études or Laurea) ;
7. la profession (réponse fermée de 0 à 4 : aucune, élève/étudiant-e, prestation de services, industrie, artisanat) ;
8. le code postal du domicile (réponse fermée avec zone libre demandant une valeur numérique respectant la structure des codes postaux suisses) ;
9. le code postal de l'école ou du travail (réponse fermée avec zone libre demandant une valeur numérique respectant la structure des codes postaux suisses) ;
10. expérience dans la pratique du sms (réponse fermée de 1 à 7 : moins de 1 an, entre 1 an et 3 ans, entre 3 ans et 5 ans, plus de 5 ans) ;
11. le nombre de sms envoyés par semaine (réponse fermée de 1 à 6 : moins de 5, 5 à 10, 10 à 20, 20 à 50, 50 à 100, plus de 100) ;
12. utilisation du T9 (réponse fermée : Oui / Non) ;
13. l'utilisation d'autres types de Cémo : le courrier électronique, le chat et les forums (réponse fermée de 1 à 5 à cocher : pas de réponse, jamais, rarement, régulièrement, presque tous les jours) ;

14. l'envoi de sms via un service web (réponse fermée de 1 à 5 à cocher : pas de réponse, jamais, rarement, régulièrement, presque tous les jours) ;
15. les types de destinataires des sms (réponse fermée à cocher : famille, collègues, amis, concours/jeux, autre : zone de commentaire limitée en nombre de caractères) ;
16. temps de lecture : nombre d'heures par semaine pour la profession/scolarisation et pour les loisirs (en dehors des sms et du courrier électronique) (réponse fermée avec zone libre demandant une valeur numérique) ;
17. types de lecture (réponse fermée, cases à cocher : choix entre des articles de journaux, des BD et/ou mangas, des pages de sites Web, des livres et autre(s) : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
18. pratique de l'écriture sur papier (réponse fermée : Oui / Non) ;
19. si pratique de l'écriture sur papier, précision du type (réponse fermée : choix entre petits mots, correspondance traditionnelle, journal intime et autre : réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
20. si pratique de l'écriture sur papier, précision de la fréquence (réponse fermée, de 1 à 5 à cocher entre pas de réponse, jamais, rarement, régulièrement, presque tous les jours) ;
21. changement de langues en fonction du destinataire (réponse fermée : Oui / Non) ;
22. si changement de langues, précision des langues utilisées (réponse ouverte avec zone de commentaire limitée en nombre de caractères) ;
23. alternance de langues dans un sms (réponse fermée : Oui / Non) ;
24. si alternance de langues dans un sms, précision des langues (réponse fermée : haut allemand, français, italien, romanche, dialecte alémanique et autre : zone de commentaire limitée en nombre de caractères) ;

En Suisse, la question d'opinion est la suivante (suivie du type de réponse) :

1. Zone de commentaire libre : « Décrivez succinctement la façon dont vous utilisez les sms » (réponse ouverte : zone de commentaire limitée en nombre de caractères).

A.3 Annexes au Chapitre 3

A.3.1 Lexicométrie

FIGURE A.4 – *Répartition du nombre d'unités lexicales en fonction de la longueur de l'unité (en nombre de caractères, dans les corpus de sms bruts et transcrits)*

	brut	transcrit
1-letter words	110.673	108.402
2-letter words	271.396	276.446
3-letter words	172.817	158.308
4-letter words	144.140	174.804
5-letter words	123.726	134.471
6-letter words	77.448	89.429
7-letter words	43.205	55.970
8-letter words	25.635	35.935
9-letter words	11.451	15.989
10-letter words	6.085	8.841
11-letter words	3.027	4.980
12-letter words	2.055	3.143
13-letter words	746	1.028
14-letter words	448	458
15-letter words	260	277
16-letter words	104	244
17-letter words	63	41
18-letter words	48	23
19-letter words	31	11
20-letter words	23	11
21-letter words	22	8
22-letter words	19	5
23-letter words	23	1
24-letter words	33	7
25-letter words	20	1

FIGURE A.5 – *Unités lexicales des 4 corpus de sms francophones dont la fréquence est supérieure à 1.000*

| | | |

FIGURE A.6 – *Mots-clés des corpus de sms par région*

Belgique	suite	La Réunion	suite	Québec	suite	Suisse	suite
tantôt	TEL	ne	ok	to	lol	becs	allé
À	Plein	lé	c	Okay	be	Becs	Chaux
biz	partout	mi	okay	pis	like	Hello	Yverdon
OK	jtd	Té	Okay	I	il	Ok	sympa
bisous	beaucoup	pa	amoin	ne	Allo	Coucou	réjouit
Bisous	à	i	byin	Pis	we	Bec	soucis
kot	Y	té	lycée	the	do	réjouis	train
Bruxelles	néerlandais	fé	kwé	n	gona	joue	apéro
Gros	petite	èk	mèm	P	will	Neuchâtel	samedi
été	Je	Ben	alé	Hey	Oki	redis	I
Louvain	o	ben	mdr	have	Québec	uni	café
plein	et	pou	Atpr	and	of	bec	Coop
kiss	anniversaire	aou	Koman	Montréal	go	natel	Fonds
jusque	bisou	di	réviser	you	was	toute	Payerne
Neuve	avec	A	Mdr	okay	just	un	Zurich
bien	sonne	Ca	tro	at	S	Lausanne	bientôt
amour	Si	Mi	Ok	it	this	Fribourg	salutations
petit	soirée	ke	Kwa	cellulaire	Appelle	Alors	Salutations
coeur	nuit	ou	Ou	is	when	on	prochain
stp	Caro	amwin	pu	Im	one	Redis	environ
SMS	Bisou	il	bises	that	Allô	soir	Bienne
examens	Merci	mwin	ariv	im	pire	Genève	Delémont
aprem	Msg	Ah	gany	if	are	volontiers	midi
demain	réponds	ek	dan	ur	how	pour	Ça
GSM	chou	sa	ali	but	want	ok	les
Kiss	end	ti	BU	Il	coin	retrouve	grippe
Ça	pour	li	sava	Haha	Je	STP	h
h	mon	XD	baigner	correct	btw	ha	Mais
Ju	Laura	in	demin	Il	some	D	Écoute
car	JTM	zot	moin	so	Good	vous	Désolée
Dis	sonner	Bises	Lé	chum	Facebook	Tchô	désolée
Liège	Vais	na	Bye	with	professeur	On	ces
Namur	chez	é	case	genre	But	se	redit
saurai	T	koné	alor	my	did	deux	Bâle
espère	MSN	lol	ki	None	think	boire	Denner
étudier	plus	plaît	apré	can	billets	à	jouerait
manques	bonjour	Lol	kaz	As	Es	À	Moutier
très	Kikou	aprèm	révisions	up	Call	souci	TSR
Suis	Tr	mé	Denis	call	hope	le	Valais
t	Hep	veu	Piouk	know	là	bibli	fondue
Réponds	toi	la	koz	i	fun	Hey	égal
Passe	Sorry	vyin	révise	char	Est	jamais	avertir
saurais	AIME	kan	annif	xx	finalement	verre	noce
de	Cha	koman	dor	métro	tente	décembre	contre
puis	CD	sé	vwar	good	need	cool	motivée
J	appel	bye	téléfone	get	LOL	aprèm	maison
gros	loup	fac	deux	for	Berri	alors	semaine
absence	hein	out	parti	what	chill	une	neige
Voilà	Mons	nou	bac	time	Eille	Ha	jeudi
Enfin	MAIL	kwa	dann	class	scriptum	noël	Cool

FIGURE A.7 – *100 structures polylexicales nominales les plus fréquentes dans le corpus de sms*

gros bisou	AN	mal à la tête	NàDetN	kiss tout plein	NAdvA	salle de bain	NdeN
bonne nuit	AN	dernière fois	AN	minutes de retard	NdeN	bon amusement	AN
bonne soirée	AN	boîte aux lettres	NàN	meilleure amie	AN	arrêt de bus	NdeN
bonne journée	AN	marché de noël	NdeN	fin de journée	NdeN	gros poutou	AN
beaux rêves	AN	happy birthday	AN	vendredi soir	NN	gros câlin	AN
joyeux anniversaire	AN	autre jour	AN	grand merci	AN	mal de tête	NdeN
semaine prochaine	NA	star academy	NN	big kiss	AN	homme de ma vie	NdeAN
bon anniversaire	AN	petit ange	AN	adresse email	NN	pleine forme	AN
appel en absence	NenN	adresse e-mail	NN	bonne semaine	AN	bonnes vacances	AN
bon week-end	AN	bébé d' amour	NdeN	mêmes choses	AN	mi amore	AN
petit coeur	AN	bisous tout plein	NAdvA	bonne idée	AN	énormes bisous	AN
bon appétit	AN	gros kiss	AN	clé usb	NN	samedi prochain	NA
saint nicolas	AN	coup de fil	NdeN	jeune homme	AN	bonne aprem	AN
kiss kiss	NN	are you	VN	changement de programme	NdeN	grève de la stib	NdeDetN
petit loup	AN	petit sms	AN	week-end prochain	NA	doux rêves	AN
autre fois	AN	même temps	AN	rayon de soleil	NdeN	samedi soir	NN
bon courage	AN	petit sms	AN	fête des mères	NdeN	bonne après-midi	AN
prochaine fois	AN	même temps	AN	quart d' heure	NdeN	adresse msn	NN
bonne fête	AN	gros gros bisous	AAN	petite soeur	AN	salut miss	NN
petit message	AN	fin de semaine	NdeN	merci d' avance	NdeN	adresse mail	NN
bonne chance	AN	gros becs	AN	cadeau de noël	NdeN	numéro de gsm	NdeN
bonne merde	AN	nouvel an	AN	grand chose	AN	petit bout	AN
petit chou	AN	appareil photo	NN	bon voyage	AN	numéro de téléphone	NdeN
autre chose	AN	harry potter	NN	coucou petit	NA	petite puce	AN
bisous tout doux	NAdvA	bisous doux	NA	coucou miss	NN	grand plaisir	AN

A.3.2 Néologismes

FIGURE A.8 – *Distribution des âges pour le corpus francophone entier*

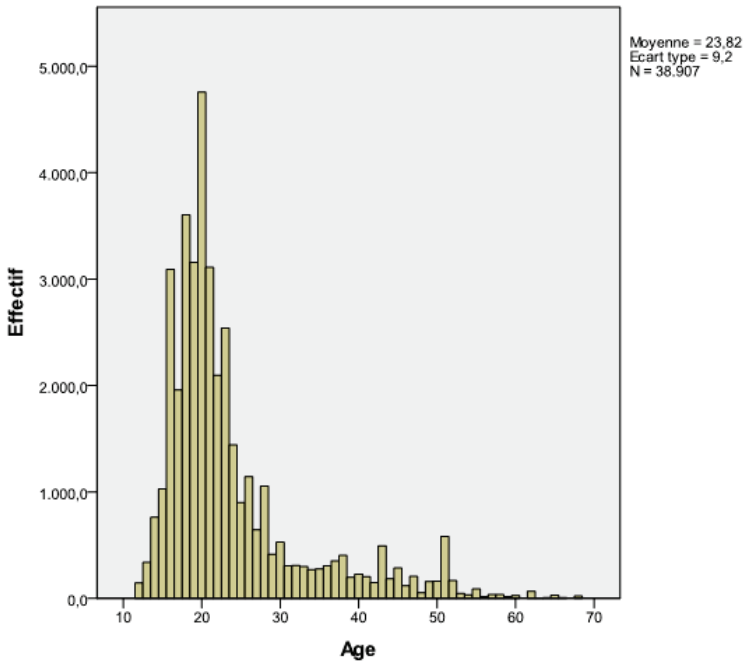


FIGURE A.9 – *Distribution des âges pour le corpus francophone entier, après rééquilibrage par catégories*

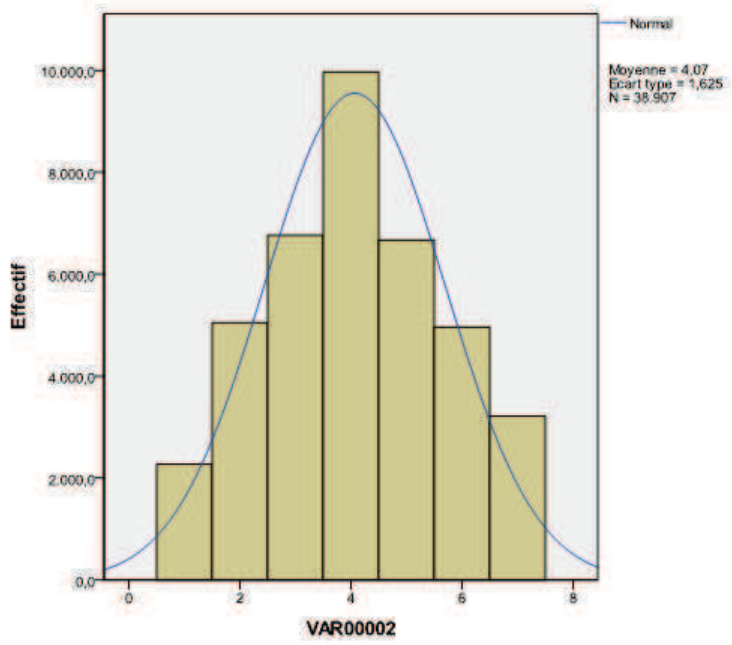


FIGURE A.10 – *Test t comparant les médianes d'utilisation de néologismes en fonction de l'âge*

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
VAR00002	Hypothèse de variances égales	2,771	,096	7,498	38905	,000	,363	,048	,268	,458
	Hypothèse de variances inégales			7,518	1232,830	,000	,363	,048	,268	,458

FIGURE A.11 – *Test t comparant les médianes d'utilisation de néologismes en fonction du nombre de sms envoyés par semaine*

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
sms per week	Hypothèse de variances égales	10,032	,002	-4,596	38910	,000	-,173	,038	-,246	-,099
	Hypothèse de variances inégales			-4,412	1226,306	,000	-,173	,039	-,249	-,096

FIGURE A.12 – *Test du khi-carré mesurant la dépendance des variables néologisme et sexe en francophonie*

Tableau croisé Sexe * Néologisme

			Néologisme		Total
			0	1	
Sexe	F	Effectif	24187	881	25068
		Résidu ajusté	-8,3	8,3	
	M	Effectif	13564	280	13844
		Résidu ajusté	8,3	-8,3	
Total		Effectif	37751	1161	38912

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	68,578[a]	1	,000		
Correction pour la continuité[b]	68,064	1	,000		
Rapport de vraisemblance	72,811	1	,000		
Test exact de Fisher				,000	,000
Nombre d'observations valides	38912				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 413,06.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.13 – *Test du khi-carré mesurant la dépendance des variables néologisme et études en francophonie*

Tableau croisé éducation * Néologisme

			Néologisme		Total
			0	1	
éducation	0	Effectif	182	7	189
		Résidu ajusté	-,6	,6	
	1	Effectif	336	5	341
		Résidu ajusté	1,7	-1,7	
	2	Effectif	3957	100	4057
		Résidu ajusté	2,1	-2,1	
	3	Effectif	11302	513	11815
		Résidu ajusté	-10,4	10,4	
	4	Effectif	21974	536	22510
		Résidu ajusté	8,2	-8,2	
Total		Effectif	37751	1161	38912

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	110,356[a]	4	,000
Rapport de vraisemblance	104,148	4	,000
Nombre d'observations valides	38912		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 5,64.

