HAL
open science

Communiquer par SMS : Analyse automatique du langage et extraction de l'information véhiculée

Eleni Kogkitsidou

To cite this version:

Eleni Kogkitsidou. Communiquer par SMS : Analyse automatique du langage et extraction de l'information véhiculée. Linguistique. Université Grenoble Alpes, 2018. Français. NNT : 2018GREAL012 . tel-01968698

HAL Id: tel-01968698

https://theses.hal.science/tel-01968698

Submitted on 3 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Communauté UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Sciences du langage, Spécialité Informatique et Sciences du langage**

Arrêté ministériel : 25 mai 2016

Présentée par

Eleni KOGKITSIDOU

Thèse dirigée par **Georges ANTONIADIS**

préparée au sein du **Laboratoire LIDILEM – EA 609**
dans **l'École Doctorale Langues, Littérature et Sciences Humaines**

Communiquer par SMS : Analyse automatique du langage et extraction de l'information véhiculée

Thèse soutenue publiquement le **27 septembre 2018,**
devant le jury composé de :

M. Georges ANTONIADIS
Professeur, Université Grenoble Alpes, Directeur de thèse
M. Cédrick FAIRON
Professeur, Université Catholique de Louvain, Rapporteur
Mme Tita KYRIACOPOULOU
Professeur, Université Paris-Est Marne-la-Vallée, Rapporteur
Mme Gudrun LEDEGEN
Professeur, Université de Rennes 2, Examinatrice et Présidente du jury
Mme Rachel PANCKHURST
Maître de Conférences HDR, Université Paul-Valéry Montpellier 3, Examinatrice
M. Matthieu QUIGNARD
Ingénieur de recherche au CNRS, UMR 5191 ICAR, Examinateur

Remerciements

J'adresse de chaleureux remerciements à mon directeur de thèse, Georges Antoniadis, pour m'avoir donné l'opportunité de réaliser ce travail, pour sa confiance et son encadrement constant pendant toutes ces années.

Je voudrais remercier sincèrement mon co-encadrant, Matthieu Quignard, pour ses suggestions, ses conseils et sa disponibilité.

Je tiens également à remercier Cédrick Fairon, Tita Kyriacopoulou, Gudrun Ledegen et Rachel Panckhurst d'avoir accepté de faire partie de mon jury, ainsi que pour leurs critiques constructives et pertinentes.

Je souhaite également remercier la région Rhône-Alpes et son programme ARC6 qui m'ont permis de réaliser mon travail de recherche.

Je remercie toute l'équipe du LIDILEM et notamment Zohra Bouhania et Isabelle Rousset pour leur aide et leur disponibilité. Un grand merci à tous mes amis de thèse, Emilie, Lucie, Agnès, AÏcha, Alex, Claire, Carole, Manon, Armelle, Yuko, Julie qui m'ont intégré dès le premier jour et m'ont permis de garder un très bon souvenir de mon séjour à Grenoble.

Je tiens à remercier tout particulièrement ma Kyria, qui la première m'a donné l'envie d'étudier le traitement automatique des langues. Je ne saurais jamais comment assez la re-

mercier, pour l'aide, le soutien et la confiance qu'elle m'a fait dès le début de mon périple.

Je voudrais remercier Claude Martineau pour les relectures et sa disponibilité permanente pour m'aider à mener cette thèse à son terme.

J'en profite pour remercier l'UFR de sociologie et d'informatique pour les sciences humaines de Sorbonne université, plus particulièrement, Claude Montacié de m'avoir offert la possibilité de travailler en tant qu'ATER au sein de son équipe pendant ces deux dernières années. Je remercie aussi mon collègue Gaël Lejeune pour son aide et encouragement constant.

Je souhaite aussi remercier tous mes très chers amis pour leur soutien et leur disponibilité à n'importe quelle heure du jour et de la nuit, peu importe les différents pays où nous étions.

Je remercie ma famille pour leur soutien, ma mère et mon père qui ont toujours encouragé mes rêves et mes projets.

Cette thèse, je la dédie à Cristian, mon meilleur ami, mon confident, mon mari. Sans ton aide, tes conseils et ton soutien à toute épreuve, cette thèse n'aurait jamais pu voir le jour.

Résumé

Cette thèse concerne l'analyse automatique des SMS et l'extraction des informations qui y sont contenues. Le point de départ de notre recherche est le constat que la plupart des messages courts, observés dans le corpus alpes4science, présentent des différences en comparaison avec le langage standard. Les différences sont mises en évidence, d'une part, par la morphologie particulière des mots et, d'autre part, par les règles de syntaxe et de grammaire qui ne sont pas respectées lorsque l'émetteur considère que cela ne nuit pas à l'intelligibilité du message. À cause des écarts par rapport à la langue standard, le traitement et l'analyse des messages bruités est toujours un défi pour les tâches du TAL. Par conséquent, réduire les écarts est un défi fondamental que nous surmontons en utilisant l'approche de la normalisation pour la conception d'outils en traitement automatique des SMS. Nous proposons un modèle de normalisation en deux étapes, fondé sur une approche symbolique et statistique. La première étape vise à produire une représentation intermédiaire du message SMS par l'application de grammaires locales, tandis que la deuxième utilise un système de traduction automatique à base de règles pour convertir la représentation intermédiaire vers une forme standard. Le résultat produit par ce modèle a été évalué, par la suite, pour la reconnaissance d'entités nommées au travers d'une série de tests appliqués à l'aide de trois autres systèmes. Les résultats obtenus ont montré que les performances de ces systèmes de reconnaissance d'entités nommées présentent des améliorations significatives lorsqu'ils sont appliqués sur les SMS automatiquement normalisés en comparaison avec le corpus brut et manuellement transcrit.

Mots-clés : communication médiée par ordinateur, langage SMS, normalisation des SMS, extraction d'informations

Abstract

This thesis focuses on SMS language and information extraction from the point of view of natural language processing. The starting point of our study is the observation of the differences that most short messages have, using the alpes4science corpora, in comparison with the standard language. The differences are highlighted by the particular morphology of words and by the syntactic and grammar rules that are not respected when the issuer considers that it would not impair the intelligibility of the message. Because of the deviations from the standard language, processing and analyzing noisy messages is still a challenge for any NLP task. Therefore, reducing the gaps is a fundamental step to overcome when designing approaches for automatic SMS processing. We propose a two-step normalization model based on a symbolic and statistical approach. The first step aims to produce an intermediate representation of the SMS by applying local grammars. The second step uses a rule-based machine translation system to convert the intermediate representation to a standard form. The obtained result from this model was evaluated, afterwards, for named entities recognition through a series of tests applied thanks to three other systems. The results have shown that these performances of named entity recognition systems are significantly improved when applied to automatically normalized SMS in comparison with raw and manually normalized corpora.

Keywords : computer-mediated communication, SMS language, SMS normalization, information extraction

Table des matières

1 Introduction **13**

2 La communication par SMS **23**

2.1 Introduction 23

2.2 Définition de la communication virtuelle écrite 24

2.2.1 Caractéristiques de la communication virtuelle écrite 26

2.2.1.1 La communication synchrone ou asynchrone 26

2.2.1.2 La communication hautement interactive 27

2.2.1.3 La communication orale/écrite 28

2.2.1.4 La participation active ou passive 31

2.3 Certaines caractéristiques de la communication par SMS 31

2.3.1 Particularités et phénomènes graphiques 32

2.3.1.1 L'abréviation 32

2.3.1.2 La phonétisation 34

2.3.1.3 L'alternance codique 35

2.3.1.4 Les émoticônes 37

2.4 Conclusion 38

3 Le projet alpes4science **39**

3.1 Introduction 39

3.2 Alpes4science 40

3.2.1 Méthodologie de recueil des données 41

3.2.2 L'enregistrement des données 42

3.2.2.1 Les problèmes rencontrés 45

3.2.3 Traitement automatique du corpus 45

3.2.3.1 Anonymisation 45

3.2.3.2 Méthode d'anonymisation 46

3.2.3.3 Transcription 49

3.2.3.4 Méthode de transcription 49

3.3 Les données récoltées 53

3.3.1 Le questionnaire 53

3.3.2 Le corpus SMS 55

3.4 Conclusion 58

4 L'analyse du corpus alpes4science 61

4.1 Introduction 61

4.2 Présentation des données socio-démographiques 62

4.2.1 Le profil du participant 63

4.2.1.1 Le sexe 63

4.2.1.2 L'age 63

4.2.1.3 Le niveau d'études 65

4.2.1.4 Langue maternelle et régionale 65

4.2.1.5 La répartition géographique 66
4.2.2 Les pratiques des participants 67
4.2.2.1 L'usage et compréhension sms 67
4.2.2.2 Nombre d'envois par semaine 69
4.3 Analyse lexicométrique 69
4.3.1 Principales caractéristiques lexicométriques du corpus 71
4.3.1.1 Mesure de la diversité/richesse lexicale 72
4.3.1.2 L'âge 76
4.3.1.3 Niveau d'études 79
4.3.1.4 Emploi des catégories grammaticales 80
4.3.1.5 Type de clavier 82
4.3.1.6 Les n-grammes 83
4.4 Analyse et typologie des pratiques langagières 84
4.4.1 Typologie des formes du corpus 87
4.5 Conclusion 90
5 Un système hybride pour la normalisation de SMS 93
5.1 Introduction 93
5.2 Motivations 94
5.2.1 Revue de travaux antérieurs 95
5.2.2 Problèmes de tokenisation pour la normalisation 97
5.3 Approche de normalisation 101
5.3.1 Représentation intermédiaire 102

5.3.1.1 Expression graphique de sentiments 104
5.3.1.2 Locutions et expressions 104
5.3.1.3 Répétition de caractères 109
5.3.1.4 Mots hors vocabulaire 110
5.3.2 Modèle de Traduction Automatique 110
5.3.2.1 Apertium : une plateforme pour la traduction automatique . . 111
5.3.2.2 Induction de dictionnaires pour la normalisation de SMS . . . 115
5.3.2.3 Désambiguïsation de formes polysémiques 118
5.3.2.4 La traduction 119
5.4 Évaluation 122
5.4.1 Les résultats d'évaluation 123
5.4.1.1 Types d'erreurs détectés 128
5.4.2 Évaluation du système sur le corpus 88milSMS 129
5.5 Conclusion 130
6 Autour de la reconnaissance d'entités nommées dans les SMS 133
6.1 Introduction 133
6.2 Les entités nommées 134
6.2.1 Une définition de l'entité nommée 138
6.2.2 Les catégories 140
6.2.3 Les difficultés de la catégorisation 143
6.2.3.1 La polysémie 143
6.2.3.2 Les frontières 146

6.2.4 Pourquoi extraire des entités nommées ? 147

6.2.4.1 Applications indirectes 148

6.2.4.2 Application directe 151

6.3 La reconnaissance d'entités nommées dans les SMS 155

6.3.1 Une typologie d'entités nommées pour les SMS 161

6.4 Conclusion 166

7 Apport de la normalisation à la reconnaissance d'entités nommées 167

7.1 Introduction 167

7.2 Systèmes d'extraction d'entités nommées appliquées aux SMS 168

7.2.1 Présentation du corpus 168

7.2.1.1 Normalisation de la casse pour le corpus normalisé 171

7.2.2 Système Baseline 172

7.2.2.1 API 172

7.2.2.2 Logiciels libres 173

7.3 Évaluation 176

7.3.1 Mesures de performance 176

7.3.2 Évaluation et analyse des résultats 177

7.3.2.1 Analyse des résultats 180

7.4 Conclusion 184

Conclusion et Perspectives 185

Bibliographie 189

Annexe A Corpus alpes4science 215

Annexe B Questionnaire du projet **233**

Annexe C Données lexicometriques du corpus alpes4science **239**

Annexe D Typologie de SMS **243**

Annexe E Graphe de reconnaissance **245**

Annexe F Corpus de normalisation **247**

Liste des tableaux **259**

Table des figures **263**

Chapitre 1

Introduction

Introduction générale

La communication est un besoin de l'homme qui a commencé à émerger depuis la préhistoire. Dans une phase primitive, la voix de l'homme était le seul moyen de communication, avant de procéder au transfert d'informations à distance avec des messagers, des signaux visuels ou sonores. Depuis la transmission des messages par le feu à la diffusion de l'écriture, qui a commencé avec l'échange de messages de texte, la communication postale a constitué au 19e siècle le moyen de communication le plus important.

La recherche de nouveaux modes de communication avec l'exploitation des lois de l'induction électromagnétique de Faraday, a conduit Samuel Morse à la conception du télégraphe en 1830. Au début du 20e siècle, la télégraphie sans fil [1] est devenue de plus en plus répandue. Après les premiers pas du télégraphe, la recherche s'est intensifiée pour rendre possible une conversation à distance permettant aux personnes éloignées de communiquer en temps réel.

1. La télégraphie sans fil (TSF), permet la transmission à distance d'un écrit en utilisant des ondes électromagnétiques (sans fil télégraphique).

Nous pouvons considérer que l'histoire de la téléphonie a commencé dans les années 1870 et que depuis elle est dans un développement continu.

Au seuil du 21e siècle, une grande partie des besoins communicationnels se réalise au moyen d'un support électronique. Grâce à la communication électronique l'homme profite d'un échange de données numériques afin de communiquer avec d'autres personnes par téléphone et par divers services fournis par des systèmes informatiques. La communication électronique se réalise au moyen de dispositifs tels que le téléphone, l'ordinateur, le fax etc. Ce développement a donné lieu à des nouvelles formes de communication interpersonnelle.

Parmi ces nouvelles formes nous repérons la communication électronique écrite, définie aussi par le terme *communication médiée par ordinateur* (*cf.* chapitre 2.2). Ce terme décrit des dispositifs numériques qui interviennent pour transmettre et recevoir des messages écrits. Il inclut les SMS (*Short Message Service*), les messageries instantanées, les jeux en ligne multi-joueurs, les courriers électroniques, etc. Le terme implique l'utilisation d'un moyen électronique comme médium de communication. Cette activité passe souvent par le téléphone portable et/ou l'internet. Parmi les messageries instantanées nous trouvons des applications mobiles multiplateforme qui fournissent un système de messagerie instantanée via Internet et via les réseaux mobiles (WhatsApp, Viber, Facebook Messenger, Google Hangouts, etc.). La conversation peut être synchrone et asynchrone puisque l'échange peut s'effectuer en temps réel ou en mode différé.

Le service mobile SMS permet à l'utilisateur d'envoyer et de recevoir un message de texte court à d'autres utilisateurs au moyen du téléphone portable. Le *Short Message Service* signifie service de message court et a fait son apparition dans les années 80. Il s'agit d'un service très utilisé par toutes les classes d'âges, notamment les jeunes, afin de transmettre rapidement des messages texte concis à tout moment (Markett *et al.*, 2006). L'Arcep (Autorité de Régulation des Communications Electroniques et des Postes) affirme qu'en 2015 la consommation des messages continuait à s'accroître avec 207 milliards de messages envoyés (dont le 98% sont des

SMS)[2]. Cependant, le rythme annuel de croissance du nombre de messages s'est nettement affaibli sur les trois dernières années : +11 milliards en 2013 , + 4 milliards en 2014 et près de +7 milliards en 2015

Le SMS a de nombreux avantages, puisque dans la majorité des cas, il faut moins de temps pour envoyer un SMS que pour passer un appel téléphonique, envoyer un e-mail ou un fax. Le coût du SMS est assez faible, selon le forfait téléphonique et il peut être transmis et reçu à tout moment. Le service de SMS est assez simple à utiliser et il est disponible sur tous les téléphones mobiles actuellement fabriqués. Selon (Fairon *et al.*, 2006), le SMS se distingue par son caractère assez ludique qui est spécialement apprécié par les jeunes qui montrent une créativité étonnante (création de mots, jeux de mots, émoticônes etc.).

Le SMS a le pouvoir de relier des utilisateurs du monde entier grâce à sa simplicité, son faible coût et sa grande rapidité. De nos jours, il est aussi largement utilisé par les entreprises dans le cadre de leur stratégie globale de communication.

2. https ://www.arcep.fr/fileadmin/reprise/observatoire/march-an2015/obs-marches-2015-def-201216.pdf

Problématique

Nous focalisons nos recherches sur la communication par SMS et le langage produit par cette communication. En effet, le langage SMS a assez vite suscité un intérêt scientifique. Cet intérêt a été initialement exprimé par de nombreuses études qui se basaient en général sur l'exploitation de petits corpus récoltés manuellement (Laursen, 2005, Hård Segerstad, 2005, Ling, 2005). En 2004, l'équipe de recherche du CENTAL [3] en Belgique a coordonné le projet *sms4science*. L'objectif du projet était la réalisation d'un corpus de grande taille pour la recherche. Pour y parvenir, l'equipe a défini une méthodologie pour la collecte des messages et des protocoles pour la préparation des corpus avant leur utilisation pour la recherche. L'étude du langage concerne une vaste gamme de disciplines telles que la linguistique, l'anthropologie, la psychologie, la sociologie, etc. Le but du projet était, dans une première phase, de fonder les bases pour un protocole de récolte de SMS authentiques, au travers desquels les chercheurs peuvent tirer des conclusions fiables et complètes sur le langage SMS.

L'utilisation des matériels authentiques permet, également, d'observer les particularités des graphies de SMS dans le but d'obtenir un point de vue plus objectif (Fairon *et al.*, 2006). Notre étude a comme point de départ le corpus de SMS, collectés et traités semi-automatiquement, dans le cadre du projet alpes4science. Le projet alpes4science fait partie du projet international sms4science et de son côté vise à contribuer à l'étude de ladite communication. La collecte s'est déroulée du 1er octobre 2010 au 31 janvier 2011 dans les Hautes-Alpes et l'Isère. Ainsi des informations variées qui sont liées à des questions démographiques et comportementales associées à l'utilisation de SMS sont disponibles dans la base de données alpes4science. Comme résultat ont été recueillis 22 054 SMS authentiques par 359 personnes dont 240 ont participé au questionnaire (96,7% de SMS). Dans la base de données nous trouvons les SMS (anonymisés, alignés, transcrits en langue standard, etc.), le lexique (mots SMS avec leurs traductions en langue standard et leurs fréquences) et des informations variées sur les expéditeurs (*cf.*

3. Centre de Traitement Automatique du Langage de l'Université catholique de Louvain, en Belgique

chapitre 3).

La plupart des messages courts présentent des différences significatives en comparaison avec les messages issus des textes formels, puisqu'ils doivent contenir au maximum 160 caractères et leurs auteurs utilisent diverses formes pour abréger les mots dans l'objectif de gagner du temps tout en réduisant l'effort fourni. Un des obstacles auquel nous devons faire face est la graphie particulière des mots SMS (fusion de mots, graphies abrégées imprévisibles, suppression de caractères, manque de ponctuation, etc.). Pour mieux observer ces particularités dans notre corpus de travail nous allons appliquer un test dans le but de mettre en évidence les phénomènes liés aux problèmes du traitement automatique du langage SMS. L'étiquetage morphosyntaxique constitue une étape fondamentale afin de pouvoir traiter davantage de données textuelles, comme dans la reconnaissance d'entités nommées, la traduction automatique, les systèmes de questions-réponses, l'extraction d'information etc.

Les approches liées au traitement automatique du langage appliquées à des textes standard atteignent de hauts niveaux de précision. Cependant, les résultats sont plus faibles lorsque l'étiquetage est appliqué à des textes courts contenant du bruit. Étape fondamentale du TAL est le processus du prétraitement d'un texte qui consiste à la normalisation d'un texte. Sproat *et al.* (2001) mentionne la nécessité d'appliquer le processus de normalisation avant tout autre traitement basique de TAL. En règle général, le prétraitement d'un texte consiste, selon Torres-Moreno (2011), à un découpage approprié d'un texte, une normalisation des mots et un filtrage adéquat de certains termes et symboles de ponctuation. En ce qui concerne les SMS, le processus de normalisation a comme objectif de convertir un texte informel dans un texte grammaticalement correct. Le prétraitement est une étape essentielle et non négligeable lorsque nous devons traiter des données mal orthographiées, fusionnées, avec des phrases agrammaticales, un encodage différent, etc. Ainsi le texte doit subir certaines modifications indispensables que nous appelons normalisation, un traitement approprié avant de procéder à une application du TAL. Par conséquent, l'absence d'une telle étape pourrait entraîner des résultats faux (Torres-Moreno, 2011). Plusieurs auteurs signalent la nécessité de la mise en place d'un processus de

normalisation qui faciliterait le traitement de SMS pour diverses taches issues du TAL (Han et Baldwin, 2011, Panckhurst, 2017, Lopez *et al.*, 2016, Tarrade, 2017, Beaufort *et al.*, 2010a, Yvon, 2010). Le questionnement central de notre étude est : Comment concevoir une méthode de normalisation capable de produire des résultats proches d'une version morphosyntaxique standard ?

Pour répondre à ce questionnement, notre recherche se base sur l'étude des particularités de la communication par SMS. Le travail de recherche se focalise sur l'analyse d'un corpus de SMS dans le but de mettre en exergue les phénomènes linguistiques qui peuvent constituer une entrave à leur traitement automatique. L'étiquetage morphosyntaxique se révèle une étape fondamentale afin de pouvoir traiter des données textuelles de type SMS. En effet, comme pour le traitement du langage standard, le rôle de l'étiquetage morphosyntaxique est d'une grande importance pour réaliser des tâches comme la reconnaissance d'entités nommées, la synthèse vocale, les systèmes de questions-réponses, les chatbots, etc.

Notre étude est centrée sur les méthodes probabilistes et celles fondées sur l'ingénierie des grammaires pour la normalisation morphosyntaxique des SMS. Notre hypothèse se fonde sur le fait que l'application d'une normalisation appliquée, à priori, sur les SMS permettra d'effectuer une analyse morphosyntaxique standard sans avoir à mettre en place un traitement spécifique.

L'extraction automatique d'informations contenues dans les SMS est la dernière étape de notre recherche. Plus précisément, nous nous intéressons à la tâche d'extraction d'entités nommées, notre hypothèse étant que la normalisation morphosyntaxique des SMS permettra d'améliorer les performances des méthodes traditionnelles pour l'identification d'entités nommées.

Plan de la thèse

Cette thèse commence avec le chapitre 2 qui est consacré à la présentation des notions nécessaires pour appréhender la communication par SMS. Il s'agit d'un chapitre introductif à la définition de la communication médiée par ordinateur, couplée avec certaines caractéristiques de cette communication.

Le chapitre 3 est consacré à la présentation du corpus de notre analyse, alpes4science, né de la collecte réalisée en 2010 dans la Région Rhône-Alpes au sein du laboratoire LIDILEM. La méthodologie établie pour la création de la base de données et les protocoles définis pour le traitement du corpus bruts sont exposés, aussi bien, la caractérisation de données du projet, que la formulation des conclusions reposant sur ces analyses. Notre investissement personnel porte sur la caractérisation de données du projet et la formulation des conclusions reposant sur ces analyses.

Le chapitre 4 présente les principales caractéristiques qui reposent sur le croisement de données socio-démographiques fournies par le questionnaire que les participants du projet ont complété lors de la collecte de messages. Les principaux éléments lexicométriques visant à définir le niveau de richesse lexicale sur l'ensemble du corpus et sur certaines partitions liées aux profils des participants sont également exposés. A la fin de ce chapitre, nous trouvons l'illustration de l'analyse typologique de ce corpus.

Le chapitre 5 traite les motivations de la nécessité de la normalisation des messages bruités avec l'exposition de travaux antérieurs et les problèmes liés à la tokenisation. L'architecture du modèle hybride proposé est divisée en deux : a) la représentation intermédiaire à travers des grammaires locales et b) la mise en place du système de traduction automatique qui est la partie centrale de ce chapitre. A la fin, du chapitre nous trouvons les résultats de l'évaluation du modèle hybride sur un échantillon du corpus.

Le chapitre 6 introduit la définition des entités nommées et expose les catégorisations que

nous trouvons dans l'état de l'art. Les différents problèmes de catégorisation figurant dans l'état de l'art de la reconnaissance des entités pour les textes courts sont présentés. Dans la suite de ce chapitre, nous trouvons les applications possibles de la tâche de reconnaissance d'entités nommées dans les SMS et la première typologie pour les entités nommées issues des SMS et des messages électroniques en général.

Le chapitre 7 est dédié à l'évaluation des performances des systèmes de reconnaissance d'entités nommées, fondés sur l'apprentissage automatique et l'ingénierie des grammaires, à partir d'une série de tests. L'objectif du chapitre est de montrer que les performances de ces systèmes présentent des améliorations significatives lorsqu'ils sont appliqués sur les SMS automatiquement normalisés.

Glossaire

Langue standard

Dans le cadre de notre recherche, nous appelons *langue standard* toute forme orthographiquement et grammaticalement correcte, écrite en français. La *langue standard* est associée à la *norme*. De son côté, la norme fait appel aux normes d'usage d'une langue qui suivent les codes orthographiques et grammaticaux. La norme correspond au bon usage de la langue, autrement dit, à ce qui est officiellement reconnu.