FIGURE A.14 – *Test du khi-carré mesurant la dépendance des variables néologisme et langue maternelle en francophonie*

Tableau croisé Langue Maternelle * Néologisme

			Néologisme		Total
			0	1	
Langue Maternelle	1	Effectif	33340	942	34282
		Résidu ajusté	7,4	-7,4	
	2	Effectif	4411	219	4630
		Résidu ajusté	-7,4	7,4	
Total		Effectif	37751	1161	38912

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	55,370[a]	1	,000		
Correction pour la continuité[b]	54,687	1	,000		
Rapport de vraisemblance	48,336	1	,000		
Test exact de Fisher				,000	,000
Nombre d'observations valides	38912				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 138,14.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.15 – *Test du khi-carré mesurant la dépendance des variables néologisme et chat en francophonie*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,141[a]	1	,707		
Correction pour la continuité[b]	,117	1	,732		
Rapport de vraisemblance	,142	1	,707		
Test exact de Fisher				,739	,366
Association linéaire par linéaire	,141	1	,707		
Nombre d'observations valides	38912				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 325,67.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.16 – *Test du khi-carré mesurant la dépendance des variables néologisme et forum de discussion en francophonie*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,102[a]	1	,750		
Correction pour la continuité[b]	,072	1	,788		
Rapport de vraisemblance	,102	1	,749		
Test exact de Fisher				,798	,394
Association linéaire par linéaire	,102	1	,750		
Nombre d'observations valides	38912				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 114,18.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.17 – *Test du khi-carré mesurant la dépendance des variables néologisme et courrier électronique en francophonie*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	3,165[a]	1	,075		
Correction pour la continuité[b]	3,059	1	,080		
Rapport de vraisemblance	3,158	1	,076		
Test exact de Fisher				,078	,040
Association linéaire par linéaire	3,165	1	,075		
Nombre d'observations valides	38912				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 535,24.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.18 – *Liste de candidats à la néologie présents dans le corpus sms pour la science (belge), classée par nombre d'occurrences*

Lexie	Type	Nb	Cotexte
aprèm ou aprem	tronc + agglut	378	*je pars au volley toute l'aprèm donc je risque de pas répondre avant 16h30/17h.*
mdr	sigl	219	*Bisou petite poulette signé "ta" petite soeur mdr souvenir souvenir!!*
dvd	emp	106	*viens chez moi on va regarder une K7 j'ai pas de DVD*
appart	tronc	40	*Voilà c'est bon on a un appart de cinq... Faut juste qu'on aille rendre ses papiers le*
bisouille	dériv	12	*Je te bisouille tout plein.*
bisounours	m-v	11	*Bisounours à toi mon babouin des steppes.*
smser	conv	8	*Et tu me SMSeras si c'est Kerry qui a gagné parce que j'aurai pas la TV à Leuven..*
dodoter	dériv	7	*J'arrive pas à dodoter : je me réveille chaque heure*
texto	dériv	5	*Je t'envoie 1 petit texto car on se parlera pas ce soir. Vais aller au dodo*
néonatalité	compo	4	*Alors la néonatalité, ça répond bien à tes attentes?*
déj	tronc	3	*Petit déj chez nous avant départ, café croissants*
overbooké	dériv	3	*ne me demande pas à quelle heure, il est toujours overbooké*
rebisous	dériv	3	*petit SMS pour te faire de gros bisous.Ici tout va bien. Rebisous*
chatter	conv	2	*Non, j'ai pas encore eu l'occasion de chatter avec lui!*
circocabaret	compo	2	*Soirée super! Circocabaret : clown plus bonne musique!*
footeux	dériv	2	*Contente que tu t'amuses mon petit footeux :)*
improshow	compo	2	*les 3 dernières représentations de l'improshow, les 8, 22 et 29*
rebooter	dériv	2	*Si je dois graver et rebooter mon PC*
aquaplouf	compo	1	*Waouw t'as même un GSM aquatique ? (aquaplouf) GÉNIAL !*
autogonflant	compo	1	*Oublie pas le matelas autogonflant de Sof. Kiss*
bisette	dériv	1	*Pas de soucis, je digère l'affaire petit à petit. Passe une belle soirée. Bisette*
bisounuit	compo	1	*Je regarde une émission sur le sida, c'est cool, ça fout le moral. Bisounuit*
bouratcha	dériv	1	*Je suis désolée pour hier, c'est con, j'étais bouratcha*
cafarisé	dériv	1	*je suis à Léon Lepage, c'est bien mais c'est un peu cafarisé*
cambresque	dériv	1	*Lundi cambresque mais week-end dégagé ou dégageable!*
casage	dériv	1	*on se prend des délires tout le temps mais aucune trace de casage*
comatage	dériv	1	*Oui, l'endroit même de nos ébats amoureux, nos comatages*
comater	dériv	1	*Tu comates? Ou je peux venir?*
constructivement	dériv	1	*on se voit samedi constructivement*
cybernamoureutism	m-v	1	*On va lancer une mode de cybernamoureutisme*
débriefing	dériv	1	*Yep,petite fondue au Bauloy...fini votre débriefing?*
décacher	dériv	1	*Peux-tu me dire comment on décache le numéro stp...*

FIGURE A.19 – *Liste de candidats à la néologie présents dans le corpus sms pour la science (belge), classée par nombre d'occurrences (suite)*

Lexie	Type	Nb	Cotexte
dédramatiseur	dériv	1	*Calamity est là,fidèle à son poste de dédramatiseuse..*
dégageable	dériv	1	*Lundi cambresque mais week-end dégagé ou dégageable!*
dégoûtage	dériv	1	*Gros dégoûtage!!! Y a plus Eternal Sunshine au Carollywood*
désaoûlé	dériv	1	*Sinon,t'as bien dormi?T'as désaoûlé?*
dispensatoire	dériv	1	*j'ai test dispensatoire en religion lundi*
donnage	dériv	1	*Et toi ça marche ton donnage de cours?*
dormage	dériv	1	*Bonjour Nuno! Ça été le dormage? Moi oui*
éclatade	dériv	1	*Bonne éclatade chez Fabienne...-*
e-mailiser	dériv	1	*aurais-tu la gentillesse de me e-mailiser ton texte*
embiéré	dériv	1	*des nouvelles concernant les péripéties de ton portable "embiéré"?*
émollience	dériv	1	*Fabrication Artificielle de Malheurs Interminables et de Longue Lymphatique*
ensapinée	dériv	1	*Fais de beaux rêves,miss Ensapinée*
étudiement	dériv	1	*Bonne merde pour ton étudiement*
géreur	dériv	1	*à bientôt Seby.. Oui la semi-géreuse*
gravage	dériv	1	*Anne n'a rien sur son CD! Le gravage a foiré!*
grunge	convers	1	*j'ai compris avec les grunges et Damien*
guili	convers	1	*te faire des guilis dans le bas du dos*
hallucinamment	dériv	1	*Hallucinamment Aimable et Rayonnante Personne?*
indéplaçable	dériv	1	*ton "congé indéplaçable sous peine de divorce"*
kitchissime	dériv	1	*plein de couleurs, qui clignotent, sur la maison, dans le jardin!Kitchissime*
labelloté	antono + dériv	1	*Bisous tout labelloté*
looker	conv	1	*N'oublie pas de looker après les extraits.*
mamme	tronc	1	*à tes reins la mer a perlé rousse à tes mammes vermeilles*
markopsien	dériv	1	*Passe une bonne journée markopsienne*
memerde	dériv	1	*C'est la memerde*
mitterrandienne	dériv	1	*Votre respectueux et dévoué serviteur. Bise mitterrandienne*
moisversaire	m-v	1	*je te souhaite 1 bon moisversaire*
phoner	conv	1	*Bon Loïc,t'as phoné le chanteur?*
reposage	dériv	1	*Bon reposage ce soir et bon cours demain*
ressourçage	dériv	1	*Ton ressourçage se passe bien? Bonne murge ce soir*
réunionite	dériv	1	*Ça s'appelle de la réunionite aiguë !*
solutionnaire	dériv	1	*si tu désires le solutionnaire pour le petit jeu d'association tu peux appeler*
soupercommuter	agglut + compo	1	*Ce soir on soupercommute puis on sort à la Casa*
souplette	dériv	1	*je viendrai- y aura un peu de souplette pour moi?*
speecher	conv	1	*Ça va, t'as bien "speeché" sur Canal C*
squatteur*	dériv	1	*Ta grande amie et squatteuse de cours de l'année dernière*
stressage	dériv	1	*Tu vois le"stressage"de ton frère devient contagieux*

FIGURE A.20 – *Liste de candidats à la néologie présents dans le corpus réunionnais et classée par nombre d'occurrences*

Lexie	Type	Nb	Cotexte
mdr	sigl	331	*Mouais mais bon. L'essence c'est de l'argent mdr*
aprèm ou aprem	tronc + agglut	291	*Coucou my love, cet aprèm pourrais-tu m'amener à mon cours 2 piano vers dix-sept heures trente stp?*
texter	dériv	137	*je te texte dans la journée si il y a possibilité qu'on se voit*
appart	tronc	12	*Le temps est pourri ! Là je rentre à l'appart. Je vais préparer mon sac.*
looker	dériv	3	*Euh non pourquoi ? Je look la télé et toi ?*
bisette	dériv	2	*Bonne soirée à toi. Bisette bisette*
déj	tronc	2	*On est à la plage, y a des petits oiseaux qui attendent nos miettes de petit déj, c'est mignon !*
smser	dériv	1	*Bon, je commence à te smser quand je suis dans la voiture pour aller dans un resto à plateau caillou.*
bichique	m-v	1	*Je t'aime fort ma bichique de France !!! Bisou*
megatop	compo	1	*Ben oui je viens ce midi j'ai espagnol en dernière heure du matin ! T'as pas cours là ou quoi ? Pas MEGATOP en tout cas!*
bisouille	dériv	1	*C'est ok ! Plein de bisouille cocotte*
loveuse	dériv	1	*Stp aidez et donnez-moi une réponse soutenue et digne d'une jeune fille tendre et loveuse.*
texto	dériv	1	*j'arrive dès que j'ai le bus et je t'envoie un texto dès que je suis à la fac. Désolée du retard*
textoter	dériv	1	*Chou je vais dodo moi! Suis crevé! Textote moi quand méme!*

FIGURE A.21 – *Liste de candidats à la néologie présents dans le corpus québécois et classée par nombre d'occurrences*

Lexie	Type	Nb	Cotexte
texter	dériv	26	*Hey, quand tu te réveilles, texte-moi, je m'emmerde tellement.*
courriel	m-v	11	*noms d'utilisateurs et mots de passe ?*
chiller	dériv	7	*chiller dans la journée.*
texto	dériv	4	*J'ai pris un verre de notre bouteille de vin à ton courage !*
retexter	dériv	3	*Je te retexte quand c'est fini.*
clubber	dériv	2	*Quand veux-tu aller clubber en décembre ? Les filles et moi, on*
cubibule	adapt	2	*d'ici ce soir de vive voix ! Bonne après-midi dans ton cubicule*
frencher	dériv	2	*Tu nous diras tu es où, je veux aller te frencher !*
physiothérpaeute	compo	2	*Lol Je ne sais pas avec quels physiothérapeutes on y va.*
watcher	dériv	2	*Oui ! Je l'ai watché. Lol*
bioanal	tronc	1	*Je dirais plutôt en bioANAL.*
booker	dériv	1	*pour les événements de jeudi et samedi, s'il te plaît ? Je dois*
bûchage	dériv	1	*Alló, comment va votre week-end ... de bûchage au boulot ?*
dédomestiqué	dériv	1	*Les familles de ceuzes qui pognent les nerfs face au règne de*
deleter	dériv	1	*Awww ! Peut-être que tu devrais le deleter de ton Facebook ?*
dentistage	dériv	1	*Merci mucho. :) Bon dentistage. :)*
overdoser	dériv	1	*J'ai overdosé sur le spaghetti. =)*
plugger	dériv	1	*Essaye de le plugger dans un poteau d'Hydro-Québec.*

FIGURE A.22 – *Liste de candidats à la néologie présents dans le corpus suisse et classée par nombre d'occurrences*

Lexie	Type	Nb	Cotexte
aprèm ou aprem	tronc + agglut	55	*Comme tu l'a dit ce sera une longue journée. Bonne aprèm, a domani.*
mdr	sigl	21	*J'ai raté mon train pour une minute mdr je te dis ces temps tout va mal.*
appart	tronc	11	*Ouf! On a trouvé notre appart. La rue n'est effectivement pas très engageante, un clochard y dort, et on a pas encore vu la chambre, mais on a déposé nos bagages pas loin.*
bisouille	dériv	4	*Oui, moi, je suis déjà à la fac pour notre exposé sur les sectes donc, je me motive pour les stats, je suis trop sérieuse! Bisouilles*
bisette	dériv	2	*Moi, j'arrive à la t chaud à dix-sept heures.Bonne nuit! Bisette :_)*
bisounours	m-v	1	*Excuse-moi, j'ai été retenue par le boulot. Du coup ni eu le temps de venir ni celui de t'avertir... :_(Bisounours!*
bubulle	dériv	1	*Ha la la, c'était assez comique. Je pense que je vais me taper une grosse bubulle!! Enfin bref. J'ai tout mais au pif, donc il ne me reste plus qu'à prier. Amène!*
craquous	dériv	1	*Coucou les craquous, vos proposition son très sympas mais ne tombent pas du tout au bon moment pour moi*
débriefing	dériv	1	*Me réjouis tellement d'y être! On se fera un débriefing à notre retour! À bientôt, bise*
préloc	tronc	1	*Salut ! Désolée de t'écrire si tard. Mais lauri m'a dit que t'allais peut-être chercher des prélocs demain ! Tu pourrais aussi nous en prendre deux et on te donne les sous samedi ? Becs*
psyoter	dériv	1	*Alors ça a été cette lecture jeudi passé ? Tu psyotes à fond ? Courage, plus qu'un et c'est bon ! Bisous bisous !*

A.3.3 Emprunts

FIGURE A.23 – *Emprunts de langue récents présents dans l'édition 2012 du Petit Robert et/ou le dictionnaire de l'argot, du populaire et du familial en ligne*

	Robert 2012	Dictionnaire *arg. pop. fam.*
aïd	Oui	Non
amigo	Non	Oui
boyfriend	Non	Oui
bug	Oui	Non
checker	Non	Oui
cheese-cake	Oui	Non
churros	Oui	Non
flyer	Oui	Non
kiss	Non	Oui
light	Oui	Non
looké	Oui	Oui
looser	Oui	Non
made in	Oui	Non
niet	Non	Oui
nugget	Oui	Non
over	Non	Oui
pashmina	Oui	Non
practice	Oui	Non
shit	Oui	Non
shoes	Non	Oui
smack	Oui	Non

FIGURE A.24 – *Répertoire des salutations les plus courantes dans les sms belges francophones (Cougnon et François, 2010)*

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

FIGURE A.25 – *Emprunts dans le corpus belge (détail)*

Langue	Position	%	Effectif
Anglais	1	90,9165303	2222
Espagnol	2	2,74140753	67
Italien	3	2,700491	66
Néerlandais	4	1,18657938	29
Allemand	5	1,06382979	26
Arabe	6	0,98199673	24
Autres	7	0,4091653	10

FIGURE A.26 – *Box-plot des emprunts du corpus belge en fonction de l'âge des auteurs de moins de 30 ans*

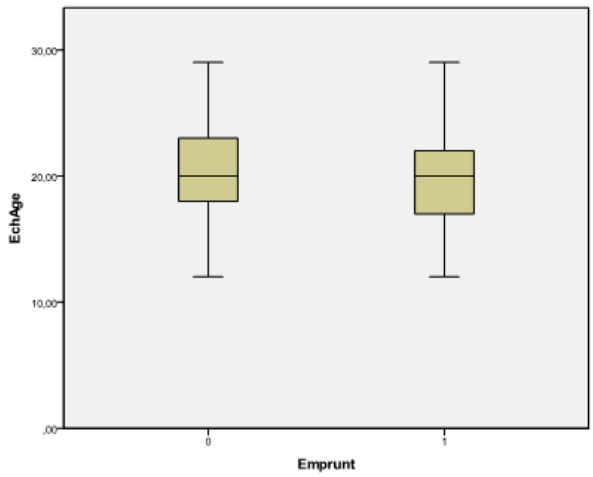


FIGURE A.27 – *Test t comparant les médianes des emprunts et non emprunts en fonction de l'âge dans le corpus belge*

Statistiques de groupe

	Emprunt	N	Moyenne	Ecart-type	Erreur standard moyenne
EchAge	0	19603	20,2275	3,72932	,02664
	1	2189	19,6839	3,51344	,07509

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
EchAge	Hypothèse de variances égales	12,687	,000	6,506	21790	,000	,54364	,08357	,37985	,70744
	Hypothèse de variances inégales			6,823	2768,289	,000	,54364	,07968	,38741	,69988

FIGURE A.28 – *Test du khi-carré mesurant la dépendance des variables emprunt et sexe dans le corpus belge*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	14,279[a]	1	,000		
Correction pour la continuité[b]	14,116	1	,000		
Rapport de vraisemblance	14,416	1	,000		
Test exact de Fisher				,000	,000
Nombre d'observations valides	26493				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 984,28.

b. Calculé uniquement pour un tableau 2x2

Mesures symétriques

		Valeur	Signification approximée
Nominal par Nominal	Phi	-,023	,000
	V de Cramer	,023	,000
Nombre d'observations valides		26493	

FIGURE A.29 – *Test du khi-carré mesurant la dépendance des variables emprunt et niveau d'étude dans le corpus belge*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	47,951[a]	3	,000
Rapport de vraisemblance	46,666	3	,000
Association linéaire par linéaire	47,461	1	,000
Nombre d'observations valides	26493		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 31,26.

Mesures symétriques

		Valeur	Signification approximée
Nominal par Nominal	Phi	,043	,000
	V de Cramer	,043	,000
Nombre d'observations valides		26493	

FIGURE A.30 – *Test t comparant les médianes d'utilisation d'emprunts en fonction du nombre de sms envoyés par semaine dans le corpus belge*

Statistiques de groupe

	Emprunt	N	Moyenne	Ecart-type	Erreur standard moyenne
sms_per_week	0	35279	3,35	1,254	,007
	1	3460	3,35	1,270	,022

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
sms_per_week	Hypothèse de variances égales	5,737	,017	-,154	38737	,877	-,003	,022	-,047	,040
	Hypothèse de variances inégales			-,153	4148,370	,879	-,003	,023	-,048	,041

FIGURE A.31 – *Emprunts dans le corpus réunionnais (détail)*

Langue	Position	%	Effectif
Anglais	1	92,8571429	182
Espagnol	2	3,57142857	7
Allemand	3	2,55102041	5
Arabe	4	0,51020408	1
Néerlandais	5	0,51020408	1

FIGURE A.32 – *Test U de Mann-Whitney comparant les médianes des emprunts et non emprunts en fonction de l'âge dans le corpus réunionnais*

Récapitulatif du test d'hypothèse

	Hypothèse nulle	Test	Sig.	Décision
1	Les médianes de EchAge sont identiques pour les catégories de Emprunt.	Test de la médiane à échantillons indépendants	,022	Rejeter l'hypothèse nulle.
2	La distribution de EchAge est identique sur les catégories de Emprunt.	Test U de Mann-Whitney à échantillons associés	,398	Retenir l'hypothèse nulle.

Les significations asymptotiques sont affichées. Le niveau de signification est .05.

FIGURE A.33 – *Test du khi-carré d'indépendance entre les variables emprunt et sexe dans le corpus réunionnais*

Effectif

		Emprunt		
		0	1	Total
Sexe	F	3263	153	3416
	M	1344	43	1387
Total		4607	196	4803

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	4,790[a]	1	,029		
Correction pour la continuité[b]	4,445	1	,035		
Rapport de vraisemblance	5,048	1	,025		
Test exact de Fisher				,030	,016
Nombre d'observations valides	4803				

FIGURE A.34 – *Test du khi-carré d'indépendance entre les variables emprunt et études dans le corpus réunionnais*

Tableau croisé Emprunt * education

Effectif

		education			
		2	3	4	Total
Emprunt	0	281	1945	2381	4607
	1	17	85	94	196
Total		298	2030	2475	4803

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	2,572[a]	2	,276
Rapport de vraisemblance	2,369	2	,306
Association linéaire par linéaire	2,004	1	,157
Nombre d'observations valides	4803		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5.
L'effectif théorique minimum est de 12,16.

FIGURE A.35 – *Test Mann-Whitney comparant les médianes des variables emprunt et nombre de sms dans le corpus réunionnais*

Récapitulatif du test d'hypothèse

	Hypothèse nulle	Test	Sig.	Décision
1	Les médianes de sms_per_week sont identiques pour les catégories de Emprunt.	Test de la médiane à échantillons indépendants	,000	Rejeter l'hypothèse nulle.
2	La distribution de sms_per_week est identique sur les catégories de Emprunt.	Test U de Mann-Whitney à échantillons associés	,008	Rejeter l'hypothèse nulle.

Les significations asymptotiques sont affichées. Le niveau de signification est .05.

FIGURE A.36 – *Emprunts dans le corpus québécois (détail)*

Langue	Position	%	Effectif
Anglais	1	95,8092486	663
Espagnol	2	3,46820809	24
Italien	3	0,43352601	3
Allemand	4	0,14450867	1
Néerlandais	5	0,14450867	1

FIGURE A.37 – *Test U de Mann-Whitney comparant les médianes des emprunts et non emprunts en fonction de l'âge dans le corpus québécois*

Récapitulatif du test d'hypothèse

	Hypothèse nulle	Test	Sig.	Décision
1	Les médianes de EchAge sont identiques pour les catégories de Emprunt.	Test de la médiane à échantillons indépendants	,001	Rejeter l'hypothèse nulle.
2	La distribution de EchAge est identique sur les catégories de Emprunt.	Test U de Mann-Whitney à échantillons associés	,000	Rejeter l'hypothèse nulle.

Les significations asymptotiques sont affichées. Le niveau de signification est .05.

FIGURE A.38 – *Test du khi-carré d'indépendance entre les variables emprunt et sexe dans le corpus québécois*

Tableau croisé Sexe * Emprunt

Effectif

		Emprunt		
		0	1	Total
Sexe	F	2611	588	3199
	M	599	104	703
Total		3210	692	3902

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	5,083[a]	1	,024		
Correction pour la continuité[b]	4,840	1	,028		
Rapport de vraisemblance	5,273	1	,022		
Test exact de Fisher				,025	,013
Nombre d'observations valides	3902				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 124,67.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.39 – *Test du khi-carré d'indépendance entre les variables emprunt et études dans le corpus québécois*

Effectif

		education				
		2	3	4	o	Total
Emprunt	0	663	433	2082	32	3210
	1	113	105	465	9	692
Total		776	538	2547	41	3902

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	7,501[a]	3	,058
Rapport de vraisemblance	7,718	3	,052
Nombre d'observations valides	3902		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5.
L'effectif théorique minimum est de 7,27.

FIGURE A.40 – *Distribution du nombre de sms par semaine dans le corpus québécois*

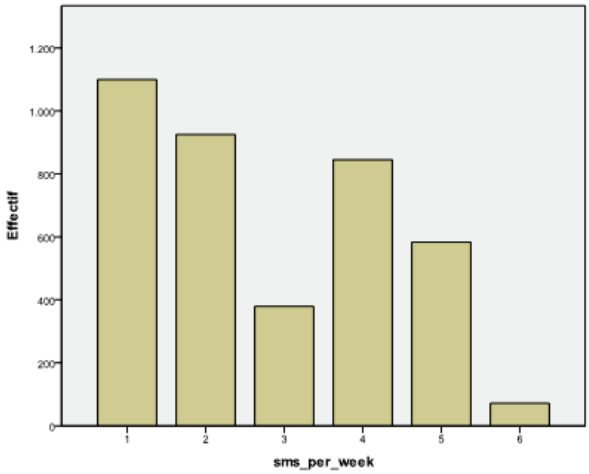


FIGURE A.41 – *Test Mann-Whitney comparant les médianes des variables emprunt et nombre de sms dans le corpus québécois*

Récapitulatif du test d'hypothèse

	Hypothèse nulle	Test	Sig.	Décision
1	Les médianes de sms_per_week sont identiques pour les catégories de Emprunt.	Test de la médiane à échantillons indépendants	,000	Rejeter l'hypothèse nulle.
2	La distribution de sms_per_week est identique sur les catégories de Emprunt.	Test U de Mann-Whitney à échantillons associés	,000	Rejeter l'hypothèse nulle.

Les significations asymptotiques sont affichées. Le niveau de signification est .05.

FIGURE A.42 – *Emprunts dans le corpus suisse (détail)*

Langue	Position	%	Effectif
Anglais	1	76,744186	99
Espagnol	2	10,8527132	14
Italien	3	6,97674419	9
Allemand	4	5,42635659	7

FIGURE A.43 – *Test U de Mann-Whitney comparant les médianes des emprunts et non emprunts en fonction de l'âge dans le corpus suisse*

Récapitulatif du test d'hypothèse

	Hypothèse nulle	Test	Sig.	Décision
1	Les médianes de EchAge sont identiques pour les catégories de Emprunt.	Test de la médiane à échantillons indépendants	,741	Retenir l'hypothèse nulle.
2	La distribution de EchAge est identique sur les catégories de Emprunt.	Test U de Mann-Whitney à échantillons associés	,252	Retenir l'hypothèse nulle.

Les significations asymptotiques sont affichées. Le niveau de signification est .05.

FIGURE A.44 – *Test du khi-carré d'indépendance entre les variables emprunt et sexe dans le corpus suisse*

Tableau croisé Emprunt * Sexe

Effectif

		Sexe		
		0	1	Total
Emprunt	0	992	2420	3412
	1	30	99	129
Total		1022	2519	3541

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	2,049[a]	1	,152		
Correction pour la continuité[b]	1,776	1	,183		
Rapport de vraisemblance	2,136	1	,144		
Test exact de Fisher				,166	,089
Association linéaire par linéaire	2,049	1	,152		
Nombre d'observations valides	3541				

FIGURE A.45 – *Test du khi-carré d'indépendance entre les variables emprunt et études dans le corpus suisse*

Tableau croisé Emprunt * education

Effectif

		education				
		0	2	3	4	Total
Emprunt	0	135	265	942	2070	3412
	1	8	7	50	64	129
Total		143	272	992	2134	3541

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	10,476[a]	3	,015
Rapport de vraisemblance	9,947	3	,019
Association linéaire par linéaire	3,374	1	,066
Nombre d'observations valides	3541		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5.
L'effectif théorique minimum est de 5,21.

FIGURE A.46 – *Test Mann-Whitney comparant les médianes des variables emprunt et nombre de sms dans le corpus suisse*

Récapitulatif du test d'hypothèse

	Hypothèse nulle	Test	Sig.	Décision
1	Les médianes de sms_per_week sont identiques pour les catégories de Emprunt.	Test de la médiane à échantillons indépendants	,044	Retenir l'hypothèse nulle.
2	La distribution de sms_per_week est identique sur les catégories de Emprunt.	Test U de Mann-Whitney à échantillons associés	,200	Retenir l'hypothèse nulle.

Les significations asymptotiques sont affichées. Le niveau de signification est .05.

FIGURE A.47 – *Description de la variable emprunt dans le corpus global de sms*

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	0	35279	91,1	91,1	91,1
	1	3461	8,9	8,9	100,0
	Total	38740	100,0	100,0	

FIGURE A.48 – *Distribution globale des âge pour le corpus sms4science francophone*

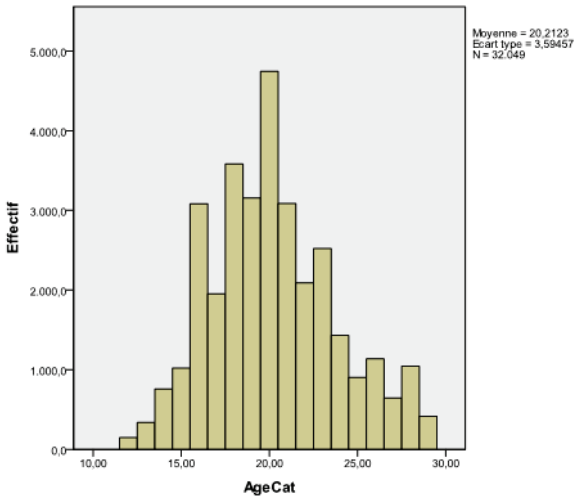


FIGURE A.49 – *Test t sur le corpus global comparant les médianes d'emprunts/non emprunts en fonction de l'âge*

Statistiques de groupe

	Emprunt	N	Moyenne	Ecart-type	Erreur standard moyenne
AgeCat	0	28967	20,2501	3,59740	,02114
	1	3082	19,8572	3,54885	,06393

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
Age Cat	Hypothèse de variances égales	,055	,814	5,771	32047	,000	,39284	,06807	,25942	,52627
	Hypothèse de variances inégales			5,835	3786,688	,000	,39284	,06733	,26084	,52485

FIGURE A.50 – *Test du khi-carré sur le corpus global calculant la dépendance entre les variables emprunt et sexe*

Tableau croisé Emprunt * Sexe

Effectif

		Sexe		
		F	M	Total
Emprunt	0	22567	12712	35279
	1	2386	1074	3460
Total		24953	13786	38739

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	34,260[a]	1	,000		
Correction pour la continuité[b]	34,042	1	,000		
Rapport de vraisemblance	34,917	1	,000		
Test exact de Fisher				,000	,000
Nombre d'observations valides	38739				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 1231,31.

b. Calculé uniquement pour un tableau 2x2

Mesures symétriques

		Valeur	Signification approximée
Nominal par Nominal	Phi	-,030	,000
	V de Cramer	,030	,000
Nombre d'observations valides		38739	

FIGURE A.51 – *Test du khi-carré sur le corpus global calculant la dépendance entre les variables emprunt et niveau d'études*

Tableau croisé Emprunt * education

Effectif

		education							
			0	1	2	3	4	o	Total
Emprunt	0	0	135	297	3588	10701	20526	32	35279
	1	1	8	44	447	1072	1880	9	3461
Total		1	143	341	4035	11773	22406	41	38740

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	58,653[a]	6	,000
Rapport de vraisemblance	49,291	6	,000
Nombre d'observations valides	38740		

a. 3 cellules (21,4%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de ,09.