Transcription

La transcription est souvent utilisée pour décrire la représentation d'un alphabet, d'unités lexicales ou phoniques au moyen de signes, d'une écriture différents. Le terme est souvent utilisé pour décrire la représentation écrite de l'oral. Selon Mondada (2000) *la transcription exploite les ressources de l'écrit pour produire une intelligibilité de l'oral fondée sur des opérations de filtrage des « bruits » ou d'autres aspects jugés non significatifs, de discrétisation du continuum sonore, d'homogénéisation dans le cadre de conventions systématiques*. Le terme *transcription* est utilisé dans le cadre de cette thèse pour faire référence à la transcription manuelle effectuée par les annotateurs lors de la phase de transformation d'un message qui contient des abréviations, des phonétisations, extensions ou autres caractéristiques propres aux SMS en sa forme standard.

Normalisation

Nous utilisons le terme *normalisation* pour faire référence à la production automatique de la forme standard d'un SMS issue de notre système hybride. C'est le processus de transformation d'une forme langagière vers sa forme standard.

Les entités nommées

Les entités nommées se définissent comme des noms propres (noms des personnes, lieux, organisations, etc.). Meur *et al.* (2004) les entités nommées se définissent en tant que *types d'unités lexicales particuliers qui font référence à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques et qui ont un nom (typiquement un nom propre ou un acronyme)*. Les entités nommées issues du corpus alpes4science, évoquant des données personnelles à caractère direct ou indirect, ont été anonymisées (cf. 3). Cependant, toute entité nommée figurant dans les exemples du chapitre 7 de cette thèse, pour des raisons de confidentialité, a été remplacée par une autre entité nommée, aléatoirement choisie.

Chapitre 2

La communication par SMS

Sommaire

2.1 Introduction **23**

2.2 Définition de la communication virtuelle écrite **24**

2.2.1 Caractéristiques de la communication virtuelle écrite 26

2.3 Certaines caractéristiques de la communication par SMS **31**

2.3.1 Particularités et phénomènes graphiques 32

2.4 Conclusion **38**

2.1 Introduction

Au seuil du 21e siècle, la communication par SMS fait déjà partie du quotidien de beaucoup de gens. La communication électronique se réalise au moyen de dispositifs, tels que le téléphone portable et l'ordinateur à travers lesquels des messages sont échangés.

Ce chapitre vise à présenter les notions nécessaires pour appréhender la communication par SMS. Dans un premier temps nous donnons la définition de la communication virtuelle

tout en mettant l'accent sur les caractéristiques de la communication médiée par ordinateur.

2.2 Définition de la communication virtuelle écrite

La communication virtuelle est aussi appelée communication médiée par ordinateur, cybercommunication ou netspeak. Ces termes renvoient à des dispositifs numériques en tant que transmetteur/médiateur pour décrire les SMS, les messagerie instantanées, les jeux en lignes multi-joueurs, les courriers électroniques, etc. Ce terme implique l'utilisation d'un moyen électronique comme médium de communication. Cette activité passe souvent par le téléphone portable et/ou l'internet. La conversation peut être synchrone et asynchrone puisque l'échange peut s'effectuer en temps réel ou en mode différé. Marcoccia (2016) nous signale l'appartenance de la communication virtuelle écrite à deux autres types : la communication médiée par ordinateur (CMO) et la communication médiée par téléphone (CMT). Cette distinction se base sur les différentes propriétés que les ordinateurs et les téléphones partagent en tant qu'instruments de communication (clavier, taille d'écran, etc.). Cependant, avec l'apparition du téléphone intelligent, nous aurons tendance à dire que cette distinction est plutôt périmée puisque, de plus en plus, les deux dispositifs électroniques partagent les mêmes propriétés.

Le fondateur et éditeur du site web *CMC Magazine*, December (1996), définit la CMO comme un processus par lequel les gens créent, échangent et perçoivent des informations en utilisant des systèmes de télécommunications en réseau qui facilitent l'encodage, la transmission et le décodage des messages. La communication virtuelle écrite est sujet d'étude de plusieurs disciplines qui pourraient étudier ses aspects (psycholinguistique, sociologie, sociolinguistique, etc.). Les premières recherches réalisées au sujet de la communication médiée par ordinateur par Panckhurst (1998) mentionnent la division entre l'oral et l'écrit. Il s'agit d'une forme de discours avec plusieurs structures, une multitude de marqueurs et une certaine complexité.

Il existe une grande variété de termes utilisés pour designer la communication numérique. Le dénominateur commun de ces termes est la dimension électronique de la communication. Le terme *communication médiée*[1] *par ordinateur* (CMO) a été proposé par Panckhurst (1997) pour être adopté par plusieurs chercheurs. *Discours électronique médié* est le terme qui a été également proposé par la même auteure pour intégrer un contexte linguistique (Panckhurst, 1999, 2006, 2007). Marcoccia (2000a) préfère le terme *communication médiatisée par ordinateur*. D'autres ont choisi le terme *électronique*, comme Anis (2003) pour *communication électronique scripturale* et Anis *et al.* (2004) ou, tout simplement, *communication électronique*. Véronis et Guimier de Neef (2006) ont adopté le terme de *nouvelles formes de communication écrite*. Le terme le plus récent pour décrire ce type de communication a été proposé par Cougnon et François (2011), Cougnon (2015). Il s'agit du terme *communication écrite médiée par ordinateur* (Cémo) qui combine de la communication écrite et de la conversation. Le choix de ce terme n'est pas fait au hasard puisque, selon l'auteur, il semble plus précis et correspond bien à *la forme de communication dont la réalisation est écrite et dont le véhicule "influant" est l'ordinateur, accepté dans son sens le plus large* (Cougnon, 2015).

En anglais, nous trouvons les termes *electronic communication*, *computer-mediated communication* ou *discours*. Ces termes sont assez génériques, comme Marcoccia (2016) le mentionne, et ne permettent pas la distinction des différents types de communication (par exemple, écrit, téléphonique, vidéo-phonique etc.) et ramènent la communication au moyen d'un ordinateur à des limites plus étroites, laissant de côté les autres technologies. Les autres termes en anglais sont le *computer-mediated communication* (Herring, 1996), le *cyberlangage* (Dejond, 2006) et le *written interactive discourse Netspeak* (David, 2001).

La communication par SMS tient une place importante au sein de cet ensemble de CMO. Ces dernières années l'échange de SMS en tant que moyen de communication au travers des

1. Initialement, le terme *médiatisée* a été proposé à la place de *médiée*. Cependant, le verbe médier est plus approprié que le verbe médiatiser lorsqu'il s'agit de la communication au moyen d'un ordinateur. Effectivement, le verbe médiatiser, selon le dictionnaire le Petit Robert signifie *diffuser par les médias*.

téléphones portables a été largement répandu. Il s'agit d'un échange court en terme de caractères, puisque l'utilisateur a à sa possession uniquement 160 caractères par message. En effet, le téléphone portable constitue le support de cette communication. Comme Maingueneau (2012) l'avait mentionné : "*Le support n'est pas accessoire*". Cette phrase met en valeur la place que le *support* et le *transport* tiennent dans la communication verbale ou écrite. Effectivement, l'instrument via lequel le message est transmis modifierait l'*ensemble d'un genre de discours*. La communication par SMS a une fonction a) interactive et sociale qui vise à renforcer et à garder les liens entre les individus et une fonction b) instrumentale afin de faciliter la communication et la coordination de leurs actions (Tsakona, 2009).

2.2.1 Caractéristiques de la communication virtuelle écrite

Selon Romiszowski et Mason (1996), nous identifions quatre aspects de la CMO : a) la communication synchrone ou asynchrone, b) la communication hautement interactive, c) la communication orale ou écrite et d) la participation active ou passive.

2.2.1.1 La communication synchrone ou asynchrone

Dans la communication *asynchrone*, le moment où un expéditeur envoie un message diffère du moment où le destinataire lit le message. Plus précisément, lorsqu'un message est publié sur un forum en ligne il sera lu ultérieurement par un autre utilisateur à un autre moment donné. Des exemples de la communication asynchrone en ligne sont : le web, le e-mail, les forums de discussion, les blogs etc. Les avantages relatifs à cette communication sont la disponibilité du temps de réflexion et la flexibilité de la disposition du temps. Par contre, cette communication peut être lente en terme de temps.

En revanche, dans la communication synchrone, les échanges se réalisent en temps réel ; nous avons souvent du texte, de l'image, du son ou une combinaison de ces éléments et la

communication s'effectue facilement. Parmi les avantages de la communication synchrone nous trouvons les échanges en direct et le sens action/réaction. AbuSeileek et Qatawneh (2013) concluent que la communication asynchrone a) produit plus de types de questions que la communication synchrone, b) produit plus de stratégies de questions, c) prend en charge les questions à réponses longues, et d) favorise des questions demandant plus de détails.

2.2.1.2 La communication hautement interactive

La communication interactive reflète beaucoup de configurations différentes qui se dégagent de la connection entre ces caractéristiques de communication impersonnelle et de la communication de masse (Mahmoud et Auter, 2009). L'interactivité est la caractéristique centrale de cette communication, de nombreux chercheurs ont, par ailleurs, focalisé leurs études sur l'interactivité dans le but de définir les dimensions, spécificités et plus généralement le caractère de la communication interactive (Kiousis, 2002, Dessus *et al.*, 1997, Péraya, 1994, Mangenot, 2009, McMillan, 2002, Downes et McMillan, 2000, Kress et Van Leeuwen, 2001). Il s'avère que la définition de l'interactivité de la communication médiée par ordinateur est assez complexe. Dans le tableau 2.1 nous citons trois classifications de définitions. Kiousis (2002) se fonde sur le modèle de la communication tri-dimensionnelle avec les trois facteurs en tant que les principales dimensions de l'interactivité : la structure technologique du médium (la vitesse, la portée, la flexibilité temporelle, etc.), les caractéristiques du contexte de communication (dépendance de troisième-ordre, présence sociale) et la perception de l'utilisateur (la proximité, la vitesse de perception, etc). McMillan et Hwang (2002) et Mahmoud et Auter (2009) se basent sur l'interactivité qui se compose de quatre éléments principaux, les fonctionnalités ou support technologique, la perception, le processus et la combinaison d'approches (pour la définition de chaque élément cf. Mahmoud et Auter (2009)).

La communication se caractérise par des processus d'interaction complexes entre les participants puisqu'elle combine la rapidité de la communication écrite, simultanément, avec la

Kiousis (2002)	structure de la technologie (medium) contexte de la communication (setting) perception de l'utilisateur
McMillan et Hwang (2002)	processus perception fonctionnalités combinaison d'approches
Mahmoud et Auter (2009)	support technologique perception d'utilisateur processus combinaison d'approches

TABLE 2.1 – Trois classifications de définitions de la CMO interactive dans Mahmoud et Auter (2009)

dynamique de la communication orale. Les possibilités d'interaction et de rétroaction sont quasiment illimitées. L'interactivité est généralement un attribut de la conversation en face-à-face, mais elle prend aussi place dans des contextes de communication assistée par ordinateur. Il s'agit d'une communication bidirectionnelle. Lorsque l'expéditeur et le destinataire échangent des messages, nous parlons d'une communication bidirectionnelle ou multivoies qui est présente dès que les messages circulent bilatéralement (Schultz, 1999). La figure 2.1 illustre un schéma simple du modèle de la communication interactive bidirectionnelle entre deux utilisateurs qui transmettent et reçoivent le contenu d'un message.

2.2.1.3 La communication orale/écrite

La communication écrite numérique se trouve à la frontière de l'oral et de l'écrit. Plusieurs travaux ont été consacrés à la définition des caractéristiques de ce discours afin de souligner ses particularités (Cougnon et Ledegen, 2008, Anis, 2000, 1999, Fairon *et al.*, 2006, Stark, 2015). Les propriétés du discours de la communication virtuelle sont souvent comparées à celles de l'oral et de l'écrit. Nous parlons donc d'un caractère hybride entre l'écrit et l'oral. Cette

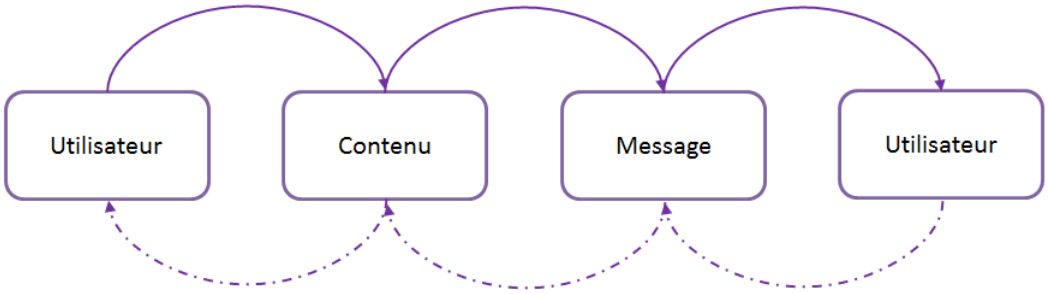

FIGURE 2.1 – Modèle de la communication interactive bidirectionnelle

hybridité provient du fait que cette communication utilise des procédés d'interaction en face à face et des procédés de la communication à distance. D'un point de vue aspectuel, il s'agit d'une communication écrite. Cependant, la communication est marquée par la spontanéité, ce qu'il lui donne cet aspect d'oralité.

Selon Maingueneau (2012), il faut noter que cette distinction n'est pas simple, elle est commode mais pas sommaire, puisqu'elle ne prend pas en compte la différence entre les écrits imprimés et manuels. En plus, elle est assez pauvre car *les techniques modernes de traitement des informations qui manipulent de manière presque "immatérielle" les sons, les lettres ou les images, les décomposent et les recomposent, les stockent et les projettent sur des écrans, puis, de là, éventuellement, sur un autre support électronique ou du papier.*

Gadet (1996) dans son analyse, autour de la distinction entre oral et écrit, souligne les trois caractéristiques de la production orale en comparaison avec la production écrite : a) l'*hétérogénéité* de l'oral (variation diastratique, diaphasique diatopique et diachronique) comparée à l'homogénéité de l'écrit qui est codifié, stabilisé et normé ; b) les *scories* (traces de son élaboration), l'oral est inondé de pauses, répétitions, hésitations, mots incomplets, interruptions etc. ; et c) l'intonation (prosodie). Panckhurst (2006) en se basant sur le constat de Gadet (1996), que l'oral *relève d'une complexité d'ordre grammatical* tandis que l'écrit d'une

complexité d'ordre lexical[2], prouve que sur une analyse linguistique de messages électroniques 75% étaient de formes nominales et moins de 10% de formes verbales. L'auteur mentionne également que sur un corpus de chat les formes verbales (32%) et nominales (44%) se rapprochent significativement.

Sur l'hybridité oral/écrit, du point de vue morphosyntaxique de la langue, la communication par SMS se positionnent Anis (1998),Collot et Belmore (1996),Yates (1996) afin d'explorer ces éléments qui contribuent à équilibrer oral et écrit. Cependant, Fairon *et al.* (2006) déclarent que la réponse à la question de l'oralité du langage SMS est effectivement, le fait que les SMS sont envahis par des phénomènes liés à la langue parlée, mais pourtant cela reste avant tout un langage écrit. De son côté, Cougnon (2015) croit à la neutralité de cette distinction.

Pour se positionner face à cette question, qui est l'objet d'étude de plusieurs chercheurs, nous considérons que dans la communication par SMS nous trouvons une pléthore de traits liés à la communication orale (hétérogénéité, scories, intonation). Cependant, comme Fairon *et al.* (2006) l'ont souligné, il s'agit, avant tout, d'un texte écrit extrêmement dépendant de son médium de communication. De ce fait, nous empruntons le terme plus approprié que Maingueneau (2012) utilise pour la distinction entre énoncés dépendants et indépendants de leur environnement. En effet, il s'agit d'un énoncé écrit de *style parlé* ; le terme **style parlé** est employé par l'auteur pour les énoncés qui présentent certaines caractéristiques d'un énoncé dépendant de l'environnement. Même si la transmission se réalise par un support graphique et la réception est différée, les SMS sont des textes écrits (support graphique) dont la réception peut être différée.

2. Ceci se traduit par la dominance de formes verbales à l'oral et de formes nominales à l'écrit.

2.2.1.4 La participation active ou passive

La communication active se définit par un échange synchrone ou asynchrone lorsque, par exemple, nous envoyons des SMS, des messages sur un forum de discussion ou en général quand nous participons à une discussion. En revanche, la participation passive se traduit par le silence de la part du participant pendant un échange.

Actuellement, l'évolution de la communication par SMS permet la réalisation des discussions entre plusieurs participants. Cependant, dans notre corpus d'étude ce paramètre est absent puisqu'il n'existe pas la notion de discussion.

La communication active et passive concerne majoritairement les discussions en forums. En effet, certains utilisateurs ont tendance à dominer les discussions tandis que la majorité se cache et évite de participer activement à la discussion. Souvent les participant des forums de discussion restent des destinataires passifs plutôt que des contributeurs actifs aux discussions. Ce phénomène est l'objet de beaucoup de recherches de la communication virtuelle (Romiszowski et Mason, 1996).

2.3 Certaines caractéristiques de la communication par SMS

L'essor de la communication et de la technologie ont contribué au grand développement de la communication écrite au moyen d'outil électroniques, tels que le téléphone mobile. Les SMS font partie de la CMO au moyen du téléphone mobile, qui en tant qu'instrument prend une place stratégique concernant la nature des productions issues de la communication par SMS.

Le *langage SMS* est le fruit de cette communication. Au sujet de ce terme, Panckhurst (1997) propose l'utilisation de la notion de l'*écrit* ou d'*écriture SMS*. Cougnon (2015), du même avis, soutient qu'il s'agit d'un nouveau code graphique, le *code SMS*, possédant de ses

propres caractéristiques avec un décodage de la langue. Elle définit ce code en tant qu'un *écrit SMS* qui utilise les caractères alphanumériques de la langue standard mais avec une valeur et pratique différente (phonétisation, abréviations, espaces etc.). Ces phénomènes sont typiques de la communication par SMS, toutefois, les exemples ont été attestés dans le corpus d'étude. Pour plus d'informations sur la typologie du langage SMS *cf.* 4.4.

2.3.1 Particularités et phénomènes graphiques

Les messages SMS combinent plusieurs procédés pour rendre un message compréhensible en utilisant le moins de caractères possible. Étant donné que le nombre limité de caractères par SMS peut entraîner une facturation supplémentaire pour transmettre un message au delà de 160 caractères, les utilisateurs devaient trouver un moyen d'être concis. De plus, la saisie d'un message au moyen d'un clavier azerty est plus lente qu'avec un clavier d'ordinateur. C'est pourquoi, la capitalisation des caractères, la ponctuation et la grammaire sont régulièrement dégradés. De nos jours, bien que les forfaits téléphoniques SMS soient illimités, les écrans sont encore petits et le besoin de transmettre rapidement un message est un problème qui persiste. De telle façon que le langage SMS est largement utilisé dans ses différentes formes qui combine plusieurs procédés pour raccourcir les phrases et les mots.

2.3.1.1 L'abréviation

L'abréviation est le fait de rendre plus court l'écriture de mots en omettant une partie des mots pour des raisons de concision et d'économie de l'espace. Il y a des abréviations universellement admises qui ne nécessitent pas d'explication. Dans la plupart de cas, des voyelles sont omises mais aussi les consonnes formant des voyelles nasales (par exemple, ct → cette, slt → salut). Il s'agit d'un phénomène graphique qui englobe diverses formes du système typologique comme les troncations, les abréviations phonétiques, les siglaisons et les acronymes.

Les abréviations sont des formes graphiques et lexicales qui ne sont pas essentiellement propres au langage SMS, puisqu'il s'agit d'un procédé qui existe depuis l'antiquité.

Les **troncations** se divisent en *apocopes* et *aphérèses*. L'apocope consiste à construire une nouvelle forme à partir de la suppression du suffixe d'un mot. Tandis que les aphérèses visent la construction d'une nouvelle forme par la suppression du préfixe d'un mot (*cf.* tableau 2.2).

Apocope		
ordi	→	ordinateur
aprem	→	après midi
cafet	→	cafeteria
dej	→	déjeuner
Aphérèse		
soir	→	bonsoir
net	→	internet
car	→	autocar

Table 2.2 – Exemples d'apocope et d'aphérèse

Les **sigles** et les **acronymes** sont le résultat d'un processus de création d'un mot à partir de chaque caractère initial des termes principaux d'une expression complexe. Comme pour les troncations les sigles ne concernent pas exclusivement le langage SMS. En effet, un sigle peut équivaloir à un nom de personne, d'une entreprise et même d'un pays ou ville, etc. Par exemple, NY → New York, DSK → Dominique Strauss-Kahn, IP → protocole internet, UGA → Université Grenoble Alpes. En ce qui concerne les SMS, nous trouvons des sigles qui appartiennent au le monde de la CMO issus d'une création graphique. Par exemple, lol → laughing out loud, mdr → mort de rire , pdr → pété de rire. Fairon *et al.* (2006) constatent que sur leur corpus d'étude les sigles figurent majoritairement en minuscules.

Les **squelettes consonantiques** consistent à construire une nouvelle unité graphique à partir de consonnes d'un mot en éliminant les voyelles. Par exemple, tkt → t'inquiète, jtm →

je t'aime, slt → salut, tjr → toujours. Sur ce sujet, Anis (2003) affirme que les consonnes en position faible dans les groupes consonantiques sont en général éliminées. Suivant l'affirmation d'Anis, Prochasson *et al.* (2009) propose de modéliser les squelettes consonantiques à partir des règles de transformations capables de a) conserver la première et de la dernière consonne ainsi que des voyelles situées avant la première et après la dernière (indépendance → in...ce) ; b) supprimer les voyelles restantes (indpndce) ; c) supprimer les consonnes l , r et h lorsqu'elles sont situées après une consonne en début de syllabe ; et d) supprimer les consonnes n et m lorsqu'elles sont situées avant une consonne en fin de syllabe (indpdce).

2.3.1.2 La phonétisation

La phonétisation des caractères est un phénomène graphique assez fréquemment observé dans des corpus de SMS et implique à la fois les lettres, les chiffres, les signes, et les graphies phonétiques plus proches (Cougnon, 2010, Cougnon et Beaufort, 2009). Appelée aussi rébus, il s'agit de séquences mêlant chiffres, lettres et signes qui s'interprètent grâce à leur valeur dénominative (Fairon *et al.*, 2006). Par exemple : à + → à plus, de grandes @→ de grandes oreilles

L'usage d'une lettre pour sa valeur phonétique est le procédé le plus fréquent dans la communication par SMS (Fairon *et al.*, 2006). Dans la même idée, le remplacement des caractère s'effectue par un chiffre qui a une valeur phonétique approximative. Avec l'utilisation du clavier azerty sur les téléphones portables le remplacement de caractères par des chiffres n'est plus un phénomène courant. L'utilisation des signes divers avec une valeur dénominative sert à la fois la rapidité et la facilité, mais elle apporte, aussi, une valeur stylistique au texte.

Dans cette catégorie nous incluons également l'**orthographe phonétique** que Fairon *et al.* (2006) et Cougnon (2010) ont introduit. Ce terme renvoie à la première orthographe qu'un enfant développe. En effet, un enfant apprend la correspondance entre les sons et les syllabes

Lettres :

koi → quoi, biz → bises, c → ce

Chiffres :

dem1 → demain, 2m1 → demain, 2mande → demande

Signes :

à + → à plus, grandes @ → grandes oreilles

de sa langue maternelle et écrit les mots selon ce qu'il entend. Par exemple, *bato* pour *bateau*. Dans le cas des SMS, il s'agit des suppressions de finales muettes et des simplifications d'une unité lexicale. Par exemple, bo → beau, fix → fixe, agreabl → agréable.

2.3.1.3 L'alternance codique

L'emprunt lexical correspond à la notion d'*alternance codique* (code switching en anglais) qui se réalise à l'intérieur d'un même échange verbal, de passages où le discours appartient à deux systèmes ou sous-systèmes grammaticaux différents (Gumperz, 1982). Le terme d'alternance codique concerne, notamment, selon la définition de Brasart (2013) : *l'usage fluide de deux langues ou plus au cours de la même conversation par un ou plusieurs locuteurs bilingues*. La linguiste spécialiste de l'alternance codique, Poplack (2004), la définit en tant que manifestation linguistique du contact et du mélange des langues, qui inclut des emprunts aux niveaux lexico-syntaxique, du transfert de la langue, de la convergence linguistique, etc.

Selon Pfaff (1979), le terme d'alternance codique devrait être considéré comme un hyponyme d'alternances *intra-phrastiques* ou de *mélange codique* (code-mixing en anglais) [3], phénomènes auxquels appartient l'emprunt linguistique. L'*emprunt* correspond à un code lexical aux

3. Dans Brasart (2013), selon l'article *Government and code-switching* de Di Sciullo *et al.* (1986) *les auteurs font la distinction entre alternance, phénomène dans lequel les deux langues restent séparées, et mélange, phénomène dans lequel des éléments lexicaux et des traits grammaticaux des deux langues apparaissent dans la même phrase.*

structures morphologiques, syntaxiques et phonologiques d'une autre langue (Doehler, 2013).

L'*alternance* et l'*insertion* (auparavant transfert) sont deux autres types d'alternance codique. L'alternance est le passage d'une langue à l'autre dans un énoncé tout en impliquant la grammaire et le lexique (Muysken, 2000). L'insertion désigne le transfert qui est défini par (Auer, 1984) comme l'alternance linguistique pour une certaine unité avec un point de retour structurellement fourni dans la première langue. Ce phénomène correspond au mélange d'éléments et non seulement à une simple juxtaposition.

Bien que l'alternance codique concerne, notamment, le bilinguisme, il s'agit d'un phénomène que nous trouvons assez régulièrement dans la communication par SMS, même si les utilisateurs ne partagent pas les même codes linguistiques bilingues. Morel et Doehler (2013) mentionnent que l'alternance codique au travers d'un téléphone portable est un moyen par laquelle *les participants s'affichent comme membres d'une communauté à frontières certes floues mais à aspiration internationalisée, voire globalisée, orientée vers un univers branché – pour ne pas dire à un monde 'connected'*. Plus précisément, Doehler (2013) a remarqué que le changement de code est le plus souvent constitué d'insertions d'éléments uniques ou de combinaisons d'éléments dans un message composé dans une autre langue, et implique typiquement (seulement) une gamme limitée d'expressions routinisées. Morel (2016) utilise le terme *bricolage plurilingue écrit* afin de décrire les pratiques plurilingues qui mettent l'accent sur la participation, la présence mutuelle et l'affiliation sociale à une communauté donnée ou un espace d'affinité. L'alternance codique met en avant a) l'expression d'actions qui servent à l'ouverture et à la clôture des messages, les remerciements (hello, hola, ciao, thanks) ; b) le caractère expressif des messages, et par extension l'expression de l'affection (plz, hey, wow, love, amore) ; et c) les traductions brèves des formes en français qui pourraient être contraignantes à l'écriture (today, news).

Exemples d'alternance codique en SMS :

Non juste pour **news** comme ça. + tard alors. Biz

Let's skype ! Si tu as fini ton BBQ ? Suis rentree de NY ;-)

Chala vais au dodo mon coeur... A tout a l'heure **mi amore** !

Home sweet home. Enfants ravis. Gros bisous

Oui pas de pb pas pret de dormir de tte façon :(bon **flight**

2.3.1.4 Les émoticônes

L'émoticône a été largement utilisée au cours des 25 dernières années. Il s'agit des symboles, que nous trouvons dans la communication électronique, utilisés dans les courriels, les forums de discussions, la messagerie instantanée, les SMS, etc. A la base les émoticônes, en tant que signes graphiques sont utilisés comme des représentations iconiques des émotions. Selon la définition de Marcoccia (2000b), l'émoticône est un signe codé qui *transmet des informations sur la dimension relationnelle et émotionnelle de l'échange initié par l'émetteur*. Il s'agit des pictogrammes qui *combinent des signes de ponctuation et des caractères d'imprimerie, représentant de manière schématique des mimiques faciales comme des sourires, des clins d'oeil, des moues de colère ou de tristesse.*

Selon Halté (2013), il s'agit des icônes de mimiques faciales dont l'objectif est d'indiquer une émotion. Cependant, le terme d'émoticônes est appliqué par Halté pour désigner *uniquement des icônes dont la fonction est d'être l'indice d'une émotion ou d'une attitude subjective portant sur l'énonciation d'un contenu.*

Les émoticônes sont utilisées pour attribuer un sentiment à un texte, puisqu'ils ont le pouvoir à travers des symboles de représenter une pléthore d'émotions telles que l'amour, la tristesse, la frustration, la colère, etc. En effet, l'émoticône peut prendre une place verbale et non-verbale selon sa position. Selon Dresner et Herring (2010) , l'émoticône remplit trois fonctions : a) le *support* quand il y a une relation de support entre le message et l'émoticône (par

exemple, *:(je suis triste*) , b) le *supplément* dans les cas où l'émoticône sert à désambiguïser le sens d'une phase, par exemple, *Il va partir ! :)* , et c) l'*antiphase* quand l'émoticône contredit ou annule le sens d'une phrase, leur fonction est ironique et sarcastique, par exemple, *Je me sens bien :(*.

2.4 Conclusion

Les grands développements techniques des années 90 ont été marqués par le rapprochement du domaine de l'informatique de celui des télécommunications. Ce dernier a vu naitre une nouvelle forme de communication, le SMS (Short Message Service), qui a été notamment décrite grâce à l'apparition de termes tels que : communication médiatisée ou médiée par ordinateur, communication écrite médiée par ordinateur (Panckhurst, 1997), cybercommunication, netspeak, etc. L'intérêt d'étudier la communication par SMS réside dans les particularités que nous observons dans ce langage.

Dans ce chapitre introductif de notre thèse nous avons pu présenter la définition et certaines caractéristiques de la communication médiée par ordinateur. Il nous a permis de établir les bases pour la suite de notre recherche et de passer à la présentation et l'étude du corpus de SMS alpes4science qui va constituer le point de départ de notre étude.

Chapitre 3

Le projet alpes4science

Sommaire

3.1 Introduction **39**

3.2 Alpes4science **40**

3.2.1 Méthodologie de recueil des données 41

3.2.2 L'enregistrement des données 42

3.2.3 Traitement automatique du corpus 45

3.3 Les données récoltées **53**

3.3.1 Le questionnaire 53

3.3.2 Le corpus SMS 55

3.4 Conclusion **58**

3.1 Introduction

Ce chapitre présente le corpus de notre analyse *alpes4science* né de la collecte réalisée en 2009 dans la Région Rhône-Alpes. Nous préciserons la méthodologie détaillée employée afin d'obtenir des données les plus authentiques. Par la suite, nous présenterons la méthodologie

établie pour la création de la base de données, aussi bien que les protocoles définis pour le traitement du corpus bruts. Notre investissement personnel porte sur la caractérisation de données du projet et la formulation des conclusions reposant sur ces analyses. Plus précisément, nous détaillerons certaines données quantitatives, les différents traitements réalisés afin de conclure sur les données récoltées, avant la réalisation de cette thèse.

3.2 Alpes4science

En terme général, la définition d'un corpus repose sur un recueil formé d'un ensemble de données sélectionnées et rassemblées (Mellet, 2002). Le corpus a comme objectif de constituer la référence pour diverses études et vise à devenir le reflet représentatif des phénomènes à analyser (Condamines, 2005). Un corpus a comme objectif de rassembler les matériaux pour répondre à une question de recherche.

Afin de pouvoir observer de près le langage SMS et travailler sur un corpus spécialisé et authentique dans le but d'avoir un point de vue plus objectif sur notre recherche, nous avons fait appel au corpus de SMS issu du projet alpes4science qui constitue le point de départ de notre recherche. Ce projet fait partie du projet international sms4science [1], un projet initié et coordonné par le Centre de Traitement Automatique du Langage (CENTAL) de l'université Catholique de Louvain en Belgique, ayant comme objectif la construction et l'étude d'un corpus international avec l'aide d'une équipe de chercheurs chargée de définir l'affinité des différents projets nationaux, afin de créer de corpus comparables dans les différentes langues et variantes régionales.

1. http ://www.sms4science.org/ ?q=en

3.2.1 Méthodologie de recueil des données

Le projet alpes4science a été signé en 2009 entre le laboratoire de Linguistique et Didactique des Langues Étrangères et Maternelle (LIDILEM) de l'université Grenoble Alpes et le Conseil Général des Hautes-Alpes dans le but de créer une base de données. La collecte s'est déroulée du 1er octobre 2010 au 31 janvier 2011 dans les Hautes-Alpes et l'Isère.

Pour participer au recueil de données les utilisateurs devaient envoyer un premier message sur le numéro mutualisé *31014*, numéro non surtaxé mais facturé au prix d'un message selon le forfait du participant, tout en ajoutant au début de chaque message le code *SMS05* (figure 3.1).

FIGURE 3.1 – Affiche du projet alpes4science (http ://www.sms4science.org/ ?q=en)

Par la suite, les messages étaient transmis à une plateforme dédiée au projet, fournie par le partenaire opérateur de télécommunications, Orange [2]. Cette démarche a permis de référencer chaque participant qui pouvait alors transférer tous les messages dont lui-même était l'auteur et qui étaient destinés à d'autres correspondants. Par la suite, un message de confirmation était automatiquement envoyé à l'émetteur en l'invitant à compléter un questionnaire optionnel

2. www.orange.fr/

sur le site du projet. Afin d'encourager et récompenser les participants plusieurs cadeaux (appareils photo numérique, ipod Shuffle, clés USB, places de cinéma, etc.) ont été offerts pendant la période de collecte par les partenaires du projet, le Conseil Général des Hautes-Alpes, l'université Stendhal, Orange et les cinémas Pathé-Gaumont de Grenoble. Les réseaux sociaux Facebook [3] et Twitter [4] ont joué, également, un rôle stratégique dans le but d'inciter les gens à participer et d'annoncer les dates et lots des tirages au sort des cadeaux, l'avancement du projet et les divers avancement du projets comme par exemple le passage du 10 000ème ou 20 000ème SMS collecté (Chabert *et al.*, 2012).

3.2.2 L'enregistrement des données

A l'aide de la plateforme *Contact Everyone* mis en place par l'opérateur Orange pour les besoins du projet, nous avons pu d'un coté recevoir les SMS envoyés sur le numéro mutualisé et de l'autre coté transmettre des messages aux participants. La plateforme a constitué le canal de communication entre les expéditeurs et les récepteurs des messages (trois chercheurs stagiaires du LIDILEM). Chaque fois qu'un participant envoyait un SMS les trois chercheurs recevaient de leur coté un mail du type : *Vous avez reçu un sms du 272453904343920 sur le 31014*. Grâce à cette procédure le numéro de l'expéditeur était crypté, pour ce cas là le chiffre *272453904343920* correspond au numéro de téléphone crypté afin de respecter le code déontologique.

A la fin, les SMS reçus par mail ont été extraits au travers d'un script php pour être par la suite stockés dans une base de données MySQL (cf. figure 3.2). Dans cette base de données nous trouvons six tables :

1. Les **SMS** avec les données liées aux messages reçus :
   — l'identifiant ;

3. https ://www.facebook.com/pages/SmsAlpins/129178067131807 ?ref=ts
4. https ://twitter.com/smsAlpins

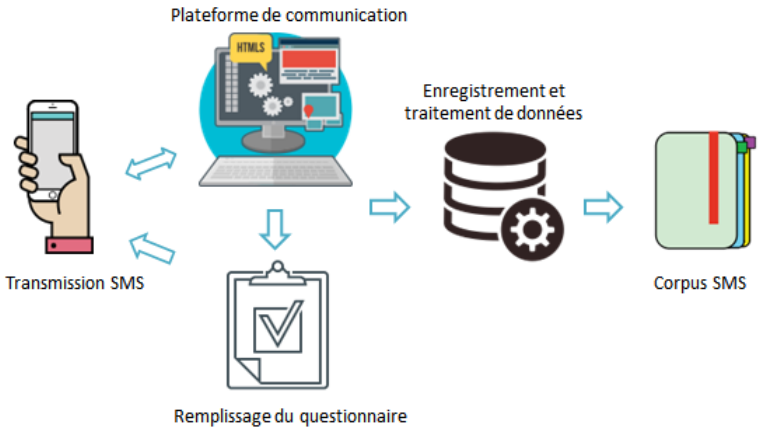

FIGURE 3.2 – Méthodologie de collecte

— la date de réception ;
— le texte original ;
— le nombre de caractères ;
— le texte anonymisé ;
— le texte découpé à transcrire ;
— le texte transcris ;
— le texte aligné ;
— les commentaires de transcripteur ;
— le nom du transcripteur.

2. L'**expéditeur** contenant les informations du questionnaire rempli par les expéditeurs :
   — l'identifiant ;
   — le numéro de téléphone ;
   — l'age ;
   — le sexe ;
   — la langue maternelle ;
   — la langue régionale ;
   — l'utilisation du langage SMS ;

— la profession ;
— le niveau d'études ;
— le département actuel ;
— le département d'origine ;
— le nombre de SMS envoyés par semaine ;
— le type de clavier utilisé ;
— les commentaires ;
— la fréquence d'usage de SMS.

3. Les **données personnelles** qui correspondent aux données anonymisées dans les SMS bruts :
   — l'identifiant de la donnée ;
   — le code de la donnée à anonymiser (CB, URL, MAIL, ADR, etc.) ;
   — l'explication du code (CB→carte bancaire, ADR→ adresse postale).
4. Le **dictionnaire** qui contient les entrées du dictionnaire du français utilisé pour la transcription des SMS :
   — l'identifiant de l'entrée du dictionnaire ;
   — la forme fléchie ;
   — la forme canonique.
5. Le **lexique** qui contient tous les mots SMS qui ont été transcrits :
   — l'identifiant de l'entrée SMS ;
   — le terme à traduire ;
   — le terme traduit ;
   — la fréquence du terme dans le corpus.
6. La table **prénom** qui contient les prénoms qui ont été anonymisés :
   — l'identifiant du prénom ;
   — le prénom original ;
   — le code qui a remplacé le prénom.

3.2.2.1 Les problèmes rencontrés

Selon Antoniadis *et al.* (2011) et Chabert *et al.* (2012) lors de la gestion de SMS certains problèmes inhérents à la plateforme utilisée ont été rencontrés. Plus précisément, la plateforme utilisée n'était pas adaptée à l'objectif du projet de constitution d'un corpus de SMS. Certains messages avaient été reçus et stockés sur la plateforme mais n'avaient pas été transmis par mail, les messages n'étaient pas forcément gratuits pour tous les détenteurs d'un forfait comportant l'envoi de SMS illimités malgré l'indication initiale *non surtaxé* et plusieurs messages longs ont été tronqués à 150-160 caractères.

3.2.3 Traitement automatique du corpus

Avec la construction du corpus de SMS nous pouvons examiner de façon adéquate le fonctionnement du langage de SMS et explorer exhaustivement des productions langagières authentiques. Deux traitements essentiels ont été effectués pour rendre le corpus de SMS opérationnel et les applications liées au TAL efficaces : l'*anonymisation* de données sensibles pour garantir la confidentialité des informations transmises et la *transcription* qui rend les SMS plus lisibles mais aussi pour pouvoir étudier le langage SMS plus facilement par la suite.

3.2.3.1 Anonymisation

Du fait que, tout traitement de données personnelles à caractère direct ou indirect doit être déclaré à la Commission nationale de l'informatique et des libertés (CNIL) une demande a été déposée. Il était donc indispensable de cacher toutes les données personnelles présentes dans les SMS. Le protocole suivi pour cette tache reste fidèle aux principes du projet international *sms4science*.

3.2.3.2 Méthode d'anonymisation

Grâce à cette interface web (figure 3.3) conçue pour les besoins de cette tâche l'annotateur pouvait avoir accès sur l'ensemble de SMS bruts et choisir entre *Déjà anonyme !* dans les cas où aucune anonymisation s'avérait nécessaire et *Choisir !* pour choisir le SMS qui contenait une entrée à anonymiser.

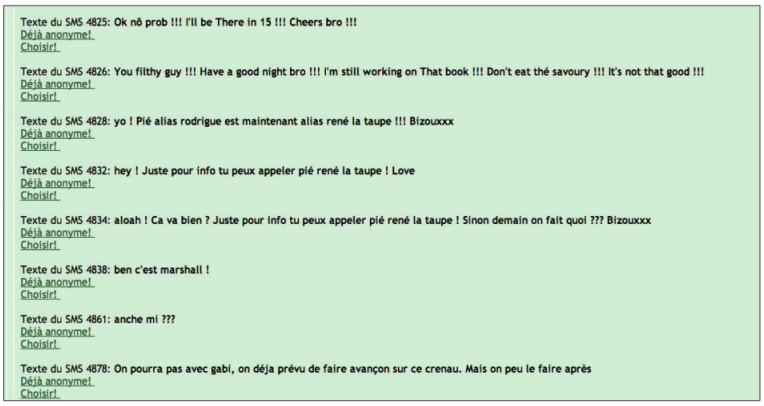

FIGURE 3.3 – Interface d'anonymisation, figure tirée de Chabert *et al.* (2012)

Dans ce dernier cas l'annotateur se dirigeait vers une autre page (figure 3.4). Certaines données comme les adresses e-mail, des sites Internet personnels et des numéro de téléphone qui ont un format fixe pouvaient être détectés automatiquement. Dans la figure 3.4 nous constatons qu'un numéro de téléphone a été détecté (en rouge), l'annotateur avait le choix entre accepter ou refuser l'anonymisation de la donnée et par la suite choisir d'anonymiser d'autres entrées personnelles potentiellement présentes sur le SMS. Dans ce cas là, la donnée à anonymiser devait être entourée par le symbole # puis choisir dans la liste déroulante, le type de donnée détectée. Si l'annotateur ne trouvait pas un type de donnée approprié à son élément, il pourrait définir un nouveau type en choisissant *Autre type de donnée* dans la liste déroulante.

Après l'identification manuelle ou automatique l'entrée détectée devrait être remplacée par

FIGURE 3.4 – Interface d'anonymisation, figure tirée de Chabert *et al.* (2012)

un code spécifique. La méthode étant assez simple consistait à remplacer l'entrée par ****(NOM DE L'ENTRÉE)_ Nombre de caractères de l'entrée**** (vf. exemple d'anonymisation 3.1).

SMS brut	Vlà le num de mon père 0672******.
SMS anonymisé	Vlà le num de mon père ***TEL_10***.
Explications	***TEL_10*** : numéro de téléphone de dix chiffres sans espaces

TABLE 3.1 – Exemple de données anonymisées

Les données anonymisées sont les suivantes :

- **Noms, prénoms et surnoms**

  Pour identifier les noms propres une base de données alimentait la plateforme. Étant donnée que certains noms et un grand nombre de diminutifs étaient présents dans les SMS, l'annotateur avait la possibilité d'enrichir la base en ajoutant de nouvelles entrées grâce à une table de correspondance qui a été créée afin de transformer le prénom d'origine en un code : *PRENOM_ x_ y*, où x correspond à un identifiant du prénom d'origine et y au nombre de lettres que le nom contient. De cette manière nous pouvons retrouver un prénom d'origine dans un SMS à partir de cette table de correspondance et ne pas perdre l'information des prénoms. Par exemple : ****SURNOM_ 2*** est cool!!" T'es bien vue par ***PRENOM_ 18_ 5*** toi :)* où ****PRENOM_ 18_ 5****

Codes	**Données anonymisées**
NOM	Nom propre
BLOG	Adresse de blog
CB	Coordonnées bancaires
URL	Adresse de site Internet personnel
MAIL	Adresse e-mail
ADR	Adresse d'habitation personnelle
RECHARGE	Numéros de recharge mobile
TEL	Numéro de téléphone
COM	Numéro de communication

TABLE 3.2 – Tableau récapitulatif de codes

correspond au prénom *Lucie*. Nous constatons également que pour ****SURNOM_2****, les surnoms et les diminutifs, ne font pas partie de cette liste.

- **Adresses postales**

  Il concerne les adresses de domicile personnel et non les lieux publics. Le format peut être assez varié et ne permet pas toujours l'identification automatique. Par exemple l'adresse : 12 r feydeau paris = ***ADR_22***.

- **Adresses e-mail**

  Les adresses mails sont facilement identifiables car elles se composent d'un format standard, cependant, une vérification était indispensable dans le cas où le symbole @ était remplacé des fois par *a+* ou *@+*.

- **Sites Internet personnels**

  Même si l'identification était facile à effectuer l'annotateur devait juger si les sites étaient *publics* ou *privés*. En effet, les sites web comme des blogs sont considérés comme des sites privés.

- **Numéros de téléphone**

  Il s'agit de numéros de téléphones fixes et mobiles, avec ou sans l'indicatif international en préfixe, après identification automatique et vérification par l'annotateur (table 3.1).

- **Digicodes, mots de passe, codes de carte bancaire**

  La détection des codes n'étaient pas une tâche facile car parmi les mots SMS nous trouvons des unités qui combinent chiffres et lettres (2m1, dem1, etc.) comme pour les digicodes et les mots passe (identification automatique et vérification par l'annotateur).

3.2.3.3 Transcription

La transcription de SMS vise à rendre un message qui contient des abréviations, des phonétisations, extensions ou autres caractéristiques propres aux SMS compréhensible par tous. Avant de passer à la transcription de SMS, sa définition de manière stricte par un protocole a été établie pour tous les éléments qui devaient être modifiés dans le message d'origine. L'objectif de cette démarche repose sur la définition de façon stricte des éléments à modifier dans les SMS bruts. Pour que la transcription soit réalisée de façon identique par les transcripteurs, la modification du SMS brut devrait être minime et réalisée lorsqu'elle s'avérait indispensable.

3.2.3.4 Méthode de transcription

La transcription a été réalisée à l'aide d'une interface qui proposait un découpage mot à mot se basant sur les ponctuations et les espaces. Par la suite, le transcripteur devait valider ou modifier ce découpage. Les transcriptions se basent sur les transcriptions préalablement proposées pour un même terme dans un nouveau SMS. De cette façon une liste des mots SMS avec leur équivalence est créée pour alimenter la plateforme (Antoniadis *et al.*, 2011, Chabert *et al.*, 2012). Le découpage au niveau des mots et la transcription de chaque terme individuellement (table 3.3) permet de créer un lexique SMS ➛ français standard et français standard ➛ SMS.

Le résultat de ce découpage (table 3.3) est la synthèse d'un corpus aligné au niveau des mots tel que nous le constatons dans l'exemple :

SMS découpé à transcrire	#Ok# #à# #tt# #de# #suite# ##!#
SMS découpé transcrit	#Ok# #à# #tout# #de# #suite# #!#

TABLE 3.3 – Exemple de découpage de SMS

<tok> <v lang="fr">Ok</v> <lang="sms">Ok</v> </tok>

<punct value=" "/>

<tok> <v lang="fr">à</v> <lang="sms">à</v> </tok>

<punct value=" "/>

<tok> <v lang="fr">tout</v> <lang="sms">tt</v> </tok>

<punct value=" "/>

<tok> <v lang="fr">de</v> <lang="sms">de</v> </tok>

<punct value=" "/>

<tok> <v lang="fr">suite</v> <lang="sms">suite</v> </tok>

<punct value=" "/>

<punct value="!"/>

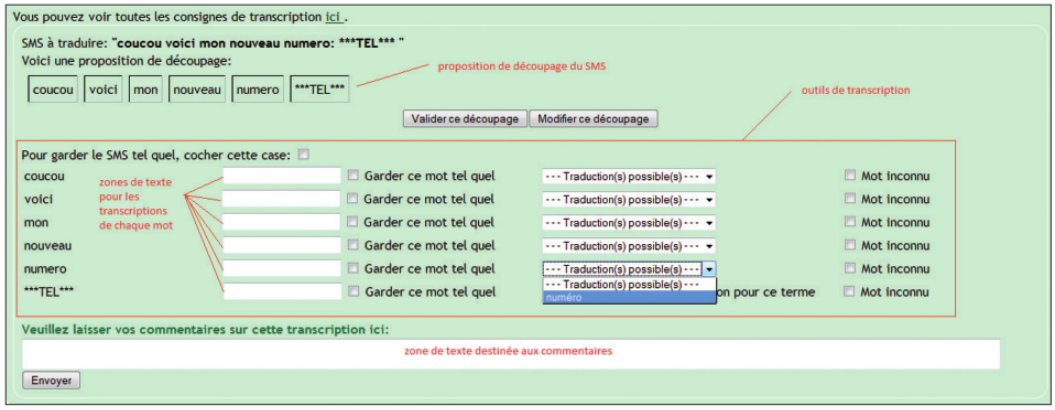

FIGURE 3.5 – Interface de transcription, figure tirée de Chabert *et al.* (2012)

Dans la figure 3.5 l'exemple d'un SMS à transcrire. Dans une première phase, nous trouvons le SMS brut avec une proposition de segmentation des mots du message. Par la suite,

l'utilisateur a le choix d'accepter ou de modifier ce découpage. Puis, si l'utilisateur souhaite garder le SMS ou un mot segmenté tel quel, il suffit de cocher la case, sinon il peut trouver dans la liste déroulante une transcription possible ou introduire une nouvelle transcription. Il a aussi la possibilité de laisser un commentaire sur la transcription réalisée. Exemple de transcription :

Corpus SMS brut	Ok. Jvou tien o ju, jdoi bosser un pe avant. A tte. Bizbiz
Corpus SMS transcrits	Ok. Je vous tiens au jus, je dois bosser un peu avant. A tout à l'heure. Bise

Table 3.4 – Exemple de transcription

Voici les règles autour de la transcription :

- **L'orthographe**

  Lors de la transcription la bonne orthographe du mot est rétablie selon le dictionnaire (cf. *Petit Robert*). La correction orthographique concerne, également, l'accord et le genre des éléments du discours.