Mesures symétriques

		Valeur	Signification approximée
Nominal par Nominal	Phi	,039	,000
	V de Cramer	,039	,000
Nombre d'observations valides		38740	

FIGURE A.52 – *Distribution globale des fréquences d'envoi de sms pour le corpus sms4science francophone*

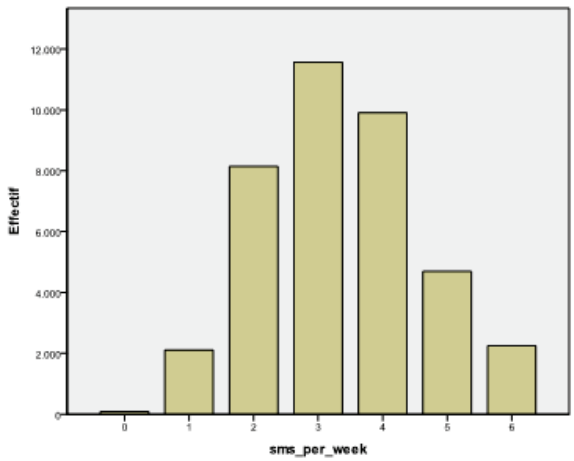


FIGURE A.53 – *Test t pour le calcul de l'égalité des médianes emprunts/non emprunts en fonction de la fréquence d'envoi des sms dans le corpus global*

Statistiques de groupe

	Emprunt	N	Moyenne	Ecart-type	Erreur standard moyenne
sms_per_week	0	35279	3,35	1,260	,007
	1	3460	3,35	1,271	,022

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
sms_per_week	Hypothèse de variances égales	4,981	,026	-,247	38737	,805	-,006	,022	-,050	,038
	Hypothèse de variances inégales			-,245	4153,811	,806	-,006	,023	-,050	,039

FIGURE A.54 – *Test d'égalité des moyennes concernant les médianes des variables langue de l'emprunt et âge dans le corpus global*

Tests d'égalité des moyennes

Emprunt

	Statistique[a]	ddl1	ddl2	Sig.
Welch	,187	3	403,568	,905
Brown-Forsythe	,192	3	735,924	,902

a. Distribution F asymptotique.

FIGURE A.55 – *Test du khi-carré calculant la dépendance entre les variables langue de l'emprunt et sexe dans le corpus global*

Tableau croisé Emprunt * Sexe

Effectif

		Sexe		Total
		F	M	
Emprunt	1	2186	980	3166
	2	81	31	112
	3	60	18	78
	4	20	19	39
	5	17	14	31
	6	17	18	35
Total		2381	1080	3461

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	18,195[a]	5	,003
Rapport de vraisemblance	17,263	5	,004
Nombre d'observations valides	3461		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5.
L'effectif théorique minimum est de 9,67.

FIGURE A.56 – *Test du khi-carré pour l'étude de dépendance entre les variables langue d'emprunt et études supérieures dans le corpus global*

Tableau croisé EducUniv * Emprunt

Effectif

		Emprunt						Total
		1	2	3	4	5	6	
EducUniv	1	1422	51	42	24	16	16	1571
	2	1738	60	36	15	15	19	1883
Total		3160	111	78	39	31	35	3454

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	7,032[a]	5	,218
Rapport de vraisemblance	7,009	5	,220
Association linéaire par linéaire	3,265	1	,071
Nombre d'observations valides	3454		

FIGURE A.57 – *Test du khi-carré pour l'étude de dépendance entre les variables emprunt anglais et études supérieures dans le corpus global*

Tableau croisé EmpruntUK * EducUniv

Effectif

		EducUniv		Total
		1	2	
EmpruntUK	1	1422	1738	3160
	2	149	145	294
Total		1571	1883	3454

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	3,500[a]	1	,061		
Correction pour la continuité[b]	3,275	1	,070		
Rapport de vraisemblance	3,486	1	,062		
Test exact de Fisher				,066	,035
Association linéaire par linéaire	3,499	1	,061		
Nombre d'observations valides	3454				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 133,72.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.58 – *Test du khi-carré pour l'étude de dépendance entre les variables langue d'emprunt et langue maternelle dans le corpus global*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,127[a]	1	,722		
Correction pour la continuité[b]	,057	1	,811		
Rapport de vraisemblance	,125	1	,724		
Test exact de Fisher				,725	,395
Association linéaire par linéaire	,127	1	,722		
Nombre d'observations valides	3461				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 21,48.
b. Calculé uniquement pour un tableau 2x2

FIGURE A.59 – *Test du khi-carré pour l'étude de dépendance entre les variables emprunt ou non emprunt et langue maternelle dans le corpus global*

Tableau croisé Emprunt * Langue_Maternelle

Effectif

		Langue_Maternelle		
		1	2	Total
Emprunt	0	30912	4298	35210
	1	3206	254	3460
Total		34118	4552	38670

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	71,817[a]	1	,000		
Correction pour la continuité[b]	71,349	1	,000		
Rapport de vraisemblance	80,637	1	,000		
Test exact de Fisher				,000	,000
Association linéaire par linéaire	71,815	1	,000		
Nombre d'observations valides	38670				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 407,29.

b. Calculé uniquement pour un tableau 2x2

Mesures symétriques

		Valeur	Signification approximée
Nominal par Nominal	Phi	-,043	,000
	V de Cramer	,043	,000
Nombre d'observations valides		38670	

FIGURE A.60 – *Détail des résidus standardisés ajustés liés au khi-carré pour l'étude de dépendance entre les variables langue d'emprunt et langue maternelle dans le corpus global*

Tableau croisé Langue maternelle * Emprunt

			Emprunt		
			0	1	Total
Langue maternelle	1	Effectif	30912	3206	34118
		Résidu ajusté	-8,5	8,5	
	2	Effectif	4298	254	4552
		Résidu ajusté	8,5	-8,5	
Total		Effectif	35210	3460	38670

A.3.4 Régionalismes

FIGURE A.61 – *Répertoire des belgicismes identifiés dans le corpus de sms belge*

à la porte	brol	encore bien	manneke	salle d'étude
à tantôt	brosser	ennuyant	manneken	sandwich
ad fundum	busé	être fâché sur	mêle-tout	sandwicherie
aimer autant	ça va avec	escabelle	ménage	sonner
aimer de	ça va encore	essuie	minerval	sorteur
américain	caillant	être occupé avec	minitrip	souper
amuser bien	calepin	facultaire	mobile-home	spéculoos
arsouille	candi	faire la file	moins quart	square
asbl	carré	faire sa	mortuaire	student
assez bien	casserole	faire sans	mouru	être calotté
astiquer	cercle	faire drôle	mutuelle	stuut
athénée	chambre du conseil	faire son	oufti	syllabus
au mieux	chatouiller	fancy fair	oui mais non	tantôt
aujourd'hui soir	chicon	farde	ouille ouille	taximan
aula	chiquelette	femme d'ouvrage	patro	tempus
avoir bien fait	clôture	fi	peket	tiesse
avoir bien le temps	clôturé	fieu	péket	togé
avoir bon	clôturer	flamouche	penne	tout partout
avoir difficile	cokotteur	flamoutch	pension	tout qui
avoir dur	commu	gamin	pépette	valves
avoir le temps long	commune	graduat	pète	
avoir x heures de fourche	copions	gsm	pété	
bac à linge	corona	guindaille	péter	
baise	cote	guindailler	piétonnier	
baptême	cotes	heure de fourche	pils	
baptiser	cougnou	histoires	poiret	
baraki	couque	interview	pour du bon	
bazar	crasse	jatte	poyon	
betch	croc	journal de classe	préfète	
bibitif	crollé	ket	prestation	
biesse	croque	kiné	prester	
bleuette	crotte	kot	prendre attention	
bleusaille	défense	kotter	promener	
blindé	déjeuner	kriek	province	
bloc	dia	là tantôt	que du contraire	
blocus	dîner	lavoir	quiquine	
bloque	dire quoi	les 24h	resonner	
bloquer	direct	lift	rhéto	
bob	dis	limaçon	romanes	
bobonne	djoks	logopède	sacoche	
bompa	échoppe	mal de tête	savoir	
bosse	égouttage	manger un bout	sale linge	

FIGURE A.62 – *Test du khi-carré mesurant la dépendance des variables belgicisme et sexe*

Tableau croisé Sexe * Belgicisme

			Belg1		
			0	1	Total
Sexe	F	Effectif	13074	2798	15872
		Résidu ajusté	-2,7	2,7	
	M	Effectif	8960	1755	10715
		Résidu ajusté	2,7	-2,7	
Total		Effectif	22034	4553	26587

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	7,038[a]	1	,008		
Correction pour la continuité[b]	6,950	1	,008		
Rapport de vraisemblance	7,066	1	,008		
Test exact de Fisher				,008	,004
Nombre d'observations valides	26587				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 1834,93.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.63 – *Box-plot des belgicismes en fonction de l'âge des auteurs (de moins de 31 ans)*

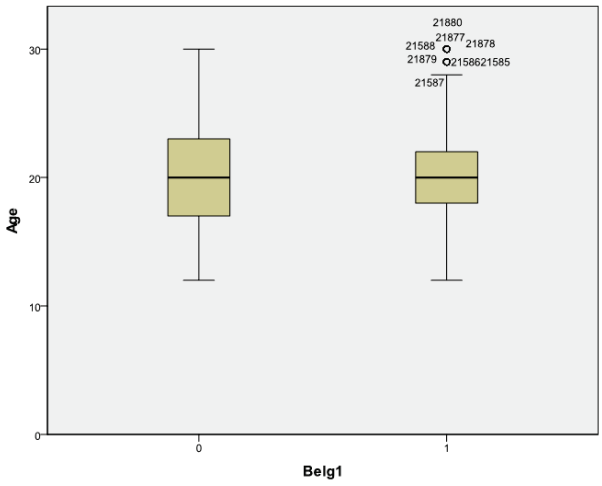


FIGURE A.64 – *Test t mesurant la dépendance des variables belgicisme et âge (moins de 31 ans)*

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
Age	Hypothèse de variances égales	253,964	,000	3,530	22330	,000	,238	,067	,106	,370
	Hypothèse de variances inégales			3,938	7200,726	,000	,238	,060	,119	,356

FIGURE A.65 – *Test du khi-carré mesurant la dépendance des variables belgicisme et utilisation du chat*

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,112[a]	1	,738		
Correction pour la continuité[b]	,099	1	,753		
Rapport de vraisemblance	,112	1	,738		
Test exact de Fisher				,749	,376
Association linéaire par linéaire	,112	1	,738		
Nombre d'observations valides	26587				

FIGURE A.66 – *Test du khi-carré mesurant la dépendance des variables belgicisme et utilisation du forum de discussion*

Tableau croisé

			Belg1		Total
			0	1	
forum_use	0	Effectif	20764	4341	25105
		Résidu ajusté	-3,0	3,0	
	1	Effectif	1270	212	1482
		Résidu ajusté	3,0	-3,0	
Total		Effectif	22034	4553	26587

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	8,794[a]	1	,003		
Correction pour la continuité[b]	8,585	1	,003		
Rapport de vraisemblance	9,189	1	,002		
Test exact de Fisher				,003	,002
Association linéaire par linéaire	8,793	1	,003		
Nombre d'observations valides	26587				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 253,79.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.67 – *Liste des 25 belgicismes les plus fréquents du corpus de sms*

Forme, cat.	Occurrences
savoir, V	1209
à tantôt/tantôt, Adv	958
dis, V	491
gsm, N	310
sonner quelqu'un, V	267
kot, N	236
souper, V	214
dire quoi, V	143
blocus, N	34
cokotteur, N	27
dîner, V et N	24
syllabus, N	24
bosse, N	17
cercle, N	17
direct, Adv	17
sandwich, N	16
farde, N	15
mal de tête, N	15
commu, N	13
pété, A	13
rhéto, N	13
tout partout, Adv	13
fieu, N	11
manger un bout, V	11
ça va avec, Expr	10

FIGURE A.68 – *Test du khi-carré mesurant la dépendance des variables utilisation de savoir et type d'occupation*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	2,656[a]	3	,448
Rapport de vraisemblance	2,579	3	,461
Association linéaire par linéaire	,005	1	,943
Nombre d'observations valides	4569		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 6,31.

FIGURE A.69 – *Test du khi-carré mesurant la dépendance des variables utilisation de savoir et pratique du courrier électronique*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	1,335[a]	1	,248		
Correction pour la continuité[b]	1,244	1	,265		
Rapport de vraisemblance	1,325	1	,250		
Test exact de Fisher				,258	,133
Association linéaire par linéaire	1,335	1	,248		
Nombre d'observations valides	4569				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 266,69.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.70 – *Répertoire des régionalismes présents dans le corpus de sms réunionnais*

battre	faire pas rien	lontan	sauvage
battre (un) carré	fénoir	loyer	savate
bec	foutant	ma fille	sonner quelqu'un
boucaner	gagner	malabar	sur le tard
bouchon	gaillard	malgachine	tamarin
boug	goûter	manzer	tantine
brèdes	grains	margouillat	thon-banane
cafre	grand matin	marmaille	tirer
cari	gratelle	massalé	totocher
case	gros marraine	ménage	tranche papaye
chauffé	gros pois	mol	transport
chouchou	gros soleil	moukate	vanguer
coco	hiver	moune	verres
combava	jusqu'à l'heure	mounoir	vert
dalon	kaf	nuite	vilain
dans le fond	kalou	ouvrir le linge	zamal
défouler	koméla	paré	
désordre	lancer	plaine	
doucement doucement	larourout	poupette	
doudou	les bas	radier	
faille	linge	rougail	

FIGURE A.71 – *Liste de 25 régionalismes les plus fréquents du corpus de sms réunionnais*

Forme, cat.	Occurrences
case, N	52
doudou, N	38
gagner, V	27
nuite, N	19
mol, A	18
tantine, N	13
bec, N	12
cafre, N	12
lontan, Adv	12
ma fille, SN	12
marmaille, N	12
plaine, N	10
tirer, V	9
dalon, N	7
foutant, N	7
moune, N	7
boug, N	6
cari, N	5
malabar, N	5
dans le fond, Adv	4
linge, N	4
transport, N	4
lancer, V	3
ouvrir le linge, V	3
sonner quelqu'un, V	3

FIGURE A.72 – *Test du khi-carré mesurant la dépendance des variables régionalisme et sexe de l'auteur à La Réunion*

Tableau croisé Sexe * Régionalisme

			Régionalisme		
			0	1	Total
Sexe	F	Effectif	3303	116	3419
		Résidu ajusté	-3,1	3,1	
	M	Effectif	1364	24	1388
		Résidu ajusté	3,1	-3,1	
Total		Effectif	4667	140	4807

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	9,664[a]	1	,002		
Correction pour la continuité[b]	9,084	1	,003		
Rapport de vraisemblance	10,675	1	,001		
Test exact de Fisher				,002	,001
Nombre d'observations valides	4807				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 40,42.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.73 – *Test U de Mann-Whitney comparant les médianes des régionalismes et non régionalismes en fonction de l'âge à La Réunion*

Récapitulatif du test d'hypothèse

	Hypothèse nulle	Test	Sig.	Décision
1	Les médianes de Age sont identiques pour les catégories de Reunionisme.	Test de la médiane à échantillons indépendants	,005	Rejeter l'hypothèse nulle.
2	La distribution de Age est identique sur les catégories de Reunionisme.	Test U de Mann-Whitney à échantillons associés	,078	Retenir l'hypothèse nulle.