  Exemple : *Bon alors ta passée une bonne jurnée ? ! => Bon alors tu as passé une bonne journée ? !*

  Dans le cas où un adjectif est abrégé, sa forme au masculin est consignée par défaut, si aucun élément du contexte ne permet de déterminer le genre.

- **Les emprunts**

  La présence de mots étrangers présents dans les SMS, sont conservés tels quels et sont traités comme les mots français.

- **Les néologismes**

  Les néologismes et les sociolectes sont transcrits dans la mesure du possible.

- **La ponctuation**

  Aucune intervention n'est survenue à l'exception des traits d'union et apostrophes.

- **Les symboles spéciaux**

  Lorsqu'un symbole spécial remplace un mot, il est subsitué par son équivalent en lettres.

Exemples : *à + => à plus*, *toi & moi => toi et moi.*

- **Les formes abrégées**

  La règle consiste à rétablir chaque forme, si l'abréviation figure dans le dictionnaire. Dans certains cas les formes abrégées peuvent être un réel défi lors de la transcription car elles sont souvent ambiguës. Si le contexte ne permet pas de déterminer la transcription, la forme abrégée est maintenue à l'identique.

  Exemple : *2kg farine, levure 2 boulanger, citrons, oeufs, sucre, sucre glace, beur, dattes : tou* ***bio*** *:-P biz => 2kg [/de /]farine, levure de boulanger, citrons, oeufs, sucre, sucre glace, beurre, dattes : tout biologique :-P bisous.*

- **Les émoticônes**

  Tous les émoticônes restent à l'identique, un commentaire est ajouté pour expliquer la signification. Par exemple pour *:** le commentaire *bisous* aurait été ajouté.

- **Les formes consécutives**

  Il s'agit de supprimer les lettres répétées afin d'obtenir le mot normalisé. Cependant, pour les formes onomatopéiques ou les mots abrégés seuls, trois lettres sont conservées pour marquer le sens de l'onomatopée.

  Exemples : *haaaaa ! => haa !*, *haaaaaate => hâte.*

- **Les acronymes**

  Les acronymes sont transcrits lorsqu'ils ne sont pas présents dans le dictionnaire. Les acronymes et signes usuels tels que DVD, SMS, GSM, etc. seront transformés en majuscules.

  Exemples : *jtm => je t'aime*, *dvd => DVD.*

- **Les éléments manquants**

  Dans le cas où un élément dans le SMS est manquant, le mot est introduit entre crochets lors de la transcription.

  Exemples : *Ben suis dans mon labo. => Ben [/je/] suis dans mon laboratoire*, *Je sais pas => Je [/ne/] sais pas.*

3.3 Les données récoltées

La collecte des SMS a fourni le recueil de 22 054 SMS authentiques par 359 personnes dont 240 ont participé au questionnaire. Pour résumer, dans la base de données nous trouvons les SMS (anonymisés, alignés, transcrits en langue standard et découpés), le lexique (mots SMS avec leurs traductions en langue standard et leurs fréquences) et des informations variées sur les expéditeurs (âge, sexe, niveau d'études, langue maternelle, profession, les données anonymisées, etc.).

3.3.1 Le questionnaire

De Singly (2012) expose l'utilité du questionnaire comme la production d'une excellente méthode pour l'observation et l'explication du comportement. En reprenant la fameuse phrase de Bourdieu *et al.* (2005) sur la définition du questionnaire, une enquête par questionnaire cherche à donner raison à *"ce que les acteurs font par ce qu'ils sont, et non pas ce qu'ils disent de ce qu'ils font"* autrement dit mettre la lumière sur la distinction entre ce que les acteurs sont et ce que les acteurs font. De son côté De Singly, 2016, complète cette analyse en soulignant que l'enquête associe le profil social des personnes interrogées (origine sociale, position sociale, diplôme, situation familiale) à leurs activités pour définir un rapport de la cause à l'effet.

A partir du questionnaire que les participants ont rempli en ligne nous disposons des informations variées qui sont liées à des questions démographiques et comportementales associées à l'utilisation de SMS dans la base de données alpes4science (cf. annexe B). Ces données composent un matériau incontournable pour la réalisation de recherches sociologiques, sociolinguistiques et d'autres comparatives sur observation de l'usage des SMS.

Le questionnaire contenait dix-sept questions et était optionnel, visant à être court et à recevoir le plus grand nombre de messages, mais obligatoire pour les participants qui souhai-

taient participer aux concours et gagner des lots. Nous constatons que 96,4% des SMS reçus correspond à des messages provenant de 240 participants qui ont rempli le formulaire. Il est nécessaire sur ce point de mentionner que toutes les précautions ont été prises pour des raisons éthiques afin de garantir l'anonymat des participants, seule information sensible, le numéro de téléphone a été demandé de façon optionnelle. Le numéro de téléphone a permis d'associer les SMS au profil du participant, une fois le questionnaire reçu un identifiant remplaçait le numéro.

En se basant sur un tronc commun de questions proposé de façon flexible et libre de la part du projet international sms4science en 2004, le questionnaire a suivi les deux grands axes de questions qui distinguent les profils et les pratiques, selon (Cougnon, 2015) :

— **Questions liées aux pratiques du participant** :
fréquence d'utilisation de SMS, fréquence de réception de SMS, abonnement illimité SMS, période d'utilisation de SMS, classification de fréquence d'envoi de SMS par destinataire, classification de fréquence de réception de SMS par expéditeur, type de clavier, langue utilisée, situation d'utilisation de SMS, changement de registre selon destinataire, utilisation de modes/systèmes de communications sur Internet.

— **Questions liées au profil du participant** :
Âge, sexe, numéro de téléphone, langue maternelle, langue régionale, niveau d'études, domaine de travail, code postal, département/pays d'origine.

Les questions du questionnaire étaient *libres* dans le sens où le participant avait la flexibilité de répondre ou pas, *fermées* avec une seule réponse à choisir parmi une liste de réponses proposées, *ouvertes* car aucune modalité de réponse n'est proposée et l'utilisateur pouvait écrire un texte libre en guise de réponse et d'échelle d'attitude (*échelle de Likert*) correspondant à une échelle qui comprend 4 à 7 degrés par laquelle l'individu exprime son degré d'accord ou

de désaccord relatif à une affirmation [5].

3.3.2 Le corpus SMS

En reprenant la définition du corpus linguistique par Sinclair (1994), Baker (1998), McEnery et Wilson (2001) le corpus est une collection de morceaux du langage lisibles et compréhensibles par la machine, sélectionnés et ordonnés en accord avec des critères linguistiques explicites dans le but d'être utilisés comme un échantillon du langage. Quant à la question de la place des SMS dans la linguistique de corpus Cougnon (2015) nous expose, dans sa thèse *Langage et sms*, 8 arguments affirmatifs sur l'étude du langage développée par la linguistique du corpus selon l'adaptation de Laviosa (2002). C'est en observant les particularités de ce langage en comparaison avec la langue générale que nous retrouvons la notion de spécialité telle qu'elle est définie par Condamines (2006). Un corpus SMS est un corpus spécialisé car il est centré sur un vocabulaire particulier, sur un certain type de textes, sur le langage des membres d'un groupe social (Bowker et Pearson, 2002).

A l'issue de la récolte, 22 054 SMS authentiques ont été enregistrés dans la base de données. Cependant, en réalité le nombre total de messages réellement retenus pour la composition du corpus est passé à 21 261 SMS après l'élimination de SMS :

— identiques expédiés depuis le même numéro figurant plusieurs fois (doublons), un seul exemplaire de chacun a été retenu ;
— contentant exclusivement des chiffres, dans la plupart de cas les messages contentaient des numéros de téléphones ;
— rédigés en une autre langue que la langue française, en effet, plusieurs SMS écrits en anglais, espagnol, allemand, italien, etc. ont été envoyés par les participants ;
— impossibles à transcrire par les annotateurs ;
— contenant le code *SMS05* et un numéro de téléphone.

5. Définition fournie par :http ://www.definitions-marketing.com/definition/echelle-de-likert/

La restriction de 160 caractères par SMS a entrainé le découpage automatique de certains messages. En effet lorsqu'un SMS contient plus de 160 caractères les opérateurs le coupent en morceaux de 153-160 caractères et envoient le message tronqué en morceaux. Par contre, les téléphones mobiles associent ces morceaux et n'en font apparaître qu'un. Cette tâche représente un problème car la première partie d'un SMS long contient bien le préfixe *SMS05*, qui permet de renvoyer les SMS reçus vers la plateforme, mais la deuxième partie ne contient pas forcement ce préfixe, une fois éloigné de la sa partie précédente qui commence par le code. Exemple de SMS brut tronqué :
Coucou,c'est chouette pr hier soir si tu as réussi a te protéger un peu.j'espere que tu as qd meme passé une bonne soirée.ce soir je vais jouer a l
Des messages contenant uniquement 66 caractères ont été, également, repérés tronqués. Des problèmes d'encodage sont à l'origine de ce fait car certains caractères spéciaux proposés par des mobiles ne figurent pas dans l'alphabet GSM 7 bits (Chabert *et al.*, 2012, Antoniadis *et al.*, 2011). Dans les deux cas de figure un message automatique d'erreur a été envoyé à l'utilisateur :
SMSPRO : ce service est réservé à des utilisateurs pré-enregistrés Vous ne pouvez pas avoir accès au service depuis votre téléphone mobile .
Ces messages ont été retenus dans le corpus mais annotés par les annotateurs comme étant découpés car nous ne pouvons pas omettre le fait qu'un tel problème nuit et rend obsolète la constitution et l'étude du corpus portant surtout sur l'analyse quantitative du corpus.

Par ailleurs, comme Cougnon (2015) l'affirme, la différence entre un corpus traditionnel et un corpus spécialisé tel que le corpus de SMS repose sur sa caractérisation quantitative. La caractérisation du corpus en terme quantitatif se trouve sur la description du nombre de phrases, formes/tokens et caractères. Toutefois, une telle analyse s'avère un grand défi pour toute sorte de corpus liés à la communication médiée par ordinateur. Du fait qu'il existe une variation imprévisible de formes (abréviations, omissions, etc.) propre à chaque utilisateur,

l'unité lexicale est difficilement identifiable et l'utilisation de la ponctuation est couramment inexistante. Dans la partie 5.2.2 nous évoquons de façon exhaustive les différents problèmes de tokenisation d'unités lexicales identifiés propre au langage SMS. A titre indicatif nous donnons les caractéristiques du corpus sur la table 3.5. Nous nous basons sur le corpus brut et transcrit de 21 261 SMS où nous avons, dans une première phase, calculé à l'aide de la ligne de commandes le nombre de caractères [6] et par la suite le nombre de tokens à l'aide de l'outil Unitex [7] Paumier (2003) pour le corpus de SMS transcrits. Nous remarquons que le corpus de transcription est plus long que le corpus brut, ce qui confirme qu'il s'agit d'un code écrit qui combine des procédés pour raccourcir les phrases.

	Corpus SMS brut	Corpus SMS transcrits
Nombre de caractères	1 383 469	1 558 648
Nombre de tokens	-	288 050

Table 3.5 – Description du corpus

Il existe une distinction en CMO entre deux modes de communication : synchrone (web chat, messagerie instantanée) et asynchrone (email, SMS, MMS, forum de discussions). Cela signifie en réalité que le moment de la production et de la réception du message n'est plus le même, à la différence de ce qui arrive en communication directe, (Bevilacqua *et al.*, 2012). Même si les SMS font partie de la communication asynchrone qui rend l'interaction directe impossible, comme Frehner (2008) le souligne, ils ont portant le *potentiel d'approximer la synchronicité et permettre une conversation écrite en temps quasi-réel à un point qui n'a jamais été connu avant*. Notre corpus ne peut qu'être un corpus de communication asynchrone puisqu'il contient des messages rédigés uniquement par l'expéditeur participant au projet. De ce fait, le corpus ne contient pas de discussions pour des raisons éthiques, en effet, pour que les SMS

6. *wc -m* : compte le nombre de caractères dans le fichier dans le terminal. La commande "wc" signifie essentiellement "le nombre de mots" et avec différents paramètres facultatifs peut être utilisée pour compter le nombre de lignes, de mots et de caractères dans un fichier texte.

7. http ://unitex.univ-mlv.fr/

de l'interlocuteur figurent parmi les SMS du corpus il est impératif d'avoir son consentement. Panckhurst et Moïse (2012) affirment sans doute, qu'un tel fait est assez contraignant pour les linguistes, sociologues et psychologues qui s'intéressent surtout aux aspects conversationnels.

Même si le corpus est marqué par certaines limites, nous partageons les idées de Fairon *et al.* (2006), Cougnon (2015) autour de la valeur qu'un corpus de SMS peut avoir : a) unique dans son genre ayant une taille importante, b) marqué par la diversité d'usagers, c) couteux en nécessité d'une longue durée de la conception jusqu'à la délivrance du corpus, d) couteux financièrement avec la nécessité publicitaire et humaine, e) difficile à construire car il demande une mise en place d'un système adéquat et d'un protocole garantissant la protection de la vie privée.

3.4 Conclusion

Ce chapitre est dédié à la présentation du projet alpes4science réalisé en 2010 dans les Hautes-Alpes et l'Isère au sein du laboratoire LIDILEM. Nous avons exposé en détail la procédure de participation d'utilisateurs aussi bien que le protocole établi pour l'enregistrement des données transmises : les réponses au questionnaire et les SMS. D'une part nous avons analysé les traitements effectuées sur le corpus de SMS afin de le rendre opérationnel : transcription et anonyisation de données sensibles et, d'autre part, nous avons effectué une analyse générale sur les SMS et les réponses au questionnaire (partie 3.3).

En dernière analyse nous résumons que le corpus de SMS issu de cette collecte constitue un ensemble de 21 261 SMS bruts accompagnés de leurs transcriptions pour constituer un corpus *monolingue*, *authentique*, et *accessible*, disponible en ligne [8]. En outre, malgré les diverses particularités que ce corpus présente il reste, en effet, un matériau précieux puisque

8. Corpus brut anonymisé de SMS disponible sur le site d'Ortolang :`https://hdl.handle.net/11403/comere/cmr-smsalpes`

les SMS sont difficiles à collecter en grande quantité. En plus, des questions éthiques et des contraintes techniques interviennent pour rendre cette tâche difficile à réaliser. Les messages récoltés nous permettront de présenter une analyse basée sur l'analyse et la classification de pratiques langagières repérées dans le corpus et l'exposition des principales caractéristiques lexicométriques.

De même, les réponses issues du questionnaire rempli par les utilisateurs nous permettent de réaliser une analyse du profil des participants que nous allons exposer dans la partie 4.2. Par ailleurs, ces données constituent un matériau incontournable pour la réalisation de recherches sociologiques, sociolinguistiques et comparatives sur l'observation de l'usage des SMS.

Chapitre 4

L'analyse du corpus alpes4science

Sommaire

4.1 Introduction **61**

4.2 Présentation des données socio-démographiques **62**

4.2.1 Le profil du participant 63

4.2.2 Les pratiques des participants 67

4.3 Analyse lexicométrique **69**

4.3.1 Principales caractéristiques lexicométriques du corpus 71

4.4 Analyse et typologie des pratiques langagières **84**

4.4.1 Typologie des formes du corpus 87

4.5 Conclusion **90**

4.1 Introduction

L'intérêt d'étudier le langage SMS réside dans les particularités que nous repérons au sein de ce discours. Le langage SMS se définit comme un aspect particulier de la communication, il s'agit d'un code écrit particulier qui combine plusieurs procédés pour raccourcir les phrases

et les mots, selon Stark (2011). En parallèle, il est proche de l'oral tout en étant une forme écrite et c'est pourquoi ce langage intéresse de nombreux chercheurs.

L'observation de ces particularités nécessite l'utilisation des données authentiques dans le but d'obtenir un point de vue plus objectif Fairon *et al.* (2006). La plupart de messages courts présente des différences importantes en comparaison avec le langage standard. En effet, les utilisateurs essaient d'utiliser diverses formes courtes pour abréger les mots dans l'objectif de gagner du temps tout en faisant le moindre effort. Par ailleurs, un des obstacles auquel nous devons faire face avec les systèmes de traitement automatique du langage est la morphologie particulière des mots SMS (fusionnement de mots, formes abrégées imprévisibles, suppression de caractères, manque de ponctuation, etc.).

Après avoir exposé en détail la constitution de la base de données issue du projet alpes4science nous explorerons, dans un premier temps, dans ce chapitre les principales caractéristiques qui reposent sur le croisement de données socio-démographiques (partie 4.2) fournies au travers du questionnaire que les participants du projet ont complété lors de la collecte des messages. Ensuite, nous verrons les principaux éléments lexicométriques (partie 4.3) visant à définir le niveau de richesse lexicale sur l'ensemble du corpus et sur certaines partitions liées aux profils des participants. La distribution des catégories grammaticales, le type de clavier employé lors de la saisie des messages et les n-grammes seront exposés par la suite. A la fin, nous illustrerons l'analyse typologique (partie 4.4) des formes que nous trouvons dans le corpus de SMS en nous basant sur la typologie modifiée proposée par Panckhurst (2009).

4.2 Présentation des données socio-démographiques

Il est d'un grand intérêt de présenter les données socio-démographiques des sujets participants avant toute analyse. Ceci consiste à détailler le public de participants lors du remplissage du questionnaire ; par ailleurs, au travers de ce public nous allons pouvoir définir à quel point

les réponses reflètent cette population. D'autre part, ceci nous permet d'identifier l'environnement social des productions langagières et d'investiguer les sujets observés, autrement dit, la façon dont ils se positionnent dans leur groupe de référence (Bulot et Blanchet, 2011). D'après la distinction que nous avons évoquée dans la partie 3.3.1 nous présenterons les deux différents axes avec leurs différentes variables concernant les 240 sujets.

4.2.1 Le profil du participant

4.2.1.1 Le sexe

Les sujets participants du projet se repartissent en 29% hommes et 71% femmes. Nous remarquons donc une forte participation des femmes. Ce constat est le même pour les autres projets de SMS, comme par exemple pour le projet *sud4science* où les hommes ne représentent seulement que 38% (Panckhurst et Moïse, 2012) (*cf.* figure 4.1). Quant aux autres projets, Cougnon (2015) affirme avec un diaporama qui concerne quatre projets (Belgique, La Réunion, Québec, Suisse) que la participation de femmes est majoritaire en supposant qu'un tel constat est probablement lié au fait que les femmes ont une tendance de répondre plus aux enquêtes ou encore à cause du noyau de collectes, c'est-à-dire les Facultés de Lettres, où nous trouvons majoritairement des femmes. De même, Moore et Tarnai (2002), Singer *et al.* (2000), Smith (2008) nous confirment que les femmes sont plus susceptibles à participer aux enquêtes que les hommes.

4.2.1.2 L'age

Comme nous pouvons le constater dans la figure 4.2 la répartition de la population au niveau de l'âge n'est pas du tout équilibrée, puisqu'elle est composée de 54,17% de personnes âgées de 18 à 23 ans. La deuxième tranche que nous trouvons par la suite est celle des 24-29 ans avec 46 personnes, qui représente 19,17%. Nous constatons une très faible participation de

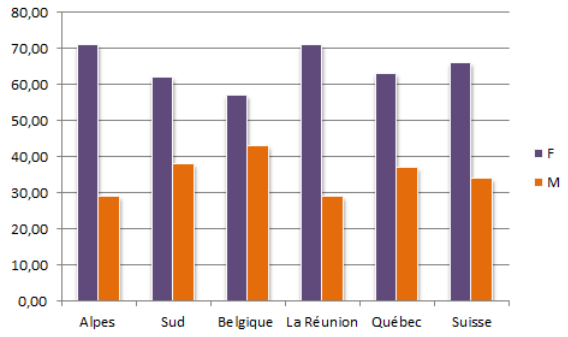

FIGURE 4.1 – Répartition du sexe dans les projets

personnes de 30 à 47 ans et une participation plus importante pour les sujets âgés de plus de 47 ans avec 7,50% en se positionnant ainsi à la troisième place avec le jeunes de moins de 17 ans (7,08%). Tel qu'il a été mentionné dans Cougnon (2015) autour l'analyse de quatre projets SMS notre projet est, en règle général, en accord avec les autres projets, c'est-à-dire que nous constatons une participation majeure de sujets de 18 à 23 ans et une faible participation de sujets de 30 à 47 ans. Le projet qui s'approche le plus de nos résultats est celui de la Suisse [1] avec les mêmes tendances globales, exceptée la participation de la tranche 30-35 où elle tient la troisième place plus importante.

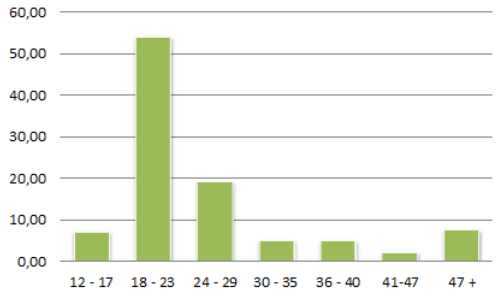

FIGURE 4.2 – Répartition des sujets selon l'âge dans le corpus

1. http://www.sms4science.ch/bin/view/Main/WebHome

4.2.1.3 Le niveau d'études

Le niveau d'études des sujets du projet est pris en compte au moyen de quatre catégories : 1) le niveau secondaire (collège/lycée), 2) les études supérieures (études après le bac), 3) le niveau universitaire avec le cycle inférieur (études jusqu'au bac+3) et 4) le niveau universitaire avec le cycle supérieur (master ou plus). La figure 4.3 illustre la distribution des niveaux d'études des sujets. Il y a une claire prédominance de sujets ayant une éducation universitaire du cycle inférieur et notamment du cycle supérieur. Cependant un tel constat peut se justifier puisque, comme nous l'avons évoqué, le centre de cette démarche était le domaine universitaire. Notons aussi que deux participants ne se sont pas prononcés.

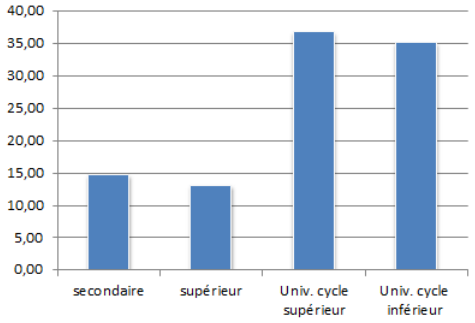

FIGURE 4.3 – Répartition des sujets selon le niveau d'études

4.2.1.4 Langue maternelle et régionale

La majorité des sujets a désigné comme langue maternelle la langue française, pourtant certains participants ont déclaré les langues suivantes : chinois, vietnamien, grec, arabe, polonais, catalan, allemand. En ce qui concerne les langues régionales nous trouvons le patois matheysin, du nord (ch'ti), ardéchois, alsacien, créole, normand et occitan.

4.2.1.5 La répartition géographique

Le lancement de la collecte a été effectué dans la région Rhône-Alpes, cependant des sujets provenant d'autres régions ont pu participer au projet. La figure 4.4 nous donne la localisation géographique des participants en fonction de leur lieu d'origine. On remarque une concentration de participants provenant du sud-est avec une grande participation de l'Isère, des Hautes-Alpes, de l'Ardèche et de la Savoie, un fait qui pourrait favoriser la recherche sur la linguistique de terrain.

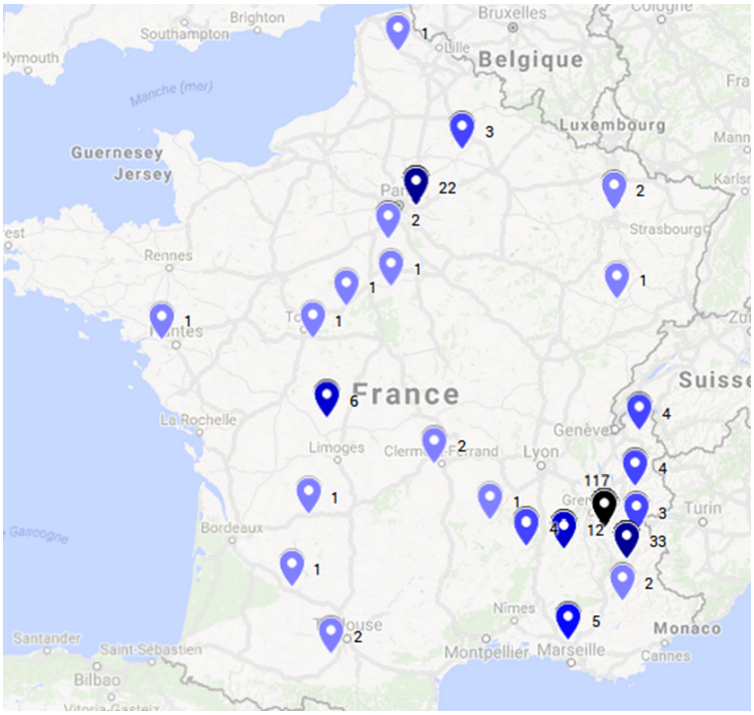

FIGURE 4.4 – Répartition géographique des sujets selon leur lieu d'origine

4.2.2 Les pratiques des participants

4.2.2.1 L'usage et compréhension sms

Pour caractériser les réponses que les sujets ont fournies nous avons procédé à un étiquetage de chaque réponse afin de formuler et de comprendre les raisons d'utilisation de SMS aussi bien que les traits de cette communication. Nous avons donc défini onze étiquettes :

— La *communication* qui inclut la transition et la réception d'information ou simplement la demande de nouvelles.

— La *nécessité* qui reflète le besoin de l'émetteur de transmettre via la voie écrite une information.

— L'*évitement de téléphoner* que pour certaines raisons les personnes déclarent préférer ne pas appeler leur interlocuteur, de ce fait ils privilégient le SMS pour communiquer.

— La *toute occasion* qui concerne ceux qui déclarent qu'ils utilisent ce moyen dans toute situation et tout le temps ou même sans situation particulière.

— Le *RDV* pour la planification de rendez-vous, sorties, soirées, loisirs et autres.

— Le *message court* afin de définir l'intention de transmettre des petits messages, courts, brefs avec juste quelques mots.

— L'*urgence* car plusieurs personnes déclarent utiliser les messages quand ils sont pressés, en cas d'une urgence ou quelque chose d'important mais pas formel à communiquer.

— La *rapidité* en accord souvent avec l'étiquette message court est employé pour souligner le besoin immédiat de communiquer, surtout en absence de réponse à un appel téléphonique.

— La *discrétion/indisposition* car ils affirment que dans certaines situations comme les endroits où ils sont entourés de monde, publiques, bruyants (train, métro, bus, etc.), où téléphoner c'est interdit (par exemple dans la salle de cours) la seule solution de communication reste le message écrit.

— Le *sans urgence* à l'opposé de l'étiquette *urgence* pour certaines personnes communiquer

par SMS est un moyen de transmettre leurs messages quand il y a pas une urgence imminente, la réponse attendue n'est pas urgente ou même ils utilisent les SMS en toute occasion sauf s'il s'agit d'une urgence.

— Les *jeux/concours* selon certaines personnes l'utilisation de messages écrits est exclusive pour leur participation à des jeux ou concours.

Pour mieux visualiser la distribution de ces étiquettes dans les réponses des participants à la question de l'usage de SMS nous proposons la consultation de la table 4.1. Nous arrivons à conclure que comme attendu la communication par SMS est en accord avec le besoin de la communication, il s'agit d'un désir de partager des événements de la vie courante, donner et demander des nouvelles et de poser une question pour obtenir ou transmettre une information. Un autre trait que nous repérons est celui de la discrétion et de l'indisponibilité qui se caractérise comme un sujet le décrit : *Pour joindre quelqu'un tout en sachant que l'appeler le dérangerait, ou simplement par manque de temps pour appeler.* Par ailleurs, nous observons qu'une grande partie (9,77% de participants) considère que les SMS sont un moyen pour organiser des sorties et prendre des rendrez-vous entre amis et avec la famille.

Étiquette	%
Communication	37,99
Nécessité	3,92
Évitement de téléphoner	5,58
Toute occasion	8,11
RDV	9,77
Message court	5,86
Urgence	7,27
Rapidité	5,03
Discrétion/indisposition	12,85
Sans urgence	3,07
Jeux concours	0,55

TABLE 4.1 – Répartition du pourcentage d'étiquettes attribuées à l'usage de SMS

La deuxième partie de cette analyse concerne la compréhension des SMS et plus préci-

sément si la compréhension des codes et abréviations est facile. La réponse à cette question démontre que 57,5% des sujets estiment que comprendre un message contenant des codes et des abréviations est difficile et que 42,5% comme facile.

4.2.2.2 Nombre d'envois par semaine

Les sujets participants affirment envoyer plus de 50 messages par semaine (environ 33%) et seulement 15% des participants déclarent envoyer environ 5 messages par semaine. Il faut noter sur ce point que pendant la période de la collecte les forfaits de téléphones bénéficiant d'un envoi de SMS illimités étaient un phénomène relativement rare et assez coûteux. Le tableau ci-dessous résume le nombre d'envois de messages par semaine par les participants.

SMS par semaine	%
>50	33,6
>20	30,4
>10	20,1
>5	15,6

Table 4.2 – Répartition du nombre d'envoi de SMS par semaine

4.3 Analyse lexicométrique

L'utilisation de données discursives pour l'analyse quantitative du discours constitue un élément d'un grand intérêt pour les chercheurs de nombreuses disciplines des sciences humaines et sociales (linguistique, psychologie, sociologie, politique, histoire, etc.). L'*Analyse Automatique du discours*, que Pecheux a introduit en 1969, marque le début d'un développement continu pour l'examen et la rectification d'idées fondamentales autour de la théorie du discours et de son analyse. Il devient, dorénavant, le leader de l'analyse discursive en français (Helsloot et Hak, 2000, Hank et Helsloot, 1995).

Pour Williams (2014) la lexicométrie *n'est pas une méthode uniforme mais une panoplie d'approches qui cherche à capturer les différentes propriétés du corpus.* Dans une tentative de définir ce terme nous citons également la définition de Frehner (2008) qui englobe un ensemble de méthodes *permettant d'opérer, à partir d'une segmentation, des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire.*

Comme Frehner (2008) le mentionne, l'utilisation de textes en langue naturelle doit se réaliser avec un strict minimum de transformations afin de pouvoir procéder à une analyse de tout type de données orales ou écrites. En effet, le corpus ne doit pas subir une préparation quelconque. Dans le cas d'un corpus de SMS, il est impossible de travailler en format brut. Effectivement, la première étape pour réaliser un traitement lexicométrique est de réaliser une segmentation automatique du texte en occurrences de formes graphiques. Néanmoins, nous identifions plusieurs problèmes de segmentation des unités contenues dans les SMS, comme par exemple l'utilisation d'abréviations non standard de mots composés (mdr → mort de rire), le manque d'espaces et de signes de ponctuation et le mélange de chiffres et de caractères. Une analyse exhaustive de ce type de problèmes sera exposée dans la partie 5.2.2.

Les outils permettant d'effectuer des analyses lexicométriques, dans une première phase, réalisent la tokenisation du corpus textuel en unités, *les occurrences/tokens*, qui par la suite servent de base aux décomptes ultérieurs et permettent de mieux appréhender le vocabulaire de ce corpus.

Nous résumons les fonctionnalités basiques d'un outil lexicométrique aussi bien que les possibilités offertes par l'approche des corpus de textes :

— Les *concordances (KWIC*[2]) permettent la visualisation du contexte gauche et droit d'une forme à partir d'une recherche de motifs lexicaux simples ou complexes.
— La recherche de *patterns morphosyntaxiques* permet la recherche d'un pattern composé de formes ou d'étiquettes morphosyntaxiques.

2. Key Word In Context

— L'*index hiérarchique* qui fournit une liste avec les mots du corpus classés par ordre de fréquence.
— L'*index alphabétique* qui fournit une liste avec les mots du corpus classés par ordre alphabétique permettant, également, de rapprocher les utilisations du singulier et du pluriel d'un même substantif, les différentes flexions d'un verbe ou adjectif (Leimdorfer et Salem, 1995).
— Les *segments répétés* permettent la recherche de formes qui apparaissent à plusieurs endroits du texte.
— Les listes de *spécificités* avec les mots statistiquement surutilisés ou sous-utilisés dans chacune des parties du corpus.
— Les *cooccurrences* afin de calculer les associations spécifiques, autrement dit les unités qui apparaissent plus fréquemment dans son entourage.
— La *densité lexicale* utilisée avec le calcul du rapport type/occurrences, type/token ratio (TTR) en anglais, aussi bien que ses variantes.
— L'*édition* et *exportation HTML* pour chaque texte du corpus.

4.3.1 Principales caractéristiques lexicométriques du corpus

Comme nous l'avons déjà signalé, l'utilisation du corpus brut de SMS s'avère impossible dans l'objectif de pouvoir effectuer une quelconque analyse lexicométrique et caractériser le corpus brut de façon quantitative. De ce fait, nous ne pouvons que nous baser que sur le corpus transcrit manuellement par les annotateurs(*cf.* partie 3.2.3.3).

Les données des participants qui nous permettent de réaliser les analyses de cette étude sont issues du questionnaire que nous avons exposé dans la partie 3.3.1 dont les modalités ont été décrites dans le cadre du projet alpes4science. Pour effectuer les analyses lexicométriques avec croisement des données démographiques nous nous basons sur 96,4 % des messages que ces participants ont envoyés. Les informations démographiques des participants sont résumées

dans le tableau 4.3. Les données sociologiques tirées des réponses fournies par les participants du projet sont exposées en détail dans la partie 4.2.

	Participants
Caractéristique	**(*N*=240)**
Sexe	29% H :71% F
Age	14 - 69 ans
Niveau d'études	
Secondaire	35
Supérieur	39
Univ. cycle supérieur	88
Univ. cycle inférieur	84

TABLE 4.3 – Informations démographiques de participants

Une préparation des données a été effectuée au niveau des SMS dans l'intention de rendre l'ensemble de messages homogène, puisque lors de la transcription fournie par les annotateurs certains éléments complémentaires ont été ajoutés aux messages transcrits, qui n'était pas initialement présents dans les messages bruts : a) désanonymisation de données sensibles, b) élimination d'éléments manquants introduits par les annotateurs entre *[/.../]*, par exemple : Je *[/ne/]* sais pas. Pour réaliser les analyses concernant l'emploi de catégories grammaticales l'ajout de l'annotation morphosyntaxique se relève indispensable. Du fait que la taille de notre corpus est assez importante nous nous basons sur un échantillon du corpus de 1000 SMS, choisis aléatoirement, que nous avons annoté automatiquement et révisé manuellement dans le but de garantir la fiabilité de résultats.

4.3.1.1 Mesure de la diversité/richesse lexicale

La production écrite, en terme de chiffres, s'avère plus importante que la production orale qui se caractérise plutôt par une certaine rapidité et spontanéité. De même, Martinon (1927) l'avait affirmé dans son ouvrage : *on ne parle pas tout à fait comme on écrit, pas plus qu'on*

ne peut écrire tout à fait comme on parle, et la formule vous parlez comme un livre n'est un compliment que dans la bouche des ignorants. Il est donc intéressant d'observer les résultats d'une telle mesure dans le langage SMS qui se trouve à l'intervalle entre oral et écrit. Du même avis Anis (1998), Collot et Belmore (1996), Yates (1996) (dans Cougnon (2015)) soulignent une note flexible, *hybride*, tout en équilibrant parfois entre l'oral et l'écrit. Sur ce sujet Fairon *et al.* (2006) se positionnent afin de clarifier la réponse à la question de l'oralité du langage SMS ; effectivement, ils déclarent que le langage SMS est submergé de phénomènes typiques à la langue parlée, mais pourtant cela reste avant tout une langue écrite. De la même façon, Cougnon (2015) conclut que, sans doute, il s'agit d'un écrit avec une volonté de lui attribuer un caractère plutôt *neutre*. Sur le même axe Véronis et Guimier de Neef (2006) et Panckhurst (2007), clarifient qu'il ne s'agit que d'une *question de registres et de fréquences*, car en effet, selon le destinataire et la situation de communication (familière, courante) le discours peut devenir formel ou informel que cela soit oral ou écrit.

La diversité lexicale concerne les mots utilisés dans un texte par un énonciateur, ces mots reflètent la capacité de l'individu à accéder et à récupérer les mots cibles à partir d'un lexique pour la construction d'un ensemble d'unités linguistiques (Fergadiotis *et al.*, 2013), autrement dit, il s'agit du nombre de mots différents employés par un individu afin de composer ces énoncés.

Quant à la richesse lexicale, Muller (1969) déclare dans son ouvrage *La statistique lexicale* qu'il s'agit d'une notion *relative* qui ne peut qualifier un texte en tant que *riche* ou *pauvre* que par comparaison avec d'autres textes. De la même manière, il définit la richesse lexicale exclusivement par le nombre de vocables : *Cette façon de voir considère le texte comme un ensemble clos et achevé (même s'il s'agit d'un fragment ou d'une tranche), formé de N mots, et dont on mesure la richesse par le nombre des V vocables qui y figurent* (Muller, 1977, p. 116).

La mesure la plus couramment utilisée pour calculer la diversité du vocabulaire est le

rapport de type-occurrence (Type/Token Ratio en anglais) (Templin, 1957). L'occurrence représente le nombre total de mots (tokens) et le type correspond aux mots qui apparaissent pour la première fois dans le texte. Par exemple, admettons que nous avons un texte composé de 4 000 mots, c'est-à-dire de 4 000 tokens, parmi ces mots il y en a certains qui se répètent et seulement 1 000 sont des différents mots, les types. Si nous calculons donc le rapport entre les types et les tokens (table 4.4) nous obtenons comme résultat 25 %.

Rapport Type/Token = (nombre de types/nombre de tokens) * 100
(1 000/4 000)*100 = 25 %

Table 4.4 – Exemple de formule TTR

En appliquant le rapport sur l'ensemble de 21 260 SMS nous obtenons 305 675 tokens et 12 328 types ce qui correspond ainsi à un rapport entre type/token de 4,08% [3] (table 4.5).

SMS	Tokens	Types	TTR	Std TTR
21 260	301 807	12 328	4,08%	37,35%

Table 4.5 – Principales caractéristiques lexicométriques du corpus alpes4science

Avec la volonté de vouloir évaluer la richesse lexicale de ce corpus, nous utilisons comme point de repère les analyses effectuées par Cougnon (2015), telles qu'elles figurent dans son ouvrage portant sur le langage SMS, en nous offrant la possibilité d'extraire les principales caractéristiques lexicométriques des différents corpus du projet *sms4science* (table 4.6) :

Ainsi, nous nous permettons de comparer ce résultat avec les 4 corpus de SMS francophones et de conclure que selon l'indice du rapport TTR notre corpus présente une richesse lexicale similaire à celle du corpus belge et que le corpus québécois garde la première place. Le rapport

3. Les résultats des analyses ont été produites de façon automatique grâce au logiciel WordSmith, un outil pour l'analyse lexicale, développé par Mike Scott à l'Université de Liverpool disponible en ligne : http ://lexically.net/wordsmith/. Les unités ont été segmentées au niveau des espaces et de ponctuation.

Projet	Tokens	Types	TTR	Std TTR
Belgique	688 550	24 265	4,0%	40,7%
La Réunion	233 285	13 080	5,6%	39,1%
Québec	60 809	7 056	11,6%	43,1%
Suisse	94 398	7 395	7,8%	40,3%

TABLE 4.6 – Principales caractéristiques lexicométriques de 4 projets SMS francophones dans Cougnon (2015)

type/token standardisé [4] (Scott, 1996) a été utilisé par Cougnon (2015) pour calculer le rapport pour les 1 000 premiers mots, puis calculer à nouveau pour les 1000 mots suivants, et ainsi jusqu'à la fin du corpus, c'est-à-dire que nous obtenons un rapport de type/token moyen basé sur des blocs de texte de 1000 mots consécutifs.

Selon ce rapport, le corpus occupe la dernière place avec une faible richesse lexicale, mais ces chiffres sont-ils vraiment représentatifs vu l'hétérogénéité de la taille des corpus ?

En effet, plus il y a de types en comparaison avec le nombre de tokens, plus la variété lexicale dans le texte est grande. Par conséquent, le rapport varie très largement en fonction de la longueur du texte d'analyse ; ainsi un nombre faible d'occurrences donne un TTR élevé. Même si, selon Heaps (1978), tel qu'il a été rapporté par Fergadiotis *et al.* (2013), *lorsque la longueur de l'échantillon augmente, il est moins probable qu'un locuteur produise de nouveaux mots puisque le nombre d'éléments lexicaux pouvant être activés à un moment donné est considéré fini.* De ce fait, il est possible que quand la longueur du texte augmente, la probabilité de trouver des répétitions comme par exemple des mots grammaticaux à fréquence très élevée augmente, de telle sorte que le rapport peut diminuer (Gayraud, 2001). Par conséquent, (Xanthos, 2013, p 232) déclare que : *Les mesures de diversité lexicale dérivent plus ou moins directement de la variété, soit le nombre de mots distincts (ou types) dans un corpus ; or,*

4. Cette métrique est utilisée lorsqu'il est difficile de comparer le TTR des petits textes avec des textes plus grands. En effet, quand un texte est grand le nombre de nouveaux mots types diminue. Afin de remédier à cela, WordSmith peut calculer le TTR basé sur tous les 1000 mots et produire la moyenne du TTR.

la variété dépend de façon évidente, du moins pour sa valeur maximale, du nombre de mots successifs (ou tokens) qui composent le corpus. A ce titre, elle ne permet pas de comparer directement des corpus de longueurs différentes – une critique à laquelle n'échappe pas le rapport types-tokens (RTT) V/N, sans doute l'indice le plus utilisé dans l'histoire de la mesure de la diversité lexicale.

C'est ainsi que, contrairement à Cougnon (2015), nous avons fait le choix de ne pas comparer le corpus avec tout autre genre de texte (littéraire, journalistique, etc.) car nous considérons que toute conclusion pourrait sembler non fiable à cause de la taille et le genre de corpus.

Dans ce cas-là, afin d'éviter un résultat inverse qui pourrait modifier la fiabilité d'analyse, nous procédons à un découpage du corpus par tranche d'âge et d'éducation en distribuant de façon équitable les SMS tout en distribuant un nombre égal de SMS pour chaque catégorie (1000 SMS pour la tranche d'âge et le niveau d'études).

4.3.1.2 L'âge

Si nous considérons le rapport type/token pour mesurer la diversité lexicale en fonction de l'âge des sujets, nous obtenons les résultats suivants, illustrés dans le tableau 4.7.

Age	Tokens	Types	TTR	Std TTR
14 - 19	12 220	1 591	13,02%	34,83%
20 - 25	14 998	1 887	12,58%	35,37%
26 - 35	12 447	1 924	15,46%	36,98%
36 -	12 936	2 097	16,21%	38,25%

TABLE 4.7 – La diversité lexicale en fonction de l'âge

Nous constatons, également, à partir du diagramme de la figure 4.5 que la diversité lexicale est évolutive en fonction de la tranche d'âge. Notamment, entre les sujets de 14 à 19 ans et les sujets âgés de plus de 36 ans nous observons une variété lexicale plus importante pour

ces derniers. Pour nos conclusions nous nous basons sur le rapport standardisé, car comme illustré dans la figure 4.5 le rapport TTR peut nous conduire à une fausse conclusion, puisque comme nous pouvons le remarquer dans le tableau récapitulatif 4.7 il y a une concentration de tokens plus élevée pour les sujets de 20 à 25 ans, ce qui signifie que la taille du texte est plus importante malgré les précautions prises.

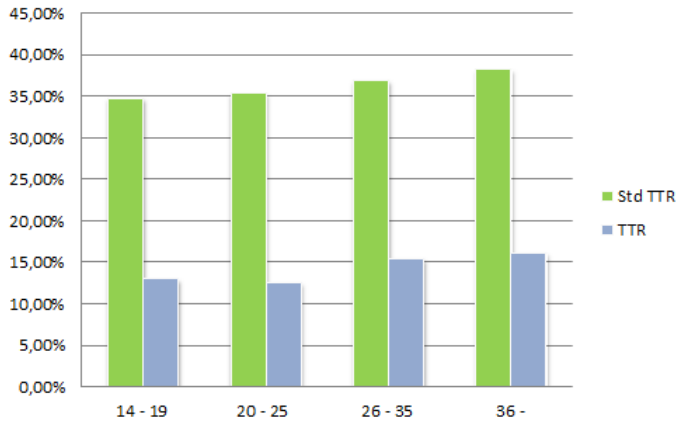

FIGURE 4.5 – Rapport Type/Token et Type/Token standardisé en fonction de l'âge

Pour cette raison, nous avons pris la décision d'appliquer deux mesures supplémentaires : une relativement récente et indépendante de la taille du texte et une autre souvent utilisée dans la mesure de la diversité lexicale car elle équilibre les inégalités dans le but de réévaluer nos données.

1) Il s'agit de la mesure *Moyenne Mobile du Rapport de Types/Tokens* (Moving-Average Type–Token Ratio (MATTR) en anglais) introduite par Covington et McFall (2010) [5]. La mesure offre un certain nombre d'avantages par rapport à la mesure classique de *TTR*, en effet, elle n'est fondée sur aucune théorie statistique relative à l'ajustement du rapport type/token pour la taille du texte, elle est donc indépendante de la longueur du texte et elle est calculable rapidement. L'algorithme sélectionne une longueur de fenêtre de n tokens, et le TTR pour les tokens de *1* à n est estimé. Ensuite, le TTR est estimé pour les tokens de *2* à $(n+1)$, puis de

5. Les résultats ont été produits grâce au logiciel libre disponible en ligne : www.ai.uga.edu/caspr

3 à (*n+ 2*), et ainsi de suite pour l'ensemble de l'échantillon. Le score final est la moyenne des TTR estimés Fergadiotis *et al.* (2013, 2015).

2) L'indice de diversité lexicale proposé par Carroll (1964) appelé Rapport Types/Tokens Corrigé (Corrected Type-Token Ratio en anglais), est souvent utilisé pour mesurer la diversité lexicale car il équilibre les inégalités. L'indice est représenté par le nombre de mots différents (types) divisé par la racine carrée de deux fois le nombre de mots (tokens) dans l'échantillon, une mesure de la diversité lexicale qui reste approximativement indépendante de la taille de l'échantillon, selon Richards (1987) comme il a été rapporté par Carroll (1964).

La figure 4.6 montre les résultats obtenus par l'application de ces deux mesures.

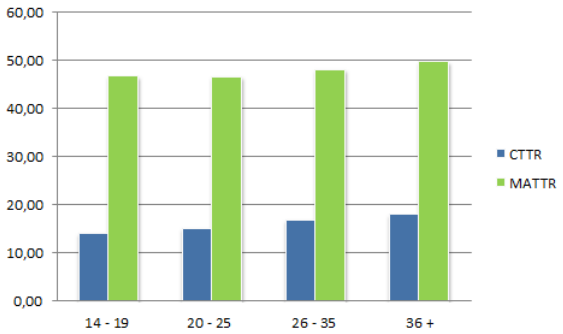

FIGURE 4.6 – Rapport Type/Token Corrigé et Moyenne Mobile du Rapport de Types/Tokens en fonction de l'âge

Contrairement, au calcul de la diversité lexicale produite grâce au TTR (cf. figure 4.5) où nous avons constaté une régression pour les sujets âgés de 20 à 25 ans comparés aux sujets de 14 à 19 ans, nous constatons ici que nous obtenons une courbe développante croissante. Par ailleurs, en accord avec Gayraud (2001) l'âge est un élément qui influencerait significativement l'augmentation de la diversité lexicale. Nous confirmons, également, le constat que nous avons évoqué précédemment, c'est-à-dire une évolution pour toutes les tranches d'âge et une minime régression selon le score de MATTR pour les sujets âgés entre 20 et 25 ans. Nous concluons ainsi que les sujets plus âgés déploient un vocabulaire plus riche et varié en comparaison avec

les jeunes.

4.3.1.3 Niveau d'études

Nous proposons dans cette partie de mener une investigation sur un échantillon de 1 000 messages, en examinant les caractéristiques lexicales de quatre groupes (études secondaires, études supérieures, études universitaires cycle inférieure et études universitaires cycle supérieur) de sujets selon leurs différents niveaux d'appartenance d'éducation (table 4.3). Quand nous confrontons les mesures pour nos quatre groupes, ainsi que leur performance aux tests complémentaires, nous trouvons des différences importantes. Ces résultats sont regroupés dans le tableau 4.8.