Les significations asymptotiques sont affichées. Le niveau de signification est .05.

FIGURE A.74 – *Test du khi-carré mesurant la dépendance des variables régionalisme et utilisation du chat à La Réunion*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	2,363[a]	1	,124		
Correction pour la continuité[b]	2,105	1	,147		
Rapport de vraisemblance	2,350	1	,125		
Test exact de Fisher				,143	,074
Association linéaire par linéaire	2,363	1	,124		
Nombre d'observations valides	4807				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 63,08.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.75 – *Test du khi-carré mesurant la dépendance des variables régionalisme et utilisation du forum à La Réunion*

Tableau croisé forum_use * Régionalisme

			Régionalisme		
			0	1	Total
forum_use	0	Effectif	3490	120	3610
		Résidu ajusté	-2,9	2,9	
	1	Effectif	1177	20	1197
		Résidu ajusté	2,9	-2,9	
Total		Effectif	4667	140	4807

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	8,689[a]	1	,003		
Correction pour la continuité[b]	8,115	1	,004		
Rapport de vraisemblance	9,758	1	,002		
Test exact de Fisher				,003	,001
Association linéaire par linéaire	8,688	1	,003		
Nombre d'observations valides	4807				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 34,86.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.76 – *Répertoire des québécismes identifiés dans le corpus de sms québécois*

à matin, Adv.	3
à soir, Adv.	32
abusif, A	1
adonner, V.	3
autographier, V.	1
avant-midi, N.	3
avoir de l'allure, Exp.	2
avoir la chienne	1
en baboune, Exp.	1
bagel, N.	1
bande, N,	1
beigne, N.	2
bien le bonjour, Exp.	1
blonde, N.	8
cave, A.	1
char, N.	17
dépanneur, N.	3
dîner, N. et V.	3
dispendieux, A.	1
épluchette, N.	1
être crampé de rire, Exp.	4
fin, A.	4
funny, A.	1
good luck, Exp.	3
habit, N.	1
meeting, N.	4
nono, A.	1
pis, Adv.	18
pouding, N	1
safe, A.	2
sloche, N.	1
smatte, A.	1
sofa, N.	2
souper, N. et V.	70

FIGURE A.77 – *Test du khi-carré mesurant la dépendance des variables régionalisme et sexe de l'auteur au Québec*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	1,610[a]	1	,205		
Correction pour la continuité	1,323	1	,250		
Rapport de vraisemblance	1,721	1	,190		
Test exact de Fisher				,235	,123
Nombre d'observations valides	3907				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 22,34.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.78 – *Test du khi-carré mesurant la dépendance des variables régionalisme et pratique du chat au Québec*

Tests du Khi-deux

	Valeur	Ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	1,360[a]	1	,244		
Correction pour la continuité[b]	1,115	1	,291		
Rapport de vraisemblance	1,429	1	,232		
Test exact de Fisher				,272	,145
Association linéaire par linéaire	1,360	1	,244		
Nombre d'observations valides	3907				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 27,29.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.79 – *Test du khi-carré mesurant la dépendance des variables régionalisme et pratique des forums de discussion au Québec*

Tests du Khi-deux

	Valeur	Ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	1,111[a]	1	,292		
Correction pour la continuité[b]	,889	1	,346		
Rapport de vraisemblance	1,069	1	,301		
Test exact de Fisher				,314	,172
Association linéaire par linéaire	1,111	1	,292		
Nombre d'observations valides	3907				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 26,28.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.80 – *Test du khi-carré mesurant la dépendance des variables régionalisme et sexe de l'auteur en Suisse*

Tableau croisé Sexe * Helvétisme

			Helvétisme		Total
			0	1	
Sexe	F	Effectif	2187	387	2574
		Résidu ajusté	-2,8	2,8	
	M	Effectif	918	119	1037
		Résidu ajusté	2,8	-2,8	
Total		Effectif	3105	506	3611

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	7,773[a]	1	,005		
Correction pour la continuité[b]	7,480	1	,006		
Rapport de vraisemblance	8,039	1	,005		
Test exact de Fisher				,005	,003
Nombre d'observations valides	3611				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 145,31.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.81 – *Test du khi-carré mesurant la dépendance des variables régionalisme et pratique du chat en Suisse*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	3,171[a]	1	,075		
Correction pour la continuité[b]	2,996	1	,083		
Rapport de vraisemblance	3,139	1	,076		
Test exact de Fisher				,082	,042
Association linéaire par linéaire	3,170	1	,075		
Nombre d'observations valides	3611				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 187,07.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.82 – *Test du khi-carré mesurant la dépendance des variables régionalisme et pratique des forums de discussion en Suisse*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,284[a]	1	,594		
Correction pour la continuité[b]	,201	1	,654		
Rapport de vraisemblance	,279	1	,597		
Test exact de Fisher				,613	,322
Association linéaire par linéaire	,284	1	,594		
Nombre d'observations valides	3611				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 44,84.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.83 – *Test du khi-carré mesurant la dépendance des variables utilisation de redire et occupation professionnelle en Suisse*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,086[a]	1	,770		
Correction pour la continuité[b]	,018	1	,895		
Rapport de vraisemblance	,085	1	,771		
Test exact de Fisher				,750	,438
Association linéaire par linéaire	,086	1	,770		
Nombre d'observations valides	506				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 14,09.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.84 – *Test khi-carré mesurant la dépendance des variables redire et éducation en Suisse*

Tableau croisé éducation * Helvétisme

			Helvétisme		
			1	2	Total
éducation	2	Effectif	42	14	56
		Résidu ajusté	-2,4	2,4	
	3	Effectif	125	22	147
		Résidu ajusté	-,3	,3	
	4	Effectif	267	36	303
		Résidu ajusté	1,8	-1,8	
Total		Effectif	434	72	506

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	6,757[a]	2	,034
Rapport de vraisemblance	6,010	2	,050
Association linéaire par linéaire	5,927	1	,015
Nombre d'observations valides	506		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 7,97.

FIGURE A.85 – *Comparaison des variantes régionales du corpus de sms avec leurs variantes en français standard (et différence du nombre de caractères entre variantes)*

Français Régional	Français Standard	Diff
savoir, V	pouvoir	-1
à tantôt/tantôt, Adv	tout à l'heure	-8
dis, V	/	+3
gsm, N	téléphone portable/mobile	- ?
sonner quelqu'un, V	appeler quelqu'un	-1
kot, N	chambre d'étudiant ?	- ?
souper, V	dîner	+1
dire quoi, V	dire ce qu'il en est	-11
blocus, N	période d'étude ?	- ?
cokotteur, N	collocataire	-3
dîner, V et N	déjeuner	-3
syllabus, N	manuel	+2
case, N	maison	-2
doudou, N	chéri/chérie	-1/0
gagner, V	Pouvoir/posséder/éprouver	- ?
à soir, Adv.	ce soir	-1
hey, int.	salut	-2
ça joue, exp.	ça va	+2
redire, V.	confirmer	-3
bec, N.	bisou	-2
natel, N.	téléphone portable/mobile	- ?

A.4 Annexes au Chapitre 4

A.4.1 Morphosyntaxe

FIGURE A.86 – *Typologie des variantes orthographiques 1 (inspirée par Catach : 1980)*

I. *Erreurs à dominante phonétique* (6)		
1101[1] **Confusion t-d** (3) *descendre/descentre* (4)[2] *Tu sais vider la bassine qui est dans ma chambre,la descentre et mettre l'essuie à lavé ?*	**1102 Confusion f-v** (1) *ferai/verai* (5) *haa mechan lou epargne mwa je te verai ce tu voudra*	**1103 Autres confusions** (2) *fatigué/fatiqué* (6) *J ai eu une nuit hyper courte et je sui fatiqué*

II. *Erreurs à dominante phonogrammique altérant la valeur phonique* (77)				
	Voyelles/digrammes voc.	**Semi-voyelles**	**Consonnes/digrammes cons.**	**Consonnes simples/doubles**
Omission ou adjonction	**2000** (4) *crevée/creuvée* (7) *y a peu d chances d m voir tonight car j suis creuvée*	**2001** (6) *filé/fillé* (8) *Salut c {Nom} ben vla {Nom} ma fillé ton num et cest pr voir si cest b1twa!*	**2002** (1) *enregistré/enregirstré* (9) *Oui un trop marrant que g enregirstré*	**2003** (27) *appeler/appeller* (10) *Tu peux m'appeller ou m'envoyer un sms*
Confusion	**2010** (16) *envoies/envoyes* (11) *tu m'envoyes chier*	**2011** (4) *ennuyer/ennuier* (12) *Désolée de t'ennuier avec ça!*	**2012** (17) *excuse/escuse* (13) *Bb je m escuse de te tel si longtenp*	**2013** (2) *dilemme/dilemne* (14) Ou se situer face à ce dilemne ?

III. *Erreurs à dominante phonogrammique n'altérant pas la valeur phonique* (225)				
	Voyelles/digrammes voc.	**Semi-voyelles**	**Consonnes/digrammes cons.**	**Consonnes simples/doubles**
Omission ou adjonction	**2100** (1) *garantis/guarantis* (15) *j'guarantis pas mon ambiance!*		**2102** (8) *absence/abscence* (16) *Tout était bien sauf ton abscence*	**2103** (4) *ensemble/enssemble* (17) *jme disai kn pourait se mettre tt enssemble*
				21030 (67) *acompte/accompte* (18) *Serait-il possible de ravoir un accompte lundi?*
Confusion	**2110** (85) *commence/commance* (19) *ça commance dans 4 min*	**2111** (1) *pièce/pyece* (20) *y avai 1 pyece kole*	**2112** (59)[3] *concentrer/consentrer* (21) *tu veu bien me laisser me consentrer*	

FIGURE A.87 – *Typologie des variantes orthographiques 3 (inspirée par Catach : 1980)*

V. *Erreurs concernant les homophones* (931)	
40 Homophones de discours (6) *tes/t'es* (40) *T'es messages ne sont pas tjs les plus clairs au monde* **41 Homophones lexicaux** (38) *court/cours* (41) *Elle l'est, mais un cours instant!*	**Homophones grammaticaux** (912) **422** *ce/se* (45) (42) *je te phone ce soir et d2m1 pour qu on ce voit* **423** *ces/ses* (6) (43) *T as d ses idée choux jte jure...* **425** *on/ont* (8) (44) *ont est tj au resto* **428** autres (828) *sa/ça* (45) *Bien dormi? Moi comme un prince sans ça princesse*
VI. *Erreurs concernant les idéogrammes* (3)	
501 Ajout d'apostrophe (2) *dorénavant/d'orénavant* (46) *D'orénavant tu sais quoi*	**502 Ajout de trait d'union** (1) télécommande/télé-commande (47) *J'ai une nouvelle télé-commande*

VII. *Erreurs concernant les lettres non fonctionnelles* (12)		
		614 Ajout de finales particulières (12) *cauchemar/cauchemard* (48) *je nage de nouveau en plein cauchemard*

FIGURE A.88 – *Les verbes les plus concernés par les fautes d'accord sujet-verbe du corpus de sms belge*

Verbe	Occurrences fautives
faire	115
dire	106
être	46
venir	46
aller	40
vouloir	31
pouvoir	28
devoir	25
avoir	20
partir	19
sortir	17
dormir	15
inquiéter	13
passer	12
savoir	10

FIGURE A.89 – *Les groupes verbaux les plus concernés par la confusion des modes verbaux dans le corpus de sms belge*

Groupe	Pourcentage
Groupe 3	78,77%
Groupe 1	11,33%
Auxiliaires	9,47%
Groupe 2	0,43%

FIGURE A.90 – *Les verbes les plus concernés par la confusion de l'accord 1e/2e p. de l'indicatif dans le corpus de sms belge*

aller
arriver
avouer
donner
envoyer
faire
manger
manquer
négocier
passer
penser
pouvoir
préférer
prendre
réveiller
s'ennuyer
sonner
tenter
travailler

FIGURE A.91 – *Histogramme des effectifs pour la confusion ou non confusion des temps futur et confitionnel en fonction des catégories d'âge*

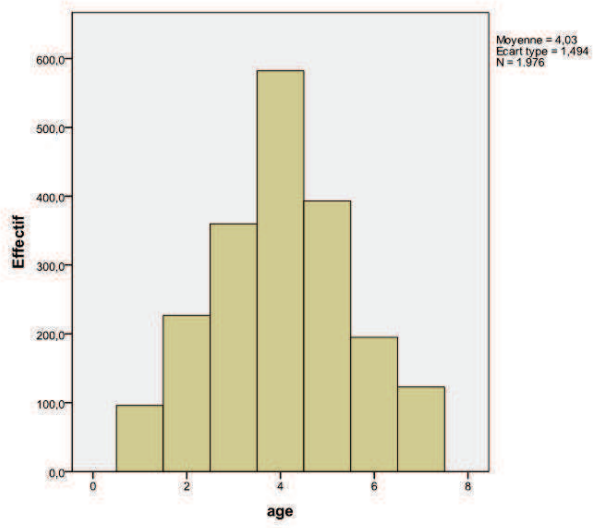


FIGURE A.92 – *Test t comparant les médianes de confusion des temps futur et conditionnel en fonction de l'âge*

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
age	Hypothèse de variances égales	5,572	,018	-3,151	1974	,002	-,473	,150	-,768	-,179
	Hypothèse de variances inégales			-2,917	112,846	,004	-,473	,162	-,795	-,152

FIGURE A.93 – *Test khi-carré mesurant la dépendance des variables confusion des temps indicatif futur et conditionnel et sexe dans les sms*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	2,050[a]	1	,152		
Correction pour la continuité[b]	1,766	1	,184		
Rapport de vraisemblance	2,019	1	,155		
Test exact de Fisher				,180	,093
Nombre d'observations valides	1976				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 41,05.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.94 – *Test khi-carré mesurant la dépendance des variables confusion des temps indicatif futur et conditionnel et niveau d'études dans les sms*

Tableau croisé Confusion Futur Conditionnel * éducation

			éducation			Total
			2	3	4	
Conf. Fut_Cond	non	Effectif	168	535	1169	1872
		Résidu ajusté	-3,2	-3,8	5,5	
	oui	Effectif	19	48	37	104
		Résidu ajusté	3,2	3,8	-5,5	
Total		Effectif	187	583	1206	1976

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	30,964[a]	2	,000
Rapport de vraisemblance	29,652	2	,000
Nombre d'observations valides	1976		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 9,84.