Niveau	Tokens	Types	TTR	CTTR	MATTR
Sec	14988	1772	11,82%	10,23%	45,4%
Sup	15271	2071	13,56%	11,85%	46,2%
Univ. cycle inférieur	15138	1950	12,88%	11,20%	47,1%
Univ. cycle supérieur	14739	2271	15,41%	13,22%	50%

TABLE 4.8 – La diversité lexicale en fonction du niveau d'études

Pour une meilleure visualisation nous pouvons également consulter le digramme 4.7 où nous trouvons un désaccord entre la mesure MATTR et celle de CTTR et par extension celle de TTR. En effet, les sujets qui ont acquis le niveau d'études supérieures présentent un lexique plus riche que ceux qui ont suivi des études universitaires du cycle inférieur en ce qui concerne les mesures TTR et CTTR, quant à la mesure MATTR nous observons une courbe croissante tout en constatant un écart minime entre les deux niveaux d'études (études supérieures et universitaires inférieures). Sans doute, nous pouvons admettre que les sujets ayant suivi des études universitaires du cycle supérieur démontrent une variété lexicale beaucoup plus importante en comparaison aux autres niveaux, toutes mesures confondues. En même temps, les sujets de niveau d'études secondaires ont le plus faible taux de variété lexicale (toutes mesures confondues) par rapport aux autres sujets.

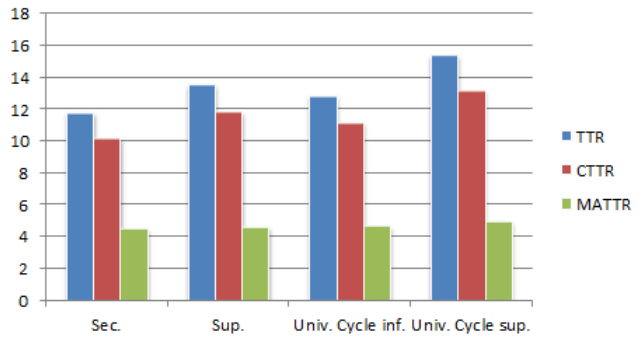

FIGURE 4.7 – Rapport Type/Token Corrigé et Moyenne Mobile du Rapport de Types/Tokens en fonction du niveau d'études

4.3.1.4 Emploi des catégories grammaticales

Nous poursuivons notre analyse avec la distribution des catégories grammaticales dans le corpus transcrit. Nos observations de ces données nous ont mené aux résultats exposés dans la figure 4.8.

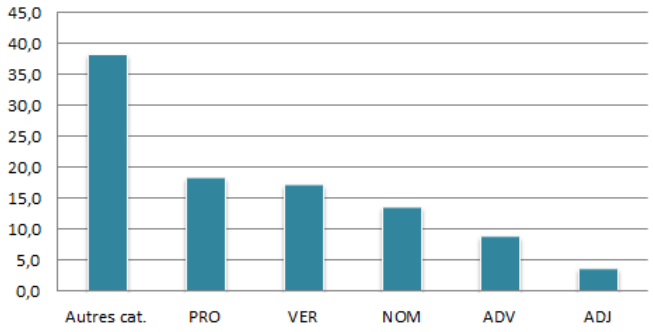

FIGURE 4.8 – Catégories grammaticales dans le corpus

En premier lieu, nous pouvons constater une grande concentration d'utilisation de pronoms (18,4%), de verbes (17,2%), aussi bien que de substantifs (13,6%), en même temps que les

adjectifs indiquent une faible utilisation (3,7%) (cf. en Annexe C.1 le tableau complet de données). De ce fait, nous observons de l'intérieur les deux catégories les plus utilisées. Plus précisément, le temps verbal le plus employé dans l'ensemble du corpus est l'indicatif du présent avec l'utilisation de 62,9% (et une utilisation majoritaire du verbe *être, avoir, aller, aimer*[6]), l'infinitif avec 14,8% et la participe passé avec 12,7%. Nous supposons qu'un tel constat est dû au fait que le temps du présent peut être mobilisé assez souvent pour évoquer un événement ou une action produite au passé ou futur ou encore par facilité de conjugaison, par exemple :
Passé : *On vient de se mettre d'accord pour partir demain vers 10 heures.*
Futur : *Je suis dans le train, j'arrive à 14h30. Je t'appelle à mon arrivée.*
Nous remarquons, également, une très faible utilisation du participe présent (0,4%) et du subjonctif du présent (1,1%).

La catégorie grammaticale dominante est notamment celle des pronoms personnels et plus précisément la première et la deuxième personne, le pronom que nous trouvons à la troisième place est le pronom démonstratif *ça*. Sur ce point, nous voulons préciser que la catégorie pronoms inclut majoritairement les pronoms personnels car elle correspond à 71,3% , les démonstratifs (18,3%), les relatifs (7,9%), les indéfinis (2,2%) et à la dernière place nous avons les possessifs (0,2%) (cf. en Annexe la figure C.2). En accord avec Barbizet et Lenoir (1968), nous confirmons pour notre cas que nous observons en règle générale une augmentation du nombre de pronoms lorsque le nombre de substantifs diminue, ceci étant un des traits du *langage normal*, le pronom normalement se qualifie comme le substitut du nom.

6. Données produites automatiquement grâce à l'outil de lemmatisation et étiquetage morpho-syntaxique TreeTagger(Schmid, 1994) incorporé au logiciel textométrique TXM (Heiden *et al.*, 2010) :http ://textometrie.ens-lyon.fr/

4.3.1.5 Type de clavier

Une autre variable qui constitue un intérêt pour notre analyse est celle du type de clavier (azerty ou alpha-numérique) employé lors de la production des messages.

FIGURE 4.9 – Clavier azerty

FIGURE 4.10 – Clavier alpha-numérique

En effet, nous nous intéressons à examiner la longueur moyenne des unités lexicales dans les messages transcrits et bruts par rapport aux deux types de clavier. Le résultat de cette analyse illustré dans la figure 4.11.

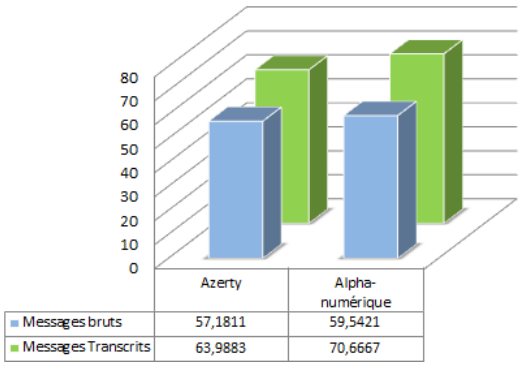

FIGURE 4.11 – Longueur moyenne d'unités lexicales par type de clavier pour les messages transcrits et bruts

Une première remarque concerne la disposition des messages bruts écrits au moyen d'un clavier *azerty* et *alpha-numérique*, nous constatons donc qu'il y a un écart de 4,12% puisque les

messages écrits avec un clavier azerty sont moins longs. De l'autre côté, les messages transcrits écrits avec un clavier azerty se comportent de la même manière que les messages plus courts et montrent un écart plus important (10,43%).

Notons encore que les messages en azerty bruts et transcrits sont plus brefs par rapport aux messages qui ont été rédigés avec un clavier alpha numérique. Nous constatons ces écarts dans la table 4.9 qui illustre le taux de réduction moyen en pourcentage pour les deux types de clavier en comparaison avec les messages bruts et leurs transcriptions. Nous arrivons à conclure ainsi que les messages bruts et écrits par un clavier alphanumérique sont plus longs. Nous voulons, également, signaler sur ce point que 18,8% des messages ne contiennent pas d'abréviations et d'allongements et que seulement 2,3% des messages du corpus se composent de formes allongées.

Taux de réduction moyen Azerty en %	Taux de réduction moyen Alpha-numér. en %
10,63%	15,74%

Table 4.9 – Taux de réduction moyen en % selon le type de clavier

4.3.1.6 Les n-grammes

Afin d'observer quelques caractéristiques à l'intérieur des messages, comme les phrases typiques que nous pouvons trouver dans ce corpus, nous avons procédé à une segmentation en *n-grammes*. Un *n-gramme* se définit comme une séquence d'éléments n contenus dans une séquence donnée, ils sont souvent utilisés pour désigner une séquence de lettres, par exemple un tri-gramme est un mot composé de trois caractères. Il est aussi possible, comme dans notre cas, d'employer le terme *n-gramme* pour désigner des séquences de mots et non de caractères. Le tableau récapitulatif 4.10 nous indique les listes de phrases de *2-grammes* à *4-grammes* les plus fréquents dans le corpus. Nous avons fait le choix de ne pas afficher les phrases composées de *n-grammes* de plus de quatre unités car les occurrences étaient égales à 2 et 1, et le résultat

ne pourrait pas être significatif.

	2-grammes		3-grammes		4-grammes	
1	c'est	149	j'espère que	19	j'espère que ça	5
2	j'ai	122	j'ai pas	17	je n'ai pas	5
3	ça va	56	je t'aime	14	s'il te plait	5
4	je vais	35	c'est bon	12	on peut se voir	4
5	je t'	31	je sais pas	12	espère que ça va	4
6	pas de	30	c'est pas	12	je suis à la	4
7	je te	29	je viens de	11	j'espère que tu	4
8	y a	29	si tu veux	10	pour l'instant je	3
9	si tu	29	tu fais quoi	9	je viens de voir	3
10	que tu	28	s'il te	9	je sais pas si	3

TABLE 4.10 – Table récapitulatif de 2 à 4-gramme)

La première remarque est un constat que Tagg (2012) a cité dans son ouvrage autour de l'analyse de communication par SMS et qui repose sur le fait que ces fragments de textes font évidement partie de morceaux de texte plus longs portant du sens, par exemple la phrase *j'espère que* peut correspondre à *j'espère que tu vas bien* ou encore à *espère que tu vas bien et*. En accord avec l'analyse que nous avons exposée dans la partie 4.2.2.1 concernant les réponses fournies par les participants du projet sur l'usage des SMS nous remarquons le caractère affectif : *je t'aime, j'espère que tu vas bien, fais de beaux rêves, que tu vas bien et que*, informatif : *j'ai vu que tu, pour l'instant je n'ai, je sais pas si, je viens de voir* et organisationnel : *on peut se voir, je suis à la, pour l'instant je* .

4.4 Analyse et typologie des pratiques langagières

Le discours électronique se caractérise par une hétérogénéité, un manque parfois de standardisation et une forte présence de créativité lexicale. Parmi ces caractéristiques nous repérons, par ailleurs, le lien entre l'oral et l'écrit tout en lui apportant, d'après Panckhurst (2009)

une dimension : *néologique* marquée par la créativité lexicale et l'utilisation d'emprunts ou *néographique* car les graphies échappent à la norme orthographique.

Toutefois, avant l'analyse de Panckhurst (2009) portant sur la classification et par extension la typologie du *discours électronique médié* afin d'apporter une typologie claire, effective et synthétique, d'autres chercheurs se sont positionnés à propos de ce sujet dans le prisme d'une analyse avec l'intention de clarifier les aspects de ce langage. Dans Anis (2003) concernant son analyse des formes linguistiques nous trouvons la distinction entre la variation graphique et certaines particularités morpho-lexicales avec un panorama assez large et bien détaillé. De manière similaire, Fairon *et al.* (2006) distinguent les phénomènes phonétiques entre la phonétisation de caractères et l'orthographe phonétique et les phénomènes graphiques des phénomènes lexicaux et morpho-syntaxiques, tout en restant assez détaillés et exhaustifs. De leur coté Véronis et Guimier de Neef (2006) résument dans quelques catégories les principales caractéristiques de formes donnant ainsi un aspect plus simple et spécifique. Parmi les derniers à avoir proposé une typologie nous trouvons Cougnon *et al.* (2013) dans une tentative de définir la typologie d'erreurs orthographiques dans les SMS ils répertorient les phénomènes graphiques provenant de stratégies réductionnelles, nous listons ici les catégories proposées :

— *Apocope* qui inclut tout type d'abréviation (âge sexe ville → ASV, aujourd'hui → ajrd)
— *Aphérèse* pour la réduction d'un mot en quelques caractères (salut → lut)
— *Syncope* pour la réduction de certains digrammes ou trigrammes et les squelettes consonantiques (problème → prbl)
— *Phonétisation* avec toutes les phonétisations possibles par substitution ou réduction (quoi → koi, demain → 2m1)
— *Casse fonctionnelle* c'est-à-dire le choix entre majuscule et minuscule
— *Chute massive de l'accentuation*

Quant à l'analyse menée par Marcoccia (2016) dans son ouvrage *Analyser la communication numérique écrite* qui porte sur les caractéristiques générales du discours numérique écrit,

nous distinguons d'un côté l'aspect formel et remarquable de l'oralité, l'abréviation et l'iconicité et de l'autre côté nous trouvons les *spécificités* de ce discours en cinq niveaux comme nous le résumons dans la table 4.11. Même s'il ne s'agit pas clairement d'une typologie mais plutôt de caractéristiques fréquemment repérées, nous constatons que les catégories des différents niveaux ne sont pas vraiment distinctes. Les spécificités morphologiques et orthographiques incluent les abréviations/procédés d'abréviation, l'écriture phonétique/phonétisation pour désigner les mêmes exemples ce qui laisse penser à un chevauchement des niveaux. De même, l'auteur utilise un exemple identique pour expliquer la création d'acronymes au niveau lexical et morphologique (mort de rire → mdr).

Typologique	**Orthographique non-standard**	**Morphologique**
symboles non-alphabétiques	abréviations	procédés d'abréviations
casse fonctionnelle	écriture phonétique	troncatures
émoticônes	prononciation non-standard	acronymes
ajout de signes de ponctuation	données prosodiques	phonétisation
Lexical	**Syntaxique**	
néologismes	composants manquants	
création d'acronymes	phrases incomplètes	

Table 4.11 – Spécificités du discours numérique selon Marcoccia (2016)

La typologie modifiée de l'écriture SMS de Panckhurst (2009), (2015) atteint les objectifs visés par l'auteur grâce à une description claire, simple et effective des phénomènes, sans mélange des catégories, qui nous semble répondre le mieux aux besoins d'une analyse typologique aussi complète et constituée en quatre catégories distinctes contenant d'autres sous-catégories (phonétiques et/ou graphiques) [7].

La *substitution* se divise en phonétique pour inclure les unités ayant subi une transformation totale (eau → o), partielle (bises → bizess) ou variationnelle (k → que) et en graphique pour décrire l'omission de l'apostrophe, la *casse fonctionnelle* ou du trait d'union (m'en → m en,

7. En Annexe le tableau typologie de phénomènes simples dans (Panckhurst, 2009, 2015, 2017)

est-ce que → est ce que), le rébus et les symboles spécifiques (à plus → à +) et les éléments variationnels (bisous → bisoux).

La *réduction* se répartit également en phénomènes phonétisés pour se diviser par la suite en abrègements morpho-lexicaux (ordinateur → ordi, mort de rire → mdr, fac → faculté, c'est → c), en variations (ui → oui) et en graphiques qui par la suite se divise en trois autres catégories : la suppression de fins de mots muettes (vous → vou), les squelettes consonantiques (pour → pr) et les agglutinations (j'arrive → jarrive).

La *suppression* ne contient que des éléments graphiques pour analyser les phénomènes de typographie et de ponctuation et décrire l'omission de signes diacritiques (ça → ca, à → a) à l'intérieur d'une unité lexicale.

L'*ajout* concerne l'ajout d'éléments graphiques pour la répétition de caractères ou de ponctuation (superrrrrrr!!!!!!), l'ajout d'émoticônes (:p), des caractères (peu → peut) et les onomatopées (bof) et aussi les éléments phonétiques qui incluent l'ajout de caractères sans modification phonétique (parlé → parler), les liaisons (j'aime → zaime) et certaines variations (oui → ouip).

4.4.1 Typologie des formes du corpus

Nous regroupons les formes suivant la typologie proposée en quatre catégories tout en apportant une modification à la dernière catégorie (ajout) afin de distinguer certains éléments que nous considérons constituer seuls une catégorie à part entière. Ces éléments sont : a) les néologismes (création de mots, emprunts), b) l'expression graphique de sentiments (émoticônes [8]) et c) la répétitions des graphèmes ou signes (étirements graphiques). Nous résumons, ainsi, une

8. Nous utilisons ici le terme *émoticônes* et pas celui d'*émoji* qui est subséquent en ce qui concerne son utilisation dans les SMS et représente des idéogrammes qui illustrent des expressions du visage, des divers objets, des lieux, et des animaux.

classification typologique simplifiée de formes SMS issues du corpus alpes4science qui se divise en cinq grandes catégories (table 4.12). L'utilisation des formes ayant subi une substitution, une réduction et suppression phonétique ou graphique dégage de l'intention de l'auteur de véhiculer un message brièvement. Pour les autres graphies produites par exemple par ajout des caractères, nous considérons qu'elles apportent un style ludique sur ce discours, en jouant avec les mots au moyen de la néographie et de l'ajout d'éléments prosodiques graphiques (étirements graphiques) afin d'apporter une intonation. L'aspect morpho-syntaxique ne figure pas dans cette classification tel qu'il est décrit par Fairon *et al.* (2006) notamment avec l'inversion de la classe grammaticale nom-verbe (par exemple : *moi vais bouquiner un peu avant de dodo*), l'omission de mots grammaticaux (par exemple : *Viens de prendre tram b condilllac à tout de suite.*) et les phénomènes discursifs (par exemple la juxtaposition de phrases : *Viens de rentrer à la maison après recup des enfants. Pas possible discuter cadeaux. J'appelle demain. Gros bisous*). Cependant, ces phénomènes-là ne sont pas toujours facilement catégorisés, si nous y ajoutons la concaténation de phénomènes. Par ailleurs la distinction d'une faute de frappe ou d'erreur orthographique n'est pas également prévisible ce qui rend la détection et la catégorisation très difficiles et représente un vrai défi.

Phonétique	*Substitution*	demain → dem1n, ce → c, claqué → clakè
	Réduction	bises → biz, ne t'inquiète pas → tkt, cafétéria → cafet, faculté → fac, ce → c
	Ajout	verte → vertounette, tranquille → tranquillou, amoureux → zamoureux
Graphique	*Substitution*	est-ce-que → est ce que, pourrais → pourré, quand → kan, à plus → à +
	Réduction	problème → pb, bisous → bisou, j'aime → jaime, comment → cmt
	Suppression	ça → ca, là → la
Néologismes		news, kiffé, like, bday, amore, speed, lol, tepu
Émoticônes		:) :-) :d :p ;) ;-)
Étirements graphiques		coucouuuu, oui!!!!!!!!, arrrrrreeeeeette, supppeeerr, qui??!!!

Table 4.12 – Typologie SMS pour le corpus alpes4science

La table 4.13 illustre les quinze unités lexicales plus fréquentes du corpus de SMS qui ont une forme d'écriture SMS. La première observation que nous faisons est le fait que les mots les plus communs sont les plus irréguliers et aussi que ces formes abrégées correspondent aux unités lexicales qui sont courtes (maximum quatre caractères). Dans une deuxième phase, nous constatons que les graphies SMS, en accord avec la typologie de formes exposée précédemment, équivalent à la suppression graphique de signes diacritiques, à la réduction phonétique et graphique avec la construction consonantique.

	Graphie SMS	Graphie std.	Freq.
1	a	à	3238
2	c	c'est	1258
3	pr	pour	778
4	pa	pas	641
5	la	là	517
6	je	je [/ne/]	473
7	sa	ça	399
8	meme	même	392
9	ke	que	382
10	g	j'ai	354
11	ds	dans	326
12	t	tu es	325
13	d	de	297
14	ms	mais	289
15	etre	être	279

Table 4.13 – Graphies SMS plus fréquentes dans le corpus

4.5 Conclusion

L'analyse que nous venons d'exposer portant sur les éléments caractéristiques du corpus de SMS avec la typologique de formes, les principales caractéristiques lexicométriques aussi bien la description de données socio-démographiques fournies par les participants du projet nous ont permis de établir la définition du corpus et ce que le corpus représente.

La distribution de données socio-démographiques nous a permis d'apprendre sur les pratiques des participants concernant l'usage qui est fait du SMS et dessiner un portrait de son besoin communicatif en distinguant notamment le désir de partager et échanger des messages dans la vie quotidienne (demander et donner des nouvelles, transmettre une information, urgence, communiquer de façon discrète, etc.). Nous avons dressé, également, le profil de participants pour arriver à conclure que la majorité des participants sont de sexe féminin, jeunes âgés de 18 à 25 ans et des personnes ayant suivies des études correspondant aux cycles inférieur et supérieur originaires majoritairement du sud-ouest du pays.

Dans une deuxième phase de notre analyse nous avons évoqué dans ce chapitre l'analyse lexicométrique afin de détecter le niveau de diversité que les messages du corpus représentent sur l'ensemble du corpus transcrit mais aussi en se basant sur des partitions du corpus en fonction de différentes tranches. Ainsi, la richesse lexicale se présente évolutive pour la population âgée de plus de 35 ans et comporte une forte utilisation de formes verbales et pronominales. A la fin, nous remarquons le caractère affectif, informatif, ludique et organisationnel des messages du corpus ; il s'agit d'une conclusion qui est issue de l'analyse que nous avons menée sur les n-grammes les plus fréquents figurant dans l'ensemble du corpus.

La dernière analyse porte sur la classification typologique qui fournit l'observation des caractéristiques fondamentales du langage de SMS telles qu'elles figurent au travers de la variété lexicale et la diversité de formes. Il s'agit d'un produit de la créativité, de l'imprévisibilité lexicale et de l'absence normative dont les personnes font usage lors de la rédaction de leurs

messages.

Ce chapitre marque la transition vers la problématique centrale de notre thèse s'articulant autour de la question suivante : Comment peut-on traiter de façon automatique le langage SMS, et par extension toute autre type de message court, qui intègre tant de diversité et de libre créativité lexicale ?

Chapitre 5

Un système hybride pour la normalisation de SMS

Sommaire

5.1 Introduction **93**
5.2 Motivations **94**
5.2.1 Revue de travaux antérieurs 95
5.2.2 Problèmes de tokenisation pour la normalisation 97
5.3 Approche de normalisation **101**
5.3.1 Représentation intermédiaire 102
5.3.2 Modèle de Traduction Automatique 110
5.4 Évaluation **122**
5.4.1 Les résultats d'évaluation 123
5.4.2 Évaluation du système sur le corpus 88milSMS 129
5.5 Conclusion **130**

5.1 Introduction

Les caractéristiques des SMS que nous avons exposées au travers de l'analyse et la typologie de pratiques langagières (partie 4.4) nous montrent la présence importante de typologies

particulières, d'abréviations, de substitution phonétiques et graphiques, d'émoticônes et de structures qui échappent à toute convention ou norme. Ces caractéristiques, en effet, constituent un obstacle face à tout outil lié au traitement du langage destiné à traiter et analyser un langage normé. Le prétraitement des messages, afin de produire des formes lexicales plus normées, pourrait améliorer significativement les performances de ces outils et contribuer à l'extraction d'informations. Ce chapitre est dédié à la conception et à la mise en œuvre d'un modèle hybride en vue de la normalisation lexicale de messages courts bruités. Il est testé et entraîné sur des SMS issus du projet alpes4science que nous avons exposé dans le chapitre 3.

Le modèle de normalisation en deux étapes est fondé sur une approche symbolique et statistique. La première partie vise à produire une représentation intermédiaire du message SMS par l'application de grammaires locales, tandis que la deuxième utilise un système de traduction automatique à base de règles pour convertir la représentation intermédiaire vers une forme standard.

Le chapitre est structuré comme suit : la première partie 5.2 contient les préliminaires qui traitent les motivations de la nécessité de la normalisation des messages bruités en exposant un ensemble de travaux antérieurs et les problèmes liés à la tokenisation. La partie 5.3 est dédiée à l'architecture du modèle hybride proposé, divisé dans la phase de la représentation intermédiaire à travers les grammaires locales et la présentation de la mise en place du système de la traduction automatique. A la fin, la dernière partie 5.4 décrit la performance du modèle hybride via les métriques mises en place pour la réalisation d'évaluations sur un échantillon de test du corpus alpes4science et 88milSMS, non utilisés dans la phase d'entraînement.

5.2 Motivations

La plupart des messages courts présentent des différences significatives en comparaison avec les messages du langage standard puisqu'ils doivent contenir au maximum 160 caractères. Leurs

auteurs utilisent diverses formes pour abréger les mots dans l'objectif de gagner du temps tout en réduisant l'effort fourni. Comme Barasa et Mous (2009) le mentionnent, le langage utilisé dans la communication par SMS est particulièrement caractérisé par la création d'une nouvelle orthographe et d'une richesse créative qui échappe aux conventions.

Les approches fondées sur l'étiquetage morphosyntaxique de textes standard atteignent de hauts niveaux de précision dans des tâches liées au traitement du langage naturel. Cependant, les résultats sont significativement mitigés lorsque l'étiquetage est appliqué à des textes courts contenant du bruit (Gadde *et al.*, 2011). Une des contraintes que nous rencontrons avec les systèmes de TAL est la graphie particulière des mots SMS (fusionnement et phonétisation de mots, formes abrégées imprévisibles, suppression de caractères, manque de ponctuation, etc.). L'étiquetage morphosyntaxique constitue une étape fondamentale afin de pouvoir traiter davantage de données textuelles, comme dans la reconnaissance d'entités nommées, la traduction automatique, les systèmes de questions-réponses, l'extraction d'information etc. (Yvon, 2010).