FIGURE A.95 – *Test khi-carré mesurant la dépendance des variables confusion des temps indicatif futur et conditionnel et pratique du chat dans les sms*

Tableau croisé Faute Fut_Cond * chat_use

			chat_use		Total
			0	1	
Faute Fut_Cond	non	Effectif	354	1518	1872
		Résidu ajusté	-2,7	2,7	
	oui	Effectif	31	73	104
		Résidu ajusté	2,7	-2,7	
Total		Effectif	385	1591	1976

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	7,458[a]	1	,006		
Correction pour la continuité[b]	6,780	1	,009		
Rapport de vraisemblance	6,723	1	,010		
Test exact de Fisher				,010	,006
Nombre d'observations valides	1976				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 20,26.
b. Calculé uniquement pour un tableau 2x2

FIGURE A.96 – *Test khi-carré mesurant la dépendance des variables confusion des temps indicatif futur et conditionnel et fréquentation de forums dans les sms*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,007[a]	1	,935		
Correction pour la continuité[b]	,000	1	1,000		
Rapport de vraisemblance	,007	1	,935		
Test exact de Fisher				1,000	,597
Nombre d'observations valides	1976				

FIGURE A.97 – *Test khi-carré mesurant la dépendance des variables confusion des temps indicatif futur et conditionnel et pratique du courrier électronique dans les sms*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,094[a]	1	,759		
Correction pour la continuité[b]	,036	1	,849		
Rapport de vraisemblance	,093	1	,760		
Test exact de Fisher				,728	,418
Nombre d'observations valides	1976				

FIGURE A.98 – *Histogramme des effectifs pour la confusion ou non confusion des modes participe passé et infinitif en fonction des catégories d'âge*

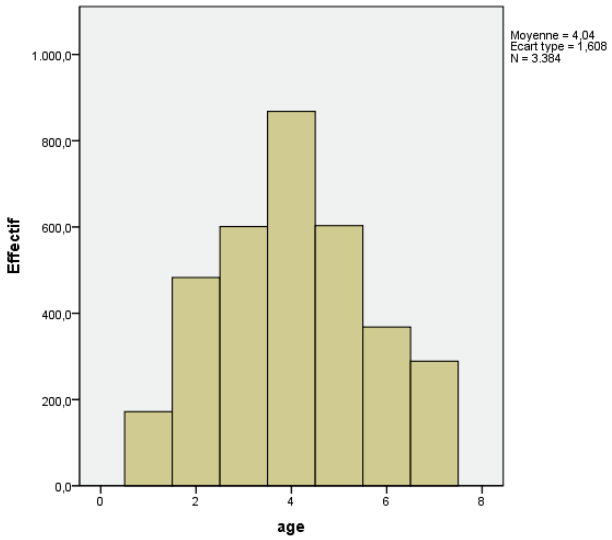


FIGURE A.99 – *Test t comparant les médianes de confusion des modes participe passé et infinitif en fonction de l'âge*

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
âge	Hypothèse de variances égales	4,517	,034	-4,743	3382	,000	-,561	,118	-,793	-,329
	Hypothèse de variances inégales			-4,503	215,639	,000	-,561	,125	-,807	-,315

FIGURE A.100 – *Test khi-carré mesurant la dépendance des variables confusion des modes participe passé et infinitif et sexe dans les sms*

Tableau croisé Faute Inf-Part * sexe

			sexe		
			F	M	Total
Faute Inf-Part	non	Effectif	1900	1289	3189
		Résidu ajusté	3,0	-3,0	
	oui	Effectif	95	100	195
		Résidu ajusté	-3,0	3,0	
Total		Effectif	1995	1389	3384

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	8,959[a]	1	,003		
Correction pour la continuité[b]	8,516	1	,004		
Rapport de vraisemblance	8,812	1	,003		
Test exact de Fisher				,003	,002
Nombre d'observations valides	3384				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 80,04.
b. Calculé uniquement pour un tableau 2x2

FIGURE A.101 – *Test khi-carré mesurant la dépendance des variables confusion des modes participe passé et infinitif et niveau d'études dans les sms*

Tableau croisé Faute Inf-Part * education

			education			Total
			2	3	4	
Faute Inf-Part	non	Effectif	291	1009	1889	3189
		Résidu ajusté	-1,2	-6,1	6,5	
	oui	Effectif	23	103	69	195
		Résidu ajusté	1,2	6,1	-6,5	
Total		Effectif	314	1112	1958	3384

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	44,565[a]	2	,000
Rapport de vraisemblance	43,525	2	,000
Nombre d'observations valides	3384		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 18,09.

FIGURE A.102 – *Test khi-carré mesurant la dépendance des variables confusion des modes participe passé et infinitif et pratique du chat dans les sms*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,001[a]	1	,972		
Correction pour la continuité[b]	,000	1	1,000		
Rapport de vraisemblance	,001	1	,972		
Test exact de Fisher				1,000	,531
Nombre d'observations valides	3384				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 36,19.
b. Calculé uniquement pour un tableau 2x2

FIGURE A.103 – *Test khi-carré mesurant la dépendance des variables confusion des modes participe passé et infinitif et utilisation du courrier électronique dans les sms*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,303[a]	1	,582		
Correction pour la continuité[b]	,215	1	,643		
Rapport de vraisemblance	,307	1	,579		
Test exact de Fisher				,667	,325
Nombre d'observations valides	3384				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 47,19.
b. Calculé uniquement pour un tableau 2x2

FIGURE A.104 – *Test khi-carré mesurant la dépendance des variables confusion des modes participe passé et infinitif et fréquentation des forums de discussion*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	1,220[a]	1	,269		
Correction pour la continuité[b]	,875	1	,350		
Rapport de vraisemblance	1,116	1	,291		
Test exact de Fisher				,306	,173
Nombre d'observations valides	3384				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 9,74.
b. Calculé uniquement pour un tableau 2x2

A.4.2 Interrogatives

FIGURE A.105 – *Graphe représentant les structures interrogatives standard du corpus de sms basées sur "est-ce que"*

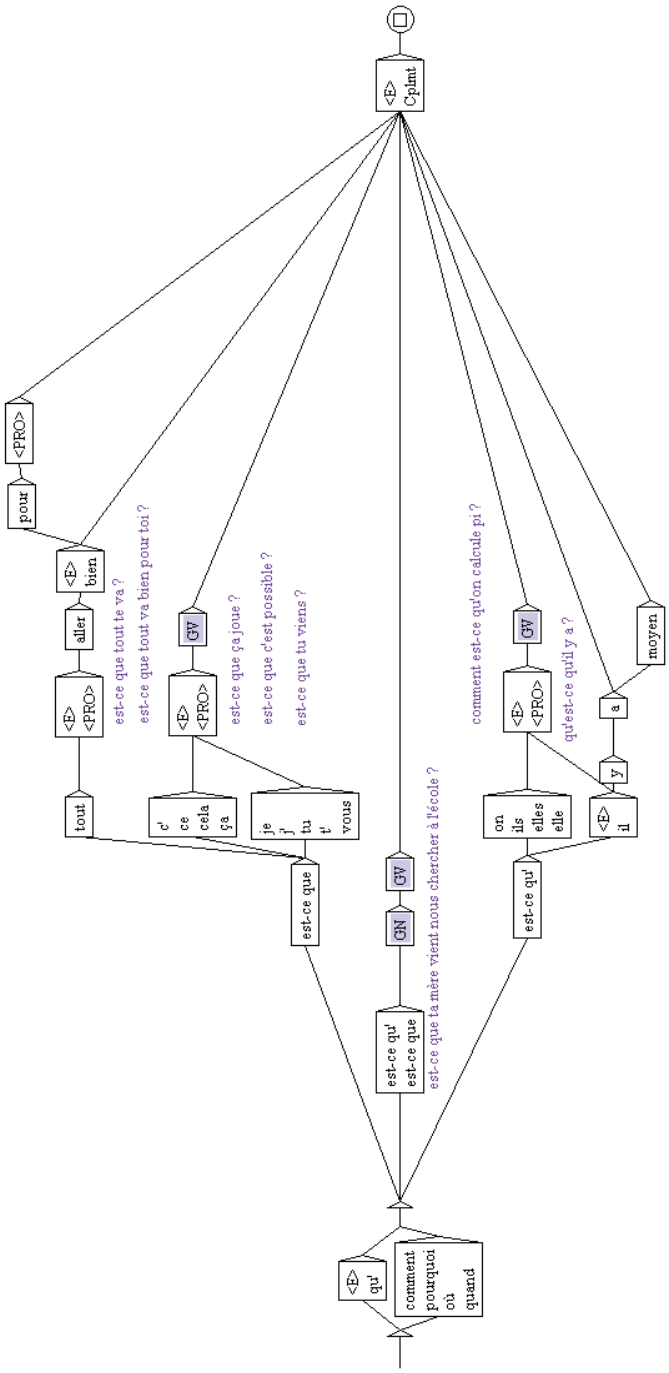


FIGURE A.106 – *Graphe représentant les structures interrogatives standard du corpus de sms basées sur un élement qu- et une inversion*

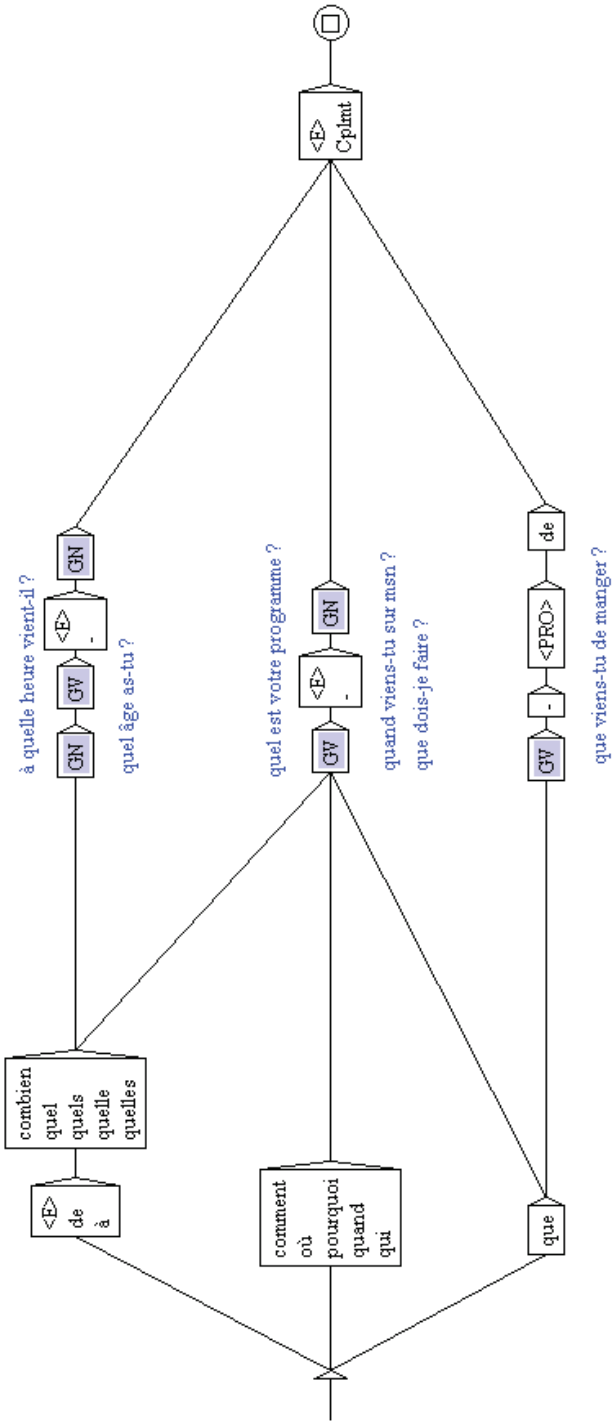


FIGURE A.107 – *Liste des verbes apparaissant dans les interrogatives standard basées sur une inversion sujet-verbe*

accepter	passer
aimer	penser
aller	pouvoir
arrêter	pratiquer
avoir	préférer
cacher	prendre
compter	recevoir
connaître	regarder
convenir	remettre
croire	remplacer
devoir	repérer
dire	répondre
dormir	savoir
écouter	se coucher
employer	se sentir
entendre	se souvenir
être	travailler
faire	venir
falloir	vivre
jouer	voir
mettre	vouloir

FIGURE A.108 – *Test t comparant les médianes d'utilisation de types d'interrogatives en fonction de l'âge*

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
AgeGroupe	Hypothèse de variances égales	34,253	,000	-11,365	11657	,000	-,487	,043	-,572	-,403
	Hypothèse de variances inégales			-10,852	2029,426	,000	-,487	,045	-,576	-,399

FIGURE A.109 – *Test khi-carré mesurant la dépendance des variables type d'interrogatives et sexe dans les sms en francophonie*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	3,314[a]	2	,191
Rapport de vraisemblance	2,550	2	,279
Nombre d'observations valides	11659		

a. 2 cellules (33,3%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de ,27.

FIGURE A.110 – *Test khi-carré mesurant la dépendance des variables type d'interrogatives et langue maternelle pour le corpus francophone global*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	40,997[a]	1	,000		
Correction pour la continuité	40,470	1	,000		
Rapport de vraisemblance	46,085	1	,000		
Test exact de Fisher				,000	,000
Nombre d'observations valides	11659				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 193,59.

b. Calculé uniquement pour un tableau 2x2

FIGURE A.111 – *Test t comparant les médianes d'utilisation de types d'interrogatives en fonction du nombre de sms envoyés par semaine*

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
sms_groupe	Hypothèse de variances égales	14,336	,000	7,340	11657	,000	,207	,028	,152	,262
	Hypothèse de variances inégales			7,482	2124,154	,000	,207	,028	,153	,261

FIGURE A.112 – *Test khi-carré mesurant la dépendance des variables type d'interrogatives et niveau d'études pour le corpus francophone global*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	28,927[a]	2	,000
Rapport de vraisemblance	29,938	2	,000
Nombre d'observations valides	11659		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 188,87.