La normalisation d'un texte devient une étape de prétraitement indispensable, de même, Sproat *et al.* (2001) indiquent notamment l'importance d'appliquer ce processus de normalisation avant tout autre traitement basique issu du TAL. La normalisation de SMS consiste à réécrire un message SMS en utilisant des formes lexicales plus conventionnelles afin de rendre ce message plus lisible aussi bien pour l'homme que pour la machine (Jose et Raj, 2014).

5.2.1 Revue de travaux antérieurs

Afin de surmonter le problème des particularités lexicales des SMS, plusieurs approches pour la normalisation lexicale des SMS ont vu le jour au cours des dernières années. Étant donné un texte $T = S_1, S_2, ..., S_n$, la tâche de normalisation lexicale est de trouver pour chaque token S_i hors-vocabulaire (*OOV : Out-Of-Vocabulary*) un token correspondant en forme standard (*IV : in-vocabulary*), par exemple : *cmb* $\rightarrow$ *combien*. Pour mieux visualiser le problème de normalisation, nous empruntons la formulation initialement introduite par Shannon (1948) afin

de décrire le processus de communication à travers un canal de communication bruité : Supposons que le texte mal-formé est T et que sa forme standard est S, l'approche de normalisation vise à trouver $\operatorname{argmax} P(S|T)$ en calculant $\operatorname{argmax} P(T|S)P(S)$, dont $P(S)$ est généralement un modèle de langage et $P(T|S)$ est un modèle d'erreurs Han et Baldwin (2011).

$$S_{max} = argmax P(S|T) \tag{5.1}$$

$$= argmax \frac{P(T|S)P(S)}{P(T)} \tag{5.2}$$

Les méthodes à base de canaux bruités (*noisy channel model*) étaient les premières à être appliquées dans la Communication Médiée par Ordinateur (CMO). Toutanova et Moore (2002) utilisent les similarités de prononciation entre les mots dans le but d'améliorer la correction orthographique. Choudhury *et al.* (2007) proposent un modèle au niveau du mot pour chaque mot en langue standard et produisent, à partir d'un modèle de Markov caché, toutes les variations possibles en langage SMS. De leur côté, Han et Baldwin (2011) exploitent une méthode à base de cascades qui détecte les mots mal-formés et génère des candidats à partir de similitudes morphophonémiques. Dans le cadre de la normalisation de tweets en anglais, Kaufmann et Kalita (2010a,b) utilisent un modèle qui fait passer les messages par une phase de prétraitement afin d'enlever le bruit et par la suite alimentent un modèle de traduction automatique pour le convertir en anglais standard. Aw *et al.* (2006) considèrent la normalisation comme un problème de traduction et proposent un outil de traduction automatique au niveau des phrases. La technique de la reconnaissance de la parole a été appliquée par Kobus *et al.* (2008), dans une première étape, pour décoder la représentation phonétique d'un mot en forme écrite. Avec une autre approche, Beaufort *et al.* (2010a) utilisent une méthode fondée sur les automates d'états finis tout en combinant la traduction automatique et les canaux bruités. Un des travaux les plus récents est le modèle qui a été proposé par Jose et Raj (2014) basé sur trois types de canaux (abréviations, graphèmes, phonétiques). Le modèle consiste à passer une phrase par quatre bases de données différentes pour identifier et corriger les mots inconnus en

donnant à l'utilisateur la possibilité de choisir parmi les meilleurs candidats.

Nous distinguons deux approches pour la transcription automatique de SMS en français (Kobus *et al.*, 2008, Beaufort *et al.*, 2010a). Comme Beaufort *et al.* (2010a) le mentionnent, les analyses de normalisation produites par leur système démontrent que lors de la normalisation le système n'est pas capable de traiter les erreurs liées au contexte qui concernent le genre (choix féminin ou masculin), le nombre (singulier ou pluriel), la personne (pronom personnel) ou le temps du verbe. Ils soulignent, d'ailleurs, que selon Kobus *et al.* (2008) les modèles basés sur les n-grammes ne sont pas capables de traiter ce type d'erreurs. Le modèle hybride que nous proposons permet, à l'aide de règles de transfert couplées avec les dictionnaires morphologiques, de résoudre ce type d'erreurs. Lopez *et al.* (2015) ont mis au point une méthode pour classifier de mots inconnus contenus dans les SMS afin d'identifier automatiquement la créativité lexicale et améliorer les approches basées sur les dictionnaires. Selon Lopez *et al.*, cette étape est fondamentale pour des tâches telles que la normalisation automatique de SMS en français standardisé. Tarrade (2017), décrit un système dédié à la normalisation de SMS issus du corpus 88milSMS et de tweets. La méthode de normalisation détermine si une unité lexicale est standard ou non et continue une chaîne de traitement basée sur une liste de candidats. Le meilleur candidat est, par la suite choisi, à partir d'un score qui dépend du phénomène linguistique qui l'a rendu non standard et de son contexte dans le texte.

5.2.2 Problèmes de tokenisation pour la normalisation

Le processus de tokenisation des langues alphabétiques se définit par la division de séquences de caractères en phrases et de phrases en tokens. Comme token nous considérons les mots, les chiffres et toute sorte de marqueur de ponctuation. Grâce à Palmer, 2000 nous disposons d'une définition simple du processus de la tokenisation d'un texte sans prendre en compte les marqueurs de ponctuation ou les chiffres : *"La tokenisation est le processus de segmentation d'une séquence de caractères dans un texte en localisant les limites de mots, les points où un*

mot se termine et un autre commence. Pour les besoins de la linguistique informatique les mots ainsi identifiés sont souvent appelés tokens. Dans les langues écrites où aucune limite de mots n'est explicitement marquée dans le système d'écriture, la tokenisation est également connue comme la segmentation de mots".

L'importance de ce processus pour les applications du TAL tels que l'étiquetage morpho-syntaxique, les analyseurs, l'extraction de mots clés, les moteurs de recherche, la normalisation du texte etc. s'identifie sur le fait qu'elles traitent des phrases et des mots. La plupart des applications de tokenisation utilisent une méthode simple qui réalise une segmentation de mots par blancs comme l'espace et les signes de ponctuation (Schmid, 2007).

En ce qui concerne les langues alphabétiques, chaque mot est souvent entouré par des espaces et parfois suivi ou précédé par des signes de ponctuation (le point, le point d'interrogation, le point d'exclamation, la virgule, le point-virgule, les deux-points, les points de suspension, les parenthèses, les crochets, les guillemets, le tiret) [1]. Nous pourrions considérer que la méthode de tokenisation la plus simple est de segmenter une séquence de caractères à la position des espaces et de supprimer toute sorte de signe de ponctuation afin d'obtenir une séquence de tokens mais cette méthode naïve manque toutefois de précision. La tokenisation des langues alphabétiques cache en réalité des problèmes beaucoup plus complexes. Nous empruntons par la suite une brève catégorisation des problèmes de tokenisation pour les langues alphabétiques émanant de (Palmer, 2000) et (Schmid, 2007).

- **Le point :** tous les points ne définissent pas la fin de phrases puisqu'ils servent à marquer les abréviations et les sigles. Exemples : G.D.F. → Gaz de France, s. o. → sans objet, n. a. → non applicable
- **Les autres signes de ponctuation :** les plus problématiques semblent être les " : " et le " ; ". Le point-virgule car il sert à séparer des propositions indépendantes mais, il peut être aussi séparateur dans une énumération. De même, les deux-points séparent

1. Martin Riegel, Jean-Christophe Pellat et René Rioul, Grammaire méthodique du français, Paris, Presses Universitaires de France, coll. « Quadrige », septembre 2009, 1108 p.

deux phrases et parfois ils sont un signe de ponctuation de la phrase interne.

- **Les mots composés détachés et à trait d'union :** même s'il a été mentionné que les tokens ne contiennent pas d'espaces, pour certaines applications il est préférable de traiter certains mots composés comme un seul token. Les prépositions complexes comme "en raison de", les composés comme "pommes de terre", "après-midi", "au-dessus", " jusqu'à ce que" sont certains exemples de mots composés constituant un lemme à part entière qui combinent divers moyens et doivent être analysés comme un seul token.

- **Les formes clitiques :** à l'opposé de mots composés où l'on doit analyser un seul token nous trouvons les formes clitiques. Les formes clitiques comme "j'ai", "parlez-vous" et "mange-t-il ?" doivent être conçues comme plusieurs tokens.

- **La coupure de mot (césure) :** il s'agit de l'opération qui consiste à segmenter avec un trait d'union conditionnel en fin de ligne un mot qui n'entrerait pas dans la justification du texte. Nous trouvons trois raisons pour lesquelles le processus de coupure de mots n'est pas trivial, en effet, quand la fin de ligne finit avec un trait d'union, comme par exemple "pré" dans le mot "prétraitement", le trait d'union et la fin de ligne doivent être supprimés. Par contre pour le mot "syntaxico-sémantique" s'il se segmente en "syntaxico-" et "sémantique", seulement, la fin de ligne devrait être éliminée. Un cas est la séquence "syntaxico- et morphosémantique" où nous devons remplacer le trait d'union par la frontière du mot.

- **L'absence d'espace :** dans certains cas l'espace après un signe de ponctuation est omis, comme résultat dans des exemples comme "heures.Le" ou "cependant,il" il devrait être divisé en trois trois séparés. En tenant compte de l'information de la fréquence des mots à partir d'un corpus, il est possible avec des méthodes statistiques de désambiguïser ces cas-là.

- **La détection de fin de phrase :** même s'il est possible d'annoter correctement la plupart des fin de phrases avec une règle relativement simple en ajoutant une ponctuation de fin de phrase comme ".", " ?" et " !", les phrases sont souvent ambiguës. La fin de

phrase peut faire partie d'une abréviation ou d'un nombre ordinal. Pire encore, elle est aussi suceptible de faire partie de la ponctuation d'une phrase et d'une abréviation, en même temps que l'abréviation se trouve à la fin d'une phrase.

La tâche de tokenisation s'avère compliquée afin de délimiter les frontières d'un mot pour les langues standard alphabétiques. Quand nous pensons au langage SMS nous réalisons que la segmentation en tokens devient un réel défi car à ces problèmes standard s'ajoutent les problèmes propres aux SMS. L'imprévisibilité d'utilisation d'espaces, de caractères spéciaux et en générale de normes sont certains d'éléments qui définissent ce langage. Un mot SMS n'est pas toujours entouré d'espaces, les signes de ponctuation sont souvent omis et les caractères spéciaux, comme les émoticônes, sont fréquemment utilisés. Par la suite nous exposons les problèmes de tokenisation que nous avons identifiés :

- **Abréviations non standard de mots composés :** tokens qui empruntent les initiales de mots composés. Exemples : mdr → mort de rire, pdr → pété de rire.
- **Détection de frontières de mots :** la plupart du temps les signes de ponctuation sont omis à la fin de la phrase.
- **Omission d'espaces et de signes de ponctuation :** l'utilisation de formes abrégées favorise l'omission d'une apostrophe ou d'espaces entre deux ou trois mots en générant une ambiguïté sémantique. Exemples : ct → cet/cette/c'est, ya → y a.
- **Autres signes de ponctuation - émoticônes :** il s'agit d'un ensemble de signes de ponctuation et de lettres qui représente une expression graphique d'émotions. Exemples : <3 → cœur, xD → personnage riant.
- **Mélange de chiffres et de caractères :** les mots SMS mélangent souvent des chiffres et des caractères. Exemple : dem1 → demain.
- **Formes consécutives de ponctuation :** ces formes sont fréquemment utilisées pour exprimer l'étonnement, l'admiration, la pensée, la tristesse et la joie. Exemples : quoi ? ? ? ? ?, :)))))), non !!!!!!!.

5.3 Approche de normalisation

Notre démarche est inspirée des méthodes de traduction automatique, à la différence des travaux précédents Bangalore *et al.* (2002), Aw *et al.* (2006), Raghunathan et Krawczyk (2009), Kaufmann et Kalita (2010a,b), qui sont limités au traitement des variations lexicales et par conséquent ont besoin de grandes quantités de données d'apprentissage étiquetées. Nous proposons un modèle de normalisation en deux étapes à l'aide des approches symboliques et statistiques : la première partie vise à produire une représentation intermédiaire du message SMS par l'application de grammaires locales, tandis que la deuxième utilise un système de traduction automatique à base de règles pour convertir la représentation intermédiaire vers une forme standard (tableau 5.1).

SMS brut	`Coucouuuuu! Où t es? Kfé? :)) Bizz`
Représentation intermédiaire	`Coucou ! Où t es ? Kfé ? ***EMOTICON*** Biz`
Traduction automatique	`Coucou ! Où tu es ? Café ? ***EMOTICON*** Bise`

TABLE 5.1 – Exemple de normalisation

Ainsi la figure 5.1 illustre les cinq phases de traitement pour la normalisation de SMS. Le premier stade est dédié au *prétraitement* qui est une étape commune aux phases d'apprentissage et de test. Elle est constituée par :

— Le *nettoyage* (par exemple : suppression guillemets début et fin de phrase, élimination des mauvais candidats : 1 lettre/symbole, transcription impossible corpus référence).

— La *normalisation des caractères* (des espaces ou de caractères équivalents).

La deuxième phase concerne l'analyse typographique et néologique qui correspond à la définition de la typologie des graphies repérées dans le corpus de SMS telle que nous l'avons exposée dans la partie 4.4.1. Cette analyse se divise en deux étapes et elle est décrite dans la partie 5.3.1 comme *Représentation intermédiaire* :

— La *reconnaissance* des expressions graphiques des sentiments (émoticônes), la répétition

de graphèmes ou de signes (formes consécutives) et les mots inconnus.

— Le *traitement d'unités reconnus* précédemment.

La troisième phase correspond à la phase de traduction au travers du système de traduction automatique Apertium et la construction de ressources linguistiques. La quatrième phase s'occupe de la mise au format de sortie du système de traduction. Et la dernière phase concerne l'évaluation mise en place avec des métriques appropriées pour évaluer la qualité de la traduction produite.

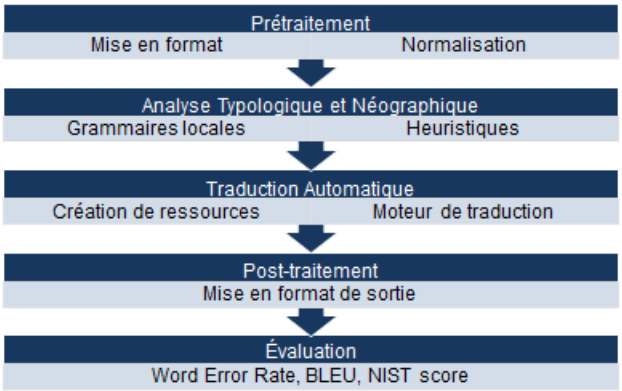

FIGURE 5.1 – Phase de traitement

5.3.1 Représentation intermédiaire

L'observation de particularités liées aux SMS nécessite l'utilisation de matériaux authentiques dans le but d'obtenir un point de vue plus objectif. Notre étude a comme point de départ le corpus de SMS, collecté et traité semi-automatiquement, dans le cadre du projet *alpes4science* (les détails concernant le projet ont été exposés dans le chapitre 3). Dans la base de données nous trouvons les SMS (anonymisés, alignés et transcrits en langue standard), le lexique (mots SMS avec leurs traductions en langue standard et leurs fréquences) et des informations variées sur les expéditeurs (âge, sexe, niveau d'études, langue maternelle, etc.).

La représentation intermédiaire du message se divise en deux processus et se base sur l'analyse typographique et néologique des SMS menée par (Anis *et al.*, 2004, Panckhurst, 2009, Fairon *et al.*, 2006, Véronis et Guimier de Neef, 2006). Tout d'abord, nous avons le processus de normalisation structurelle qui prend en charge la normalisation des séparateurs et le traitement des symboles de ponctuation ; pour accomplir cette partie nous faisons appel à l'utilisation d'heuristiques développées à partir de notre corpus d'apprentissage. Une deuxième étape a pour objectif de reconnaître et, par la suite, traiter les unités reconnues à l'aide de grammaires locales (Paumier, 2006). Cette étape est responsable du processus de normalisation pour la résolution des formes non ambiguës, les abréviations, la détection des émoticônes et des mots hors-vocabulaire, ainsi que le découpage en unités lexicales. Pour ceci nous avons conçu des réseaux de transition récursives (RTN : *Recursive Transition Networks*) appliqués en combinaison avec des dictionnaires électroniques du français standard et une base de connaissance regroupant des informations spécifiques aux mots SMS.

Pour la réalisation de cette étape nous avons utilisé le logiciel Unitex (Paumier, 2003). Il s'agit d'un logiciel libre pour l'analyse du langage naturel qui a été développé par Sébastien Paumier et l'équipe d'informatique linguistique du Laboratoire d'Informatique Gaspard-Monge (LIGM) à l'université Paris-Est Marne-la-Vallée. Le logiciel est en constant développement grâce à une communauté des linguistes et des développeurs. Unitex permet de traiter des textes en langues naturelles à l'aide des ressources linguistiques sous la forme de dictionnaires électroniques et de grammaires locales. Le concept de grammaire locale a été développé par Maurice Gross (Gross, 1993). La figure 5.2 illustre le graphe principal de la représentation intermédiaire qui regroupe un ensemble de sous-graphes que nous avons crée. Chaque sous-graphe réalise des opération spécifiques pour l'annotation d'expressions graphiques des sentiments (e), de répétition de graphèmes ou de signes (c), de mots inconnues (u) et de locutions et expressions (l).

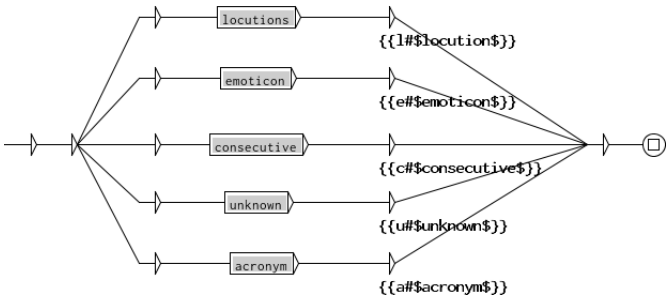

FIGURE 5.2 – Graphe principal de reconnaissance

5.3.1.1 Expression graphique de sentiments

L'expression graphique des sentiments (émoticônes) est une chaîne de caractères produite à l'aide d'un clavier qui peut être proposée pour imiter un visage exprimant une émotion particulière (Danesi, 2008). L'utilisation des émoticônes dans les interactions CMO représente une alternative à la forme verbale et constitue une représentation plus au moins imagée de la signification. Ces caractéristiques linguistiques sont codifiées sous forme de pictogrammes ou d'émoticônes pour représenter, dans la plupart des cas, des gestes, des expressions faciales et des éléments prosodiques puisque la CMO ne comporte pas ces caractéristiques (Amaghlobeli, 2012). Afin de pouvoir reconnaître toute forme d'émoticônes contenues dans les SMS, nous avons créé un dictionnaire électronique contenant plus de 300 entrées (par exemple : `o_o,.Emoticon+Meaning=surprise`). Ce dictionnaire a été utilisé pour développer une grammaire locale (figure 5.3) qui identifie et multiplie la capacité de reconnaissance des émoticônes contenues dans les SMS, par exemple : ***:)*** ***:))*** ***:) :D*** ***;) :) :D ;) :)) :D***

5.3.1.2 Locutions et expressions

Les locutions regroupent des unités qui ne sont pas actualisées individuellement, que ce soit des locutions verbales (*prendre des gants, faire du foin, casser la gueule, etc.*), adverbiales

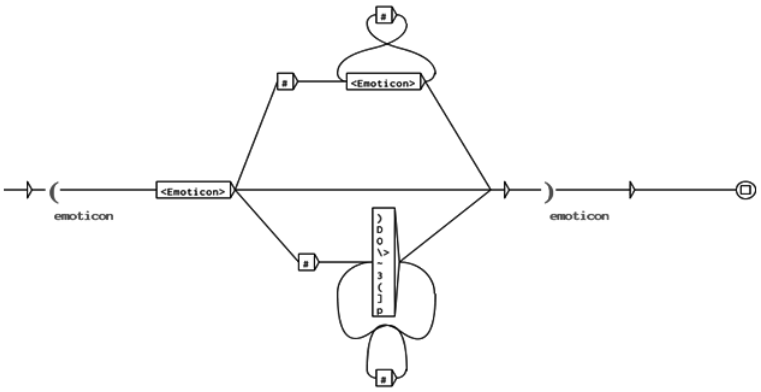

FIGURE 5.3 – Graphe de reconnaissance d'émoticônes

(*à tort, à la rigueur, etc.*), nominales (*bac à sable, maison de retraite, etc.*) et propositionnelles (*à cause de, à côté de, à partir de, etc.*). Nous avons également ajouté dans ce même graphe les suites des n-grammes les plus fréquentes (*je t'aime, j'y vais, etc.*).

La reconnaissance de ces expressions s'effectue à l'aide d'une grammaire locale (c.f. graphe en annexe E.1). Afin de parvenir à formuler cette grammaire locale nous nous sommes basés sur les n-grammes les plus fréquents du corpus afin de construire un dictionnaire des expressions avec les formes possibles de SMS (*a plus tard,à plus tard.LOC*). Le tableau 5.2 illustre un extrait du dictionnaire de locutions.

De plus, après une étude de cas de figure, nous avons construit des grammaires locales supplémentaires pour traiter la correction des erreurs fréquemment repérées contenant la préposition *à* couramment présente dans les messages sans l'accent grave, l'adverbe *là*, également sans accent, la restitution de l'article défini *le*, *la*. Nous avons isolé chaque groupe de transitions afin de présenter chaque cas.

La figure 5.4 montre l'appel au dictionnaire du tableau 5.2 en mode morphologique afin de renvoyer les formes canoniques correspondantes aux expressions reconnues dans le texte.

que ce moi,que ce est moi.LOC
que t soi,que tu sois.LOC
que t sois,que tu sois.LOC
que t soit,que tu sois.LOC
que t t,que tu te.LOC
que t te,que tu te.LOC
que tu ma,que tu m'as.LOC
que tu soi,que tu sois.EXP

TABLE 5.2 – Extrait du dictionnaire de locutions

Le mode morphologique consiste à délimiter une partie d'une grammaire avec des symboles spécifiques. La requête <LOC> utilise le dictionnaire de locutions (5.2). Les informations issues du dictionnaire du mode morphologique sont affectées à la variable *sms*. La variable de ce type est associée à la boite contenant un motif qui fait référence à des informations contenues dans un dictionnaire du mode morphologique. Dans la suite des chemins qui passent par la boite, on peut obtenir les informations attribuées à une entrée du dictionnaire, par exemple, la forme canonique (*$sms.LEMMA$*).

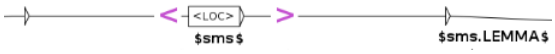

FIGURE 5.4 – Graphe avec mode morphologique

La figure 5.5 contient trois transitions : la première vise à reconnaître les formes qui commencent par la forme *j* suivies d'une apostrophe et d'un verbe, dont son initiale est une consonne, (par exemple : j'parle). La forme reconnue *j* est, par la suite, transformée en *je* accompagnée du verbe reconnu grâce à la variable d'entrée *V* (parenthèse en rouge) qui permet de sélectionner cette partie du texte à la sortie (*V*). La deuxième reconnaît n'importe quelle proposition (contenue dans la variable en rouge *pr*) suivie de *t* avec ou sans apostrophe, suivie d'un verbe dans la variable *V*, dont sa lettre initiale est une consonne (par exemple : ça t va).

La sortie de la transition donnera la proposition reconnue suivie de la forme *te* et du verbe (par exemple : ça t va → ça te va). La dernière transition commence avec un contexte négatif (crochet vert avec pont d'exclamation) pour exclure les propositions et toutes les formes verbales suivies de la forme *t'* et d'un verbe qui commence avec une consonne pour attribuer à la sortie la forme *tu* et le verbe reconnu dans la variable *V*.