FIGURE A.113 – *Effectifs relatifs aux types d'interrogatives dans les différentes régions enquêtées*

Tableau croisé Région * Interro

Effectif

		Interro		
		0	1	Total
Région	BE	7016	1210	8226
	CH	1001	119	1120
	LR	1386	64	1450
	QU	682	181	863
Total		10085	1574	11659

FIGURE A.114 – *Test khi-carré mesurant la dépendance des variables type d'interrogatives et région dans les sms*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	162,018[a]	3	,000
Rapport de vraisemblance	188,533	3	,000
Nombre d'observations valides	11659		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 116,51.

A.4.3 Négatives

FIGURE A.115 – *Test khi-carré mesurant la dépendance des variables type de négative et région dans les sms*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	129,162[a]	2	,000
Rapport de vraisemblance	146,159	2	,000
Nombre d'observations valides	10414		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 244,41.

FIGURE A.116 – *Histogramme des effectifs d'utilisation des négatives avec ou sans ne en fonction des catégories d'âge*

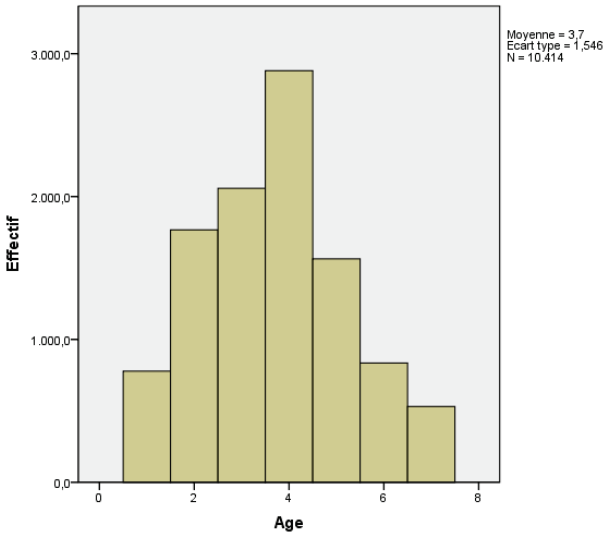


FIGURE A.117 – *Test t comparant les médianes d'utilisation de types de négatives en fonction de l'âge*

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
Age	Hypothèse de variances égales	54,157	,000	-23,538	10412	,000	-,790	,034	-,856	-,724
	Hypothèse de variances inégales			-22,386	4397,238	,000	-,790	,035	-,859	-,721

FIGURE A.118 – *Test khi-carré mesurant la dépendance des variables type de négative et sexe dans les sms*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	30,749[a]	1	,000		
Correction pour la continuité[b]	30,491	1	,000		
Rapport de vraisemblance	30,425	1	,000		
Test exact de Fisher				,000	,000
Nombre d'observations valides	10414				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 967,94.
b. Calculé uniquement pour un tableau 2x2

FIGURE A.119 – *Test khi-carré mesurant la dépendance des variables type de négative et langue maternelle dans les sms*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	42,628[a]	1	,000		
Correction pour la continuité[b]	42,167	1	,000		
Rapport de vraisemblance	45,601	1	,000		
Test exact de Fisher				,000	,000
Nombre d'observations valides	10414				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 304,20.
b. Calculé uniquement pour un tableau 2x2

FIGURE A.120 – *Test khi-carré mesurant la dépendance des variables type de négative et occupation professionnelle dans les sms*

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	169,736[a]	3	,000
Rapport de vraisemblance	153,026	3	,000
Nombre d'observations valides	10414		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 31,21.

Table des figures

2.1 Comparatif des corpus colligés par rapport aux corpus attendus en fonction de la population globale de la région échantillonnée 8

5.1 L'éducation collaborative médiée par ordinateur (Schneider, 1997, p. 9) 23
5.2 Principales études de corpus de sms en linguistique/sociolinguistique (tri décroissant par date) 26
5.3 27
5.4 Panorama des médias sociaux en 2012 (Cavazza, 2012) 34
5.5 Classement de la communication écrite médiée en fonction de ses traits principaux 39
5.6 Exemple de clavier numérique d'un téléphone portable traditionnel . 45
5.7 Racine arabe KTB sans voyellation 50
5.8 Les émoticônes sous l'angle de la coénonciation/colocution selon Pierozak 54
5.9 Les émoticônes sous l'angle de la coénonciation/colocution selon nous 54
5.10 Paramètres pour caractériser le comportement communicatif des interlocuteurs par rappport aux déterminants situationnels et contextuels (Koch et Oesterreicher, 2001 ; p. 586) 60
5.11 Définition du genre de la Cémo en fonction de quelques caractéristiques principales 62

6.1 État du projet sms4science en date du 12 octobre 2011 71
6.2 Méthodologie générale de collecte du projet sms4science 74
6.3 Répartition des sms francophones (sms4science) 80
6.4 Résumé des problèmes généraux liés au traitement des données (sms4science) 84
6.5 Les deux niveaux de transcription des sms en graphie standard . . . 93
6.6 Politique et limites de transcription des sms en graphie standard . . 94
6.7 Double transcription adoptée par l'équipe belge du projet sms4science 95
6.8 Répartition des participants et des profils de participants francophones (sms4science) 103
6.9 Harmonisation des questions de profil du tronc commun (sms4science) 108
6.10 Harmonisation des questions de pratiques du tronc commun (sms4science) 109
6.11 Répartition de l'âge dans les corpus francophones (sms4science) . . . 111
6.12 Répartition du sexe dans les corpus francophones (sms4science) . . . 112
6.13 Répartition de la profession dans les corpus francophones (sms4science) 113
6.14 Répartition du niveau d'étude dans les corpus francophones (sms4science) 115

6.15 Répartition de la fréquence d'envoi des sms dans les corpus francophones (sms4science) . . . 115
6.16 Répartition de la langue première dans les corpus francophones (sms4science) . . . 116
6.17 Répartition des autres langues pratiquées par les usagers dans les corpus francophones (sms4science) . . . 117
6.18 Répartition des pratiques de Cémo dans les corpus francophones(sms4science) . . . 119

7.1 Représentations des distributions statistiques (Carricano *et al.*, 2010) 132

8.1 Principales caractéristiques lexicométriques des 4 corpus de sms francophones . . . 136
8.2 Rapports types/tokens (TTR) et hapax/tokens (HTR) dans les corpus de sms . . . 137
8.3 Rapports types/tokens et hapax/tokens en fonction de classes d'âge . 138
8.4 Classements grammaticaux du lexique (1.000 mots les plus fréquents du français parlé – 1.000 mots les plus fréquents du corpus de sms) . 139
8.5 Moyennes du nombre de caractères par mot dans les corpus de sms brut et transcrit . . . 140
8.6 Thèmes principaux abordés par les cooccurrences issues du corpus de sms francophone . . . 143
8.7 Effectifs et pourcentages des sms contenant (1) ou non (0) un néologisme 146
8.8 Classification des éléments allogènes selon leur type de composition . 158
8.9 Cooccurrences pour les 10 éléments allogènes les plus fréquents (lignes et colonnes par ordre décroissant) . . . 159
8.10 Effectifs et classement des emprunts dans le corpus belge . . . 166
8.11 Box-plot des emprunts du corpus belge en fonction de l'âge des auteurs 167
8.12 Analyse des résidus standardisés relatifs au khi-carré (emprunts-sexe) 168
8.13 Analyse des résidus standardisés relatifs au khi-carré (emprunts-études) . . . 169
8.14 Box-plot des emprunts du corpus belge en fonction du nombre de sms envoyés par semaine . . . 169
8.15 Effectifs et classement des emprunts dans le corpus réunionnais . . . 170
8.16 Box-plot des emprunts du corpus réunionnais en fonction de l'âge des auteurs . . . 170
8.17 Analyse des résidus standardisés relatifs au khi-carré (emprunts-sexe) 171
8.18 Distribution du nombre de sms envoyés par semaine (information donnée par les auteurs des messages composant le corpus réunionnais) 172
8.19 Effectifs et classement des emprunts dans le corpus québécois 173
8.20 Box-plot des emprunts du corpus québécois en fonction de l'âge des auteurs . . . 173
8.21 Analyse des résidus standardisés relatifs au khi-carré (emprunts-sexe) 174

8.22 Box-plot des emprunts du corpus québécois en fonction du nombre de sms envoyés par semaine . . . 175
8.23 Effectifs et classement des emprunts dans le corpus suisse . . . 176
8.24 Box-plot des emprunts du corpus suisse en fonction de l'âge des auteurs 176
8.25 Analyse des résidus standardisés relatifs au khi-carré (emprunts-études) . . . 177
8.26 Hypothèses liant (H1) ou non (H0) les variables et les médianes de variable dans chaque région . . . 178
8.27 Classement des langues d'emprunt pour le corpus global sms4science francophone . . . 179
8.28 Classement des emprunts les plus fréquents pour le corpus global sms4science francophone . . . 179
8.29 Test Anova pour la comparaison des médianes (variable qualitative à plus de deux modalités) . . . 180
8.30 Exemple de concordance de la variété de *savoir* propre au français de Belgique . . . 189
8.31 Effectifs et pourcentages des sms contenant au moins un belgicisme . 190
8.32 Distribution, selon l'âge, des sms comprenant au moins un belgicisme 191
8.33 Test khi-carré mesurant la dépendance des variables savoir et sexe en Belgique . . . 192
8.34 Effectifs et pourcentage des sms contenant au moins une unité spécifique au français régional de La Réunion . . . 196
8.35 Exemple de concordance pour l'adverbe à soir dans le corpus de sms québécois . . . 200
8.36 Effectifs et pourcentage des sms contenant au moins un québécisme . 201
8.37 Test de Mann-Whitney pour la comparaison des médianes d'utilisation de régionalismes selon l'âge, au Québec . . . 202
8.38 Répertoire des helvétismes identifiés dans le corpus de sms suisse . . 203
8.39 Helvétismes les plus fréquents dans le corpus suisse francophone . . . 206
8.40 Exemple de concordance du verbe redire dans le corpus suisse 206
8.41 Effectifs et pourcentage des sms contenant au moins un helvétisme . 207
8.42 Distribution des sms comprenant au moins un helvétisme selon l'âge 208
8.43 Test khi-carré mesurant la dépendance des variables redire et sexe en Suisse . . . 209
8.44 Différences de pourcentages entre taux de sms avec éléments allogènes et régionalismes . . . 210
8.45 Hypothèses relatives à l'emploi de régionalismes, liant les variables ou les médianes de variables dans chaque région (H1 = dépendance ; H0 = indépendance) . . . 211

9.1 Distribution des variantes orthographiques de l'écrit sms selon leur type . . . 217
9.2 Typologie des variantes orthographiques 2 (inspirée par Catach, 1980) 219

9.3 Les 10 variantes les plus courantes dans l'accord sujet-verbe conjugué (corpus de sms belge, par ordre décroissant) 220
9.4 Résidus standardisés ajustés liés au test khi-carré pour les variables sujet et auxiliaire dans une construction au participe passé engendrant de la variation 222
9.5 Analyse détaillée de 4 types de variantes morphosyntaxiques dans les sms 223
9.6 Box-plot présentant la distribution des confusions (ou non-confusions) des temps futur et conditionnel en fonction des classes d'âge 225
9.7 Résultats des test khi-carré pour le calcul de dépendance des variables *confusion des temps indicatif futur et conditionnel* et *pratiques de la Cémo* 226
9.8 Box-plot présentant la distribution des confusions (ou non-confusions) de modes participe passé et infinitif en fonction des classes d'âge . . 226

10.1 Mesures d'évaluation des grammaires en vue de l'identification d'expressions dans un corpus 233
10.2 Répartition des types de structures d'interrogatives formelles 235
10.3 Répartition des types d'interrogatives moins formelles 237
10.4 Répartition des types d'interrogatives pour LOC1 240
10.5 Répartition des types d'interrogatives pour LOC2 241
10.6 Box-plot représentant les types d'interrogatives en fonction des classes d'âge 243
10.7 Analyse des résidus standardisés relatifs au test khi-carré (interrogatives-langue première) 244
10.8 Box-plot représentant les types d'interrogatives en fonction de la fréquence d'envoi de messages 245
10.9 Analyse des résidus standardisés relatifs au test khi-carré (interrogatives-niveau d'études) 245
10.10 Analyse des résidus standardisés relatifs au test khi-carré (interrogatives-région) 246
10.11 Extrait des règles lexico-syntaxiques d'identification des négatives . . 250
10.12 Distribution de la négation en fonction des variables étudiées 252
10.13 Distribution des séquences préformées en fonction du type de négatives 253
10.14 Distribution des types de verbes dans les constructions négatives avec clitique *ne* 254
10.15 Distribution des variables étudiées (pronom, auxiliaire et impératif) pour l'étude de la négation sur base des 10 verbes les plus fréquents dans les négatives en *ne* 257
10.16 Distribution de la variable *forclusif* pour l'étude de la négation sur base des 10 verbes les plus fréquents dans les négatives en *ne* 258
10.17 Analyse des résidus standardisés relatifs au khi-carré (types de négatives - région) 260
10.18 Distribution des types de négatives en fonction des âges (11 à 17) . . 261

10.19 Box-plot présentant la répartition des types de négatives en fonction des catégories d'âge . . . 261
10.20 Effectifs croisés pour les variables *âge* et *type de négative (AI, sans ne ; NAI, avec ne)* . . . 262
10.21 Analyse des résidus standardisés relatifs au khi-carré (types de négatives - sexe) . . . 262
10.22 Analyse des résidus standardisés relatifs au khi-carré (types de négatives - langue première) . . . 263
10.23 Analyse des résidus standardisés relatifs au khi-carré (types de négatives - occupation professionnelle) . . . 264

A.1 Typologie des logiciels de l'ACAO (de Vries, 2001) . . . 307
A.2 Fréquence des activités sur les services de réseautage social (source : Grunwald Associates LLC, http ://www.grunwald.com/) . . . 313
A.3 Exemple de fenêtre Mud contenant plusieurs zones de texte (World of Warcraft, 2001) . . . 317
A.4 Répartition du nombre d'unités lexicales en fonction de la longueur de l'unité (en nombre de caractères, dans les corpus de sms bruts et transcrits) . . . 328
A.5 Unités lexicales des 4 corpus de sms francophones dont la fréquence est supérieure à 1.000 . . . 329
A.6 Mots-clés des corpus de sms par région . . . 330
A.7 100 structures polylexicales nominales les plus fréquentes dans le corpus de sms . . . 331
A.8 Distribution des âges pour le corpus francophone entier . . . 332
A.9 Distribution des âges pour le corpus francophone entier, après rééquilibrage par catégories . . . 332
A.10 Test t comparant les médianes d'utilisation de néologismes en fonction de l'âge . . . 333
A.11 Test t comparant les médianes d'utilisation de néologismes en fonction du nombre de sms envoyés par semaine . . . 333
A.12 Test du khi-carré mesurant la dépendance des variables *néologisme* et *sexe* en francophonie . . . 334
A.13 Test du khi-carré mesurant la dépendance des variables *néologisme* et *études* en francophonie . . . 334
A.14 Test du khi-carré mesurant la dépendance des variables *néologisme* et *langue maternelle* en francophonie . . . 335
A.15 Test du khi-carré mesurant la dépendance des variables *néologisme* et *chat* en francophonie . . . 335
A.16 Test du khi-carré mesurant la dépendance des variables *néologisme* et *forum de discussion* en francophonie . . . 335
A.17 Test du khi-carré mesurant la dépendance des variables *néologisme* et *courrier électronique* en francophonie . . . 336

A.18 Liste de candidats à la néologie présents dans le corpus sms pour la science (belge), classée par nombre d'occurrences 337
A.19 Liste de candidats à la néologie présents dans le corpus sms pour la science (belge), classée par nombre d'occurrences (suite) 338
A.20 Liste de candidats à la néologie présents dans le corpus réunionnais et classée par nombre d'occurrences 339
A.21 Liste de candidats à la néologie présents dans le corpus québécois et classée par nombre d'occurrences 340
A.22 Liste de candidats à la néologie présents dans le corpus suisse et classée par nombre d'occurrences 341
A.23 Emprunts de langue récents présents dans l'édition 2012 du Petit Robert et/ou le dictionnaire de l'argot, du populaire et du familial en ligne 342
A.24 Répertoire des salutations les plus courantes dans les sms belges francophones (Cougnon et François, 2010) 342
A.25 Emprunts dans le corpus belge (détail) 343
A.26 Box-plot des emprunts du corpus belge en fonction de l'âge des auteurs de moins de 30 ans 343
A.27 Test t comparant les médianes des emprunts et non emprunts en fonction de l'âge dans le corpus belge 344
A.28 Test du khi-carré mesurant la dépendance des variables *emprunt* et *sexe* dans le corpus belge 344
A.29 Test du khi-carré mesurant la dépendance des variables *emprunt* et *niveau d'étude* dans le corpus belge 345
A.30 Test t comparant les médianes d'utilisation d'emprunts en fonction du nombre de sms envoyés par semaine dans le corpus belge 345
A.31 Emprunts dans le corpus réunionnais (détail) 346
A.32 Test U de Mann-Whitney comparant les médianes des emprunts et non emprunts en fonction de l'âge dans le corpus réunionnais 346
A.33 Test du khi-carré d'indépendance entre les variables *emprunt* et *sexe* dans le corpus réunionnais 346
A.34 Test du khi-carré d'indépendance entre les variables *emprunt* et *études* dans le corpus réunionnais 347
A.35 Test Mann-Whitney comparant les médianes des variables *emprunt* et *nombre de sms* dans le corpus réunionnais 347
A.36 Emprunts dans le corpus québécois (détail) 348
A.37 Test U de Mann-Whitney comparant les médianes des emprunts et non emprunts en fonction de l'âge dans le corpus québécois 348
A.38 Test du khi-carré d'indépendance entre les variables *emprunt* et *sexe* dans le corpus québécois 348
A.39 Test du khi-carré d'indépendance entre les variables *emprunt* et *études* dans le corpus québécois 349
A.40 Distribution du nombre de sms par semaine dans le corpus québécois 349

A.41 Test Mann-Whitney comparant les médianes des variables *emprunt* et *nombre de sms* dans le corpus québécois 349
A.42 Emprunts dans le corpus suisse (détail) 350
A.43 Test U de Mann-Whitney comparant les médianes des emprunts et non emprunts en fonction de l'âge dans le corpus suisse 350
A.44 Test du khi-carré d'indépendance entre les variables *emprunt* et *sexe* dans le corpus suisse 350
A.45 Test du khi-carré d'indépendance entre les variables *emprunt* et *études* dans le corpus suisse 351
A.46 Test Mann-Whitney comparant les médianes des variables *emprunt* et *nombre de sms* dans le corpus suisse 351
A.47 Description de la variable emprunt dans le corpus global de sms . . . 352
A.48 Distribution globale des âge pour le corpus sms4science francophone 352
A.49 Test t sur le corpus global comparant les médianes d'emprunts/non emprunts en fonction de l'âge 352
A.50 Test du khi-carré sur le corpus global calculant la dépendance entre les variables *emprunt* et *sexe* 353
A.51 Test du khi-carré sur le corpus global calculant la dépendance entre les variables *emprunt* et *niveau d'études* 353
A.52 Distribution globale des fréquences d'envoi de sms pour le corpus sms4science francophone 354
A.53 Test t pour le calcul de l'égalité des médianes emprunts/non emprunts en fonction de la fréquence d'envoi des sms dans le corpus global . . 354
A.54 Test d'égalité des moyennes concernant les médianes des variables *langue de l'emprunt* et *âge* dans le corpus global 355
A.55 Test du khi-carré calculant la dépendance entre les variables *langue de l'emprunt* et *sexe* dans le corpus global 355
A.56 Test du khi-carré pour l'étude de dépendance entre les variables *langue d'emprunt* et *études supérieures* dans le corpus global 356
A.57 Test du khi-carré pour l'étude de dépendance entre les variables *emprunt anglais* et *études supérieures* dans le corpus global 356
A.58 Test du khi-carré pour l'étude de dépendance entre les variables *langue d'emprunt* et *langue maternelle* dans le corpus global 356
A.59 Test du khi-carré pour l'étude de dépendance entre les variables *emprunt ou non emprunt* et *langue maternelle* dans le corpus global . . 357
A.60 Détail des résidus standardisés ajustés liés au khi-carré pour l'étude de dépendance entre les variables *langue d'emprunt* et *langue maternelle* dans le corpus global 357
A.61 Répertoire des belgicismes identifiés dans le corpus de sms belge . . . 358
A.62 Test du khi-carré mesurant la dépendance des variables *belgicisme* et *sexe* 359
A.63 Box-plot des belgicismes en fonction de l'âge des auteurs (de moins de 31 ans) 359

A.64 Test t mesurant la dépendance des variables *belgicisme* et *âge* (moins de 31 ans) . . . 359
A.65 Test du khi-carré mesurant la dépendance des variables *belgicisme* et *utilisation du chat* . . . 360
A.66 Test du khi-carré mesurant la dépendance des variables *belgicisme* et *utilisation du forum de discussion* . . . 360
A.67 Liste des 25 belgicismes les plus fréquents du corpus de sms . . . 361
A.68 Test du khi-carré mesurant la dépendance des variables *utilisation de savoir* et *type d'occupation* . . . 361
A.69 Test du khi-carré mesurant la dépendance des variables *utilisation de savoir* et *pratique du courrier électronique* . . . 362
A.70 Répertoire des régionalismes présents dans le corpus de sms réunionnais 363
A.71 Liste de 25 régionalismes les plus fréquents du corpus de sms réunionnais 363
A.72 Test du khi-carré mesurant la dépendance des variables régionalisme et sexe de l'auteur à La Réunion . . . 364
A.73 Test U de Mann-Whitney comparant les médianes des régionalismes et non régionalismes en fonction de l'âge à La Réunion . . . 364
A.74 Test du khi-carré mesurant la dépendance des variables régionalisme et utilisation du chat à La Réunion . . . 364
A.75 Test du khi-carré mesurant la dépendance des variables régionalisme et utilisation du forum à La Réunion . . . 365
A.76 Répertoire des québécismes identifiés dans le corpus de sms québécois 366
A.77 Test du khi-carré mesurant la dépendance des variables régionalisme et sexe de l'auteur au Québec . . . 367
A.78 Test du khi-carré mesurant la dépendance des variables régionalisme et pratique du chat au Québec . . . 367
A.79 Test du khi-carré mesurant la dépendance des variables régionalisme et pratique des forums de discussion au Québec . . . 367
A.80 Test du khi-carré mesurant la dépendance des variables régionalisme et sexe de l'auteur en Suisse . . . 368
A.81 Test du khi-carré mesurant la dépendance des variables régionalisme et pratique du chat en Suisse . . . 368
A.82 Test du khi-carré mesurant la dépendance des variables régionalisme et pratique des forums de discussion en Suisse . . . 368
A.83 Test du khi-carré mesurant la dépendance des variables utilisation de *redire* et occupation professionnelle en Suisse . . . 369
A.84 Test khi-carré mesurant la dépendance des variables *redire* et éducation en Suisse . . . 369
A.85 Comparaison des variantes régionales du corpus de sms avec leurs variantes en français standard (et différence du nombre de caractères entre variantes) . . . 370
A.86 Typologie des variantes orthographiques 1 (inspirée par Catach : 1980) 371
A.87 Typologie des variantes orthographiques 3 (inspirée par Catach : 1980) 372

A.88 Les verbes les plus concernés par les fautes d'accord sujet-verbe du corpus de sms belge 372
A.89 Les groupes verbaux les plus concernés par la confusion des modes verbaux dans le corpus de sms belge 372
A.90 Les verbes les plus concernés par la confusion de l'accord 1e/2e p. de l'indicatif dans le corpus de sms belge 373
A.91 Histogramme des effectifs pour la confusion ou non confusion des temps futur et confitionnel en fonction des catégories d'âge 374
A.92 Test t comparant les médianes de confusion des temps futur et conditionnel en fonction de l'âge 374
A.93 Test khi-carré mesurant la dépendance des variables *confusion des temps indicatif futur et conditionnel* et *sexe* dans les sms 375
A.94 Test khi-carré mesurant la dépendance des variables *confusion des temps indicatif futur et conditionnel* et *niveau d'études* dans les sms 375
A.95 Test khi-carré mesurant la dépendance des variables *confusion des temps indicatif futur et conditionnel* et *pratique du chat* dans les sms 375
A.96 Test khi-carré mesurant la dépendance des variables *confusion des temps indicatif futur et conditionnel* et *fréquentation de forums* dans les sms 376
A.97 Test khi-carré mesurant la dépendance des variables *confusion des temps indicatif futur et conditionnel* et *pratique du courrier électronique* dans les sms 376
A.98 Histogramme des effectifs pour la confusion ou non confusion des modes participe passé et infinitif en fonction des catégories d'âge . . 376
A.99 Test t comparant les médianes de confusion des modes participe passé et infinitif en fonction de l'âge 377
A.100 Test khi-carré mesurant la dépendance des variables *confusion des modes participe passé et infinitif* et *sexe* dans les sms 377
A.101 Test khi-carré mesurant la dépendance des variables *confusion des modes participe passé et infinitif* et *niveau d'études* dans les sms . . 378
A.102 Test khi-carré mesurant la dépendance des variables *confusion des modes participe passé et infinitif* et *pratique du chat* dans les sms . . 378
A.103 Test khi-carré mesurant la dépendance des variables *confusion des modes participe passé et infinitif* et *utilisation du courrier électronique* dans les sms 378
A.104 Test khi-carré mesurant la dépendance des variables *confusion des modes participe passé et infinitif* et *fréquentation des forums de discussion* 378
A.105 Graphe représentant les structures interrogatives standard du corpus de sms basées sur "est-ce que" 379
A.106 Graphe représentant les structures interrogatives standard du corpus de sms basées sur un élement qu- et une inversion 380
A.107 Liste des verbes apparaissant dans les interrogatives standard basées sur une inversion sujet-verbe 381

A.108 Test t comparant les médianes d'utilisation de types d'interrogatives en fonction de l'âge . . . 381
A.109 Test khi-carré mesurant la dépendance des variables *type d'interrogatives* et *sexe* dans les sms en francophonie . . . 382
A.110 Test khi-carré mesurant la dépendance des variables *type d'interrogatives* et *langue maternelle* pour le corpus francophone global . . . 382
A.111 Test t comparant les médianes d'utilisation de types d'interrogatives en fonction du nombre de sms envoyés par semaine . . . 382
A.112 Test khi-carré mesurant la dépendance des variables *type d'interrogatives* et *niveau d'études* pour le corpus francophone global . . . 383
A.113 Effectifs relatifs aux types d'interrogatives dans les différentes régions enquêtées . . . 383
A.114 Test khi-carré mesurant la dépendance des variables *type d'interrogatives* et *région* dans les sms . . . 383
A.115 Test khi-carré mesurant la dépendance des variables *type de négative* et *région* dans les sms . . . 384
A.116 Histogramme des effectifs d'utilisation des négatives avec ou sans *ne* en fonction des catégories d'âge . . . 384
A.117 Test t comparant les médianes d'utilisation de types de négatives en fonction de l'âge . . . 384
A.118 Test khi-carré mesurant la dépendance des variables *type de négative* et *sexe* dans les sms . . . 385
A.119 Test khi-carré mesurant la dépendance des variables *type de négative* et *langue maternelle* dans les sms . . . 385
A.120 Test khi-carré mesurant la dépendance des variables *type de négative* et *occupation professionnelle* dans les sms . . . 385

Cahiers du Cental

La collection « Cahiers du Cental » est une publication du Centre de traitement automatique du langage de l'Université catholique de Louvain
http://www.uclouvain.be/cental

Hors-série

Purnelle G., Fairon C. et Dister A. (éds) (2004), *Le poids des mots, Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, 2 vols, Presses universitaires de Louvain, Louvain-la-Neuve, 1219 p.

Cahiers du Cental

1. Didier J.-J., Hambursin O., Moreau Ph. et Seron M. (éds) (2006), *« Le français m'a tuer », Actes du colloque L'orthographe française à l'épreuve du supérieur*, Bruxelles 27 mai 2005, Presses universitaires de Louvain, Louvain-la-Neuve, v-113 p.
2. Mertens P., Fairon C., Dister A. et Watrin P. (éds) (2006), *Verbum ex machina, Actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles* (2006), Leuven, 10-13 avril 2006, Presses universitaires de Louvain, Louvain-la-Neuve, 2 vols, xviii-951 p.

3.1. Fairon C., Klein J.R. et Paumier S. (2006), *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête « Faites donc de vos SMS à la science »*, Presses universitaires de Louvain, Louvain-la-Neuve, viii-123 p.

3.2. Fairon C., Klein J.R. et Paumier S. (2006), *Le corpus SMS pour la science. Base de données de 30.000 SMS et logiciel de consultation*, CD-Rom, Presses universitaires de Louvain, Louvain-la-Neuve, v-44 p.

4. Fairon C., Naets, H., Kilgarriff A. et de Schryver G.-M. (éds) (2007), *Building and Exploring Web Corpora*, *Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval*, Presses universitaires de Louvain, Louvain-la-Neuve, viii-167 p.
5. Constant M., Dister A., Emirkanian L. et Piron S. (éds) (2008), *Description linguistique pour le traitement automatique du français,* Presses universitaires de Louvain, Louvain-la-Neuve, v-246 p.
6. Nakamura T., Laporte É., Dister A. et Fairon C. (éds) (2010), *Les tables. La grammaire du français par le menu. Mélanges en hommage à Christian Leclère,* Presses universitaires de Louvain, Louvain-la-Neuve, xvii-383 p.
7. Granger S. et Paquot M. (eds) (2010), *eLexicography in the 21st century : New challenges, new applications, Proceedings of eLex 2009*, ix-455 p.
8. Cougnon L.-A. (2015), *Langage et sms. Une étude internationale des pratiques actuelles.* Presses universitaires de Louvain, Louvain-la-Neuve, 416 p.

Commandez les *Cahiers du Cental* en ligne / Order *Cahiers du Cental* online :
http://pul.uclouvain.be