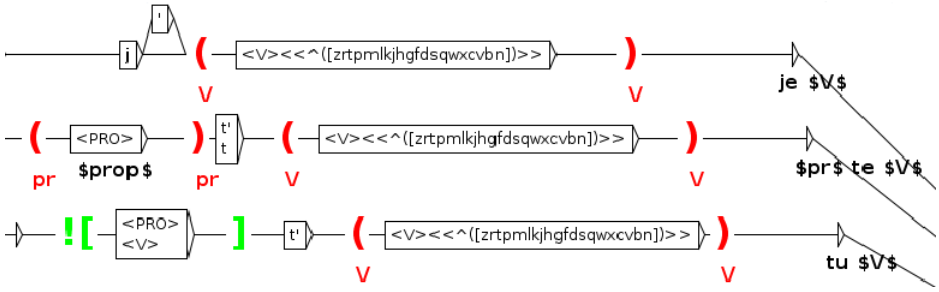

FIGURE 5.5 – Graphes de transitions de formes avec apostrophe

La figure 5.6 correspond au sous-graphe *hloc :LAWord*. Il s'agit d'un graphe qui comporte trois transitions et sert à reconnaître et corriger les formes qui contiennent l'article *l'* suivi d'une forme lexicale reconnue au travers du sous-graphe. Ainsi, la première reconnaît des formes nominales, adjectivales et verbales au féminin n'ayant pas comme lettre initiale une voyelle pour attribuer la forme *la* (par exemple : l'voiture → la voiture). La seconde reconnaît des formes nominales, adjectivales et verbales au masculin n'ayant pas comme initiale une voyelle pour attribuer la forme *le* (par exemple : l'verre → le verre). La dernière a été construite pour reconnaître les mots inconnus qui ont une consonne pour lettre initiale et leur attribuer, de facto, l'article *le* (par exemple : l'fer → le fer).

Le graphe de la figure 5.7 est le sous-graphe *:plus.grf* qui a été créé dans le but de reconnaître et normaliser les formes contenant le symbole + selon le cas. Il reconnaît les formes verbales suivies de + pour attribuer le mot *plus* (par exemple : ira + vite → ira plus vite) et les formes : *en* + → *en plus*. Il normalise, aussi, les différentes combinaisons possibles, contenant ce symbole, repérées dans le corpus *à plus tard* (++ → à plus tard, a + → à plus tard).

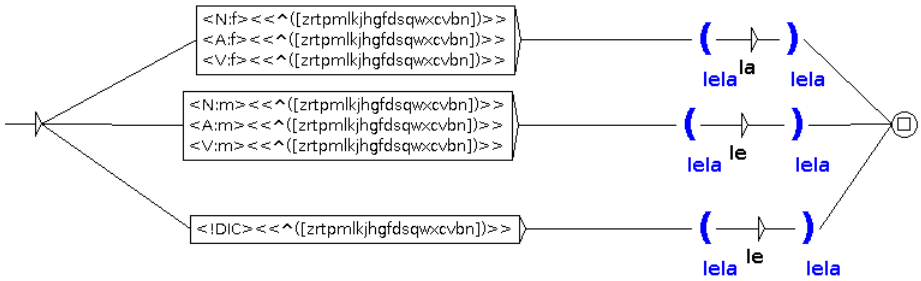

FIGURE 5.6 – Graphe de transitions pour les articles *le* et *la*

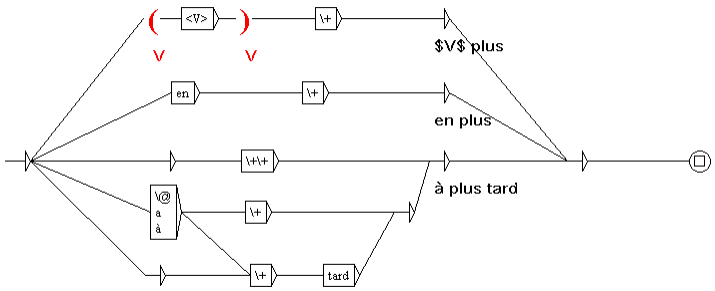

FIGURE 5.7 – Graphe de reconnaissance du caractère *+*

Les transitions de la figure 5.8 concernent la normalisation de la préposition *à*. La première transition consiste à détecter le caractère *a* suivi d'une forme verbale, stockée dans la variable *verbinf*, de l'indicatif et lui attribuer par la suite la forme *à* et le verbe dans la même forme (par exemple, a partir → à partir). De plus, la grammaire reconnaît les verbes transitifs et certaines locutions à partir de la liste qui est stockée dans la variable *verbtrans* suivis par la forme *a* (par exemple difficile à, avoir tendance à) et attribue la sortie de la variable et la forme *à*. Dans la dernière transition, en reconnaissant la typologie d'une heure du type *a* plus numéro plus *h* plus numéro (par exemple à *8h30*) la grammaire corrige la forme erronée.

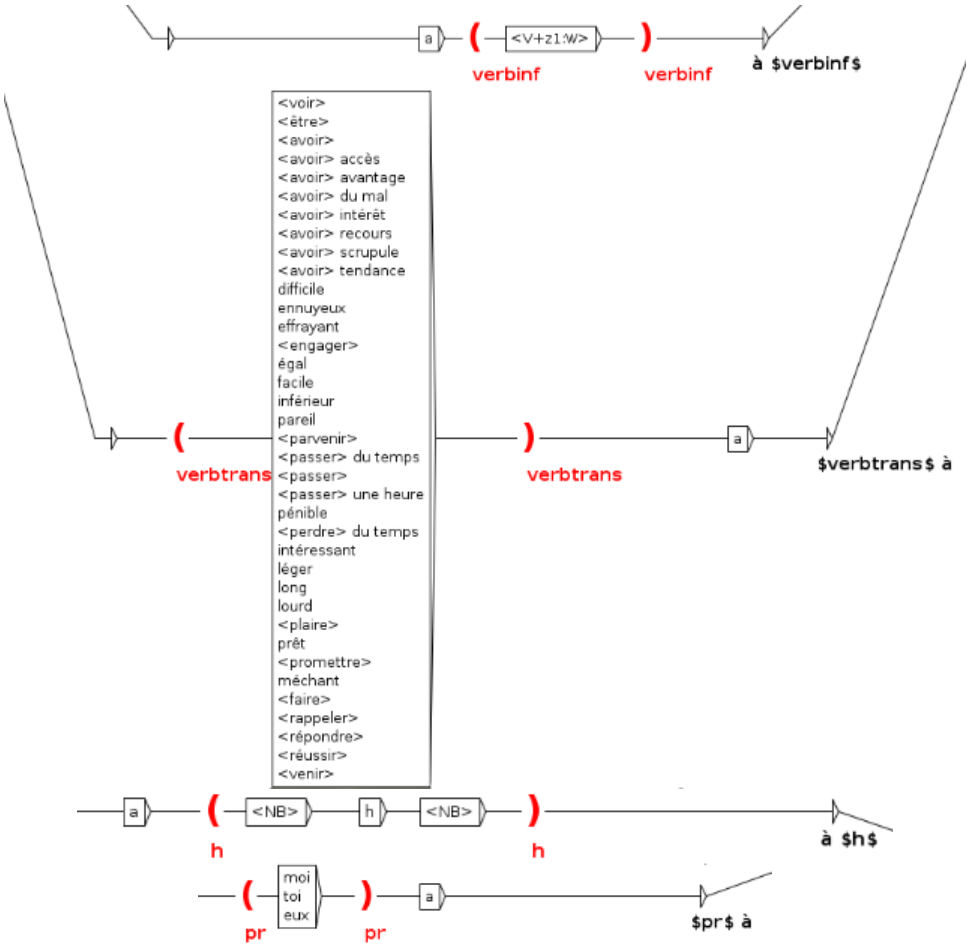

FIGURE 5.8 – Graphe de transitions pour la forme *à*

5.3.1.3 Répétition de caractères

La répétition de caractères constitue un phénomène très courant dans la CMO dans le but de souligner les sentiments exprimés (Brody et Diakopoulos, 2011). De leur côté Kalman et Gergle (2014) en étudiant la fréquence et l'utilisation de répétitions de lettres dans la communication par courriel, remarquent que les répétitions tendent à imiter un morphème étendu dans la conversation parlée, à donner de l'emphase ou à imiter des sons. La reconnaissance d'unités

lexicales hors-vocabulaire contenant des caractères répétés (par exemple : *cooool*→*cool*), s'effectue à l'aide d'une grammaire locale (figure 5.9) qui identifie les mots inconnus avec une ou plusieurs répétitions de lettres : cooool (oui), elle (non). Le graphe s'appuie sur l'utilisation des dictionnaires DELAF de l'anglais (Monceaux, 1995) et du français (Courtois, 1990), ainsi que sur un dictionnaire complémentaire (≈ 50 000 entrées) constitué par des mots connus qui ont des caractères répétées consécutives. Ce dernier dictionnaire a été élaboré à partir d'un corpus de résumés courts en français issu de DBpedia [2].

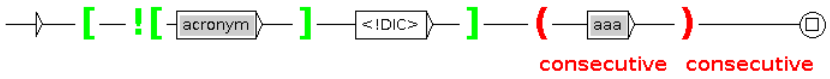

FIGURE 5.9 – Graphe de reconnaissances d'unités contenant des formes consécutives

5.3.1.4 Mots hors vocabulaire

Il s'agit d'un graphe qui reconnaît les mots hors vocabulaire se basant sur les dictionnaires fournis par Unitex (dictionnaires Unitex pour la langue française et dictionnaires de SMS), de telle sorte, que tous les mots qui ne sont pas inclus dans les dictionnaires appliqués sont considérés comme inconnus.

5.3.2 Modèle de Traduction Automatique

La transformation de la représentation intermédiaire vers la forme standard du message s'effectue à l'aide d'un système de traduction libre basé sur des règles de transfert lexical.

2. short_abstracts_fr ≈ 50 millions de mots, projet DBpedia http ://fr.dbpedia.org

Apertium est une plateforme libre de traduction automatique, connue initialement pour traiter des paires de langues assez proches. La conception de cette plateforme est fondée sur la traduction mot à mot avec une désambiguïsation lexicale (étiquettes morphosyntaxiques), un traitement lexical robuste et structuré basé sur des règles simples bien formulées (Forcada *et al.*, 2009).

5.3.2.1 Apertium : une plateforme pour la traduction automatique

Apertium [3] est une plateforme gratuite/code source ouvert dédiée à l'implémentation des systèmes de traduction automatique basés sur des règles (Forcada *et al.*, 2009). Elle a été initialement conçue pour la production de traductions de paires de langues proches, comme par exemple : espagnol-catalan, espagnol-galicien. Cependant, actuellement elle est étendue à la traduction d'autres paires de langues qui ne sont pas vraiment proches (espagnol-français) (Ramırez-Sánchez *et al.*, 2006). Ainsi, nous trouvons actuellement dans la plateforme 43 paires de langues disponibles en version stable : http ://wiki.apertium.org/wiki/Main_Page.

La plateforme fournit un moteur superficiel de traduction automatique modulaire, des données linguistiques et une variété d'outils pour gérer ces données linguistiques nécessaires à la traduction. Le moteur Apertium est défini, selon Forcada *et al.* (2009), Ramırez-Sánchez *et al.* (2006), Tyers et Sánchez-Martínez (2010), comme un pipeline, une suite, comprenant les étapes ou les modules suivants (figure 5.10) :

— Le *déformateur* qui consiste à segmenter le texte qui sera traduit du format d'information. Autrement dit, il s'agit d'ignorer lors du processus de traduction des balises HTML ou d'autres informations liées à la mise en page des documents issus de divers outils de traitement de texte. Ainsi, une fois le format capturé le reste des modules pourront le considérer comme un espace entre les mots.

— L'*analyseur morphologique* est dédié à effectuer la partie de tokenisation du texte en

3. http ://www.apertium.org

formes et par la suite attribuer à chaque forme les informations lexicales appropriées, c'est-à-dire le lemme, la catégorie lexicale et l'information morphologique de flexion.

— L'*étiquetage morphosyntaxique* définit la forme lexicale la plus appropriée qui correspond à la forme ambiguë du texte en faisant appel au modèle de Markov caché de premier ordre.

— Le *transfert lexical* est le module responsable d'associer à chaque forme lexicale de langue source la forme lexicale de la langue cible correspondante après l'avoir cherchée dans le dictionnaire bilingue au format XML.

— Le *transfert structurel* se réalise en même temps que le transfert lexicale pour détecter les motifs de formes lexicales qui doivent être traités pour la réorganisation de mots, l'accord, etc. à l'aide d'automates d'états finis.

— Le *générateur morphologique* remplit la tâche de fournir la forme en langue cible pour chaque lemme en langue cible, il donne sa forme fléchie.

— Le *post-générateur* réalise selon la langue cible des opérations orthographiques comme par exemple *de+le = du* ou encore marque l'élision de certaines voyelles finales (*si+il = s'il*, *le+homme = l'homme*).

— Le *réformateur* a pour tâche de réaliser la fonction inverse de l'étape du *déformateur*, c'est-à-dire restaurer le format initial du texte.

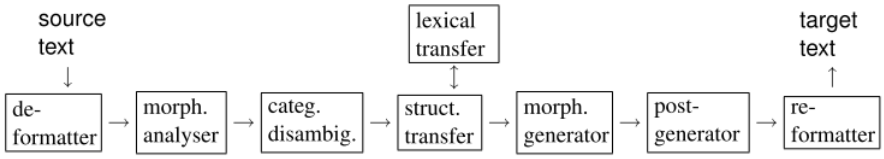

FIGURE 5.10 – Les huit modules du système Apertium dans Ramırez-Sánchez *et al.* (2006)

En ce qui concerne les ressources linguistiques, nous résumons la nécessité, pour la partie qui concerne l'analyse morphologique, 1) un dictionnaire monolingue de la langue source, 2)

un dictionnaire bilingue pour le module du transfert lexical, 3) un dictionnaire monolingue qui sera utilisé lors de la tâche du générateur morphologique, et 4) les règles de transfert afin de réaliser le transfert structurel.

1. **Dictionnaire monolingue**

   Le dictionnaire monolingue correspond à la langue source ou la langue cible de la traduction, les deux sont nécessaires pour alimenter le module dans les différentes étapes et produire la traduction. Dans les dictionnaires monolingues, comme nous pouvons le constater sur l'exemple d'une entrée du dictionnaire monolingue français (figure 5.12), chaque entrée est accompagnée d'un exemple de flexion pour les mots réguliers et irréguliers, simples et composés. Les informations lexicales contenues dans ces exemples de flexion permettrons de fléchir le mot en accord avec l'exemple de flexion attribué.

```
<pardef n="abeille__n">
  <e><p><l></l><r><s n="n"/><s n="f"/><s n="sg"/></r></p></e>
  <e><p><l>s</l><r><s n="n"/><s n="f"/><s n="pl"/></r></p></e>
</pardef>
```

FIGURE 5.11 – Exemple de flexion "abeille" dans le dictionnaire monolingue *fra.dix*

   L'exemple de flexion "abeille" que nous avons cité (figure 5.11) est utilisé pour former la flexion de noms en français sans ajout de caractère pour former le féminin singulier mais avec l'ajout du caractère "s" pour former le pluriel. Ainsi, des mots comme *voiture* et *voitures* peuvent être définis grâce à l'exemple de flexion abeille.

   La figure 5.12 montre un extrait du dictionnaire où nous pouvons constater les quatre exemples de flexion qui se situent dans la balise *par* pour *paradigme*. Nous pouvons également constater que la balise *i* englobe le radical pour chaque lexème, commun dans toutes les formes fléchies.

   Une fois les deux dictionnaires monolingues, c'est-à-dire celui de la langue source et de la langue cible, enrichis avec la traduction correspondante, respectivement, dans le système de traduction automatique, il faut lier ces deux dictionnaires en ajoutant l'entrée

```
<e lm="école">
    <i>école</i>
    <par n="abeille__n"/>
</e>
<e lm="écologiste">
    <i>écologiste</i>
    <par n="artiste__n"/>
</e>
<e lm="économe" a="ariel">
    <i>économe</i>
    <par n="académique__adj"/>
</e>
<e lm="économiser">
    <i>économis</i>
    <par n="abaiss/er__vblex"/>
</e>
```

FIGURE 5.12 – Entrées du dictionnaire monolingue *fra.dix*

correspondante dans le dictionnaire bilingue (Espla-Gomis *et al.*, 2011).

2. **Dictionnaire bilingue**

   Le dictionnaire bilingue contient les équivalences des mots et/ou des expressions entre deux langues différentes dans un format XML. A chaque entrée du dictionnaire il est possible d'attribuer une ou plusieurs traductions. Il ne s'agit pas d'un dictionnaire bilingue unidirectionnel, autrement dit, une entrée peut être soit source soit cible pour la traduction. Ceci est indiqué à la figure 5.13 dans la balise *r='LR'* pour une traduction de gauche à droite et *r='RL'* pour une traduction de droite à gauche, le cas contraire aucune mention est ajoutée.

   Pour chaque entrée nous remarquons la balise *<e>...</e>* et la balise * ... * qui signifie paragraphe, les mots de la langue source sont marqués par *<l>...</l>* tandis que ceux de la langue cible par *<r>...</r>*. Nous notons également des éléments indiquant la catégorie grammaticale, le genre et le nombre.

3. **Règles de transfert**

   Les règles de transfert correspondent aux règles qui sont mobilisées lors de la phase du transfert structural, il s'agit de la façon dont les mots sont réordonnés dans les phrases (par exemple la construction nom adjectif pour le français devient adjectif nom

```
<e r="LR"><p><l>concours<s n="n"/><s n="m"/><s n="sp"/></l>
            <r>certamen<s n="n"/><s n="m"/><s n="ND"/></r></p></e>
<e r="LR"><p><l>concours<s n="n"/><s n="m"/><s n="sp"/></l>
            <r>concurso<s n="n"/><s n="m"/><s n="ND"/></r></p></e>
<e r="RL"><p><l>concours<s n="n"/><s n="m"/><s n="sp"/></l>
            <r>certamen<s n="n"/><s n="m"/><s n="pl"/></r></p></e>
<e r="RL"><p><l>concours<s n="n"/><s n="m"/><s n="sp"/></l>
            <r>certamen<s n="n"/><s n="m"/><s n="sg"/></r></p></e>
<e r="RL"><p><l>concours<s n="n"/><s n="m"/><s n="sp"/></l>
            <r>concurso<s n="n"/><s n="m"/><s n="pl"/></r></p></e>
<e r="RL"><p><l>concours<s n="n"/><s n="m"/><s n="sp"/></l>
            <r>concurso<s n="n"/><s n="m"/><s n="sg"/></r></p></e>
<e><p><l>confirmer<s n="vblex"/></l>
      <r>confirmar<s n="vblex"/></r></p></e>
<e r="RL" ><p><l>confirmer<s n="vblex"/></l>
             <r>aseverar<s n="vblex"/></r></p></e>
<e r="RL"><p><l>confirmer<s n="vblex"/></l>
            <r>refrendar<s n="vblex"/></r></p></e>
```

FIGURE 5.13 – Entrées du dictionnaire bilingue *fra-es.dix*

en anglais : fruits rouges → red fruits) et aussi de la manière dont l'accord en genre, le singulier et le pluriel, etc. sont gérés. Les règles peuvent aussi être utilisées pour insérer ou supprimer des éléments lexicaux.

5.3.2.2 Induction de dictionnaires pour la normalisation de SMS

Après avoir exposé les éléments principaux du système de traduction, nous allons présenter l'adaptation du système, avec l'induction des dictionnaires, pour la réalisation de la traduction de *sms français* vers le *français standard*. Le système nécessite deux types de ressources linguistiques : deux dictionnaires morphologiques monolingues (*smsfra, fra*) et un dictionnaire bilingue (*smsfra-fra*). Le dictionnaire *smsfra* est une extension du dictionnaire français (*fra*) fourni par Apertium avec des ajouts supplémentaires, par exemple pour le mot *smsfra impec* figurant dans le dictionnaire monolingue :

<e lm="impec">impec<par n="impec__adj"/></e>

En accord avec la méthode d'induction automatique pour l'induction de dictionnaires bilingues du projet ReTraTos (Caseli et Nunes, 2007), la technique d'induction de dictionnaires bilingues consiste, tout d'abord, à utiliser deux corpus parallèles. Pour notre cas, il s'agit

d'utiliser le corpus de SMS transcrits manuellement et celui de SMS bruts issus du projet alpes4science. Le corpus parallèle contient 14 174 paires de phrases parallèles, correspondant aux 2/3 du corpus des SMS (alpes4science), qui constituent ainsi nos exemples de traduction. A l'aide d'un outil fourni par le système de traduction (Caseli et Nunes, 2007), dans une première phase l'étiquetage morphologique pour chaque unité (lemmatisation, catégories lexicales, informations morphologiques flexionnelles) a été effectué grâce au dictionnaire morphologique monolingue fourni par le système pour la langue française. Ce dictionnaire construit à partir de ressources linguistiques [4] contient environ 28 700 lemmes y compris des unités multi-mots et 275 000 formes (Tyers et Sánchez-Martínez, 2010). Dans une deuxième phase, l'outil a procédé à la désambiguïsation syntaxique afin d'attribuer les étiquettes morphosyntaxiques pertinentes fondés sur le modèle statistique de Markov caché (MMC).

Pour construire le dictionnaire bilingue *smsfra-fra* nous avons utilisé deux corpus constitués chacun de 14 174 SMS [5]. Le premier corpus contient les SMS qui ont été transcrits manuellement, tandis que le deuxième est la représentation intermédiaire des SMS en format brut (*cf.* 5.3.1). Les deux corpus ont été ensuite étiquetés morphosyntaxiquement avec Apertium [6] et alignés à l'aide de GIZA++ (Och et Ney, 2003), un outil d'alignement de mots combinant des modèles statistiques et heuristiques.

La table 5.3 nous permet d'observer l'exemple de traduction de *smsfra-fra* aligné pour chaque forme figurant dans la phrase "*non juste pour news comme ça. + tard alors. biz*". Le chiffre à coté de chaque entrée indique la place correspondante de l'entrée dans la phrase initiale, de ce fait les entrées qui n'ont pas d'équivalence sont marquées avec vide (ex. news / avoir() des() nouvelles()). Nous trouvons, également, l'étiquette morphosyntaxiuque attribuée à chaque entrée à la forme canonique (ex. de<pr>+le<det><def><mf><pl>). En annexe F nous pouvons trouver un extrait du dictionnaire.

4. *Eurfa* : http ://kevindonnelly.org.uk/eurfa/
5. Ceci correspond aux deux tiers du corpus alpes4science.
6. Nous avons utilisé les dictionnaires monolingues *smsfra* et *fra* avec le modèle de probabilité pour l'étiquetage morphosyntaxique du français fourni par Apertium.

FIGURE 5.14 – Illustration d'alignement mot-à-mot

L'étape suivante consiste à compiler ces deux corpus alignés et à fournir les traductions possibles pour la langue cible, tout en attribuant des caractéristiques pour chaque entrée (lemme, étiquette morphologique etc.). Ensuite, nous avons le fusionnement de ces deux dictionnaires de traduction, la génération des entrées bilingues et le traitement des différences morphosyntaxiques (pour plus de détails sur la procédure d'induction des dictionnaires bilingues, voir Caseli *et al.* (2006)). Finalement, une validation manuelle pour corriger d'éventuelles erreurs lors de la phase d'alignement a été effectuée. En annexe nous trouvons des extraits du dictionnaire monolingue *smsfra* et *fra* et bilingue *smsfra-fra*.

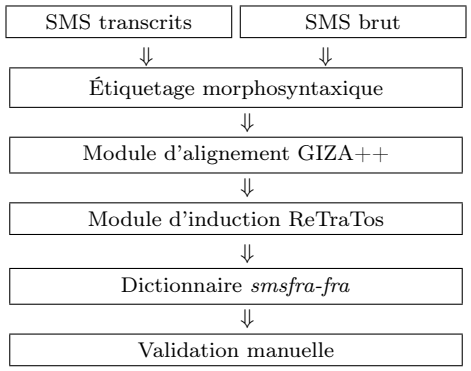

FIGURE 5.15 – Schéma d'induction du dictionnaire bilingue

Le schéma de la figure 5.15 synthétise les étapes de traitement pour l'induction du dictionnaire bilingue *smsfra-fra*. Exemple d'entrée dans le dictionnaire *smsfra-fra* :

<e> <l>impec<s n="adj"/></l><r>impeccable<s n="adj"/></r> </e>

Exemple de SMS :
non juste pour news comme ça . + tard alors . biz
Exemple de transcription :
non juste pour avoir des nouvelles comme ça . à plus tard alors . bise
Alignement avec GIZA++ :
Sentence pair (2) source length 15 target length 12 alignment score : 5.04267e-09
non juste pour news comme ça . + tard alors . biz
non (1) juste (2) pour (3) avoir () des () nouvelles () comme (5) ça (6) . (7) à () plus (8) tard (9) alors (10) . (11) bise (12)
SMS étiqueté :
non<adv> juste<adv> pour<pr> news comme<rel><adv> ça<prn><tn><nt><sg> .<sent> tard<adv> alors<adv> .<sent> biz
SMS transcrit étiqueté :
non<adv> (1) juste<adv> (2) pour<pr> (3) avoir<vblex><inf> () de<pr>+le<det><def><mf><pl> () nouvelle<n><f><pl> (4) comme<rel><adv> (5) ça<prn><tn><nt><sg> (6) .<sent> (7) à<pr> () plus<adv> () tard<adv> (8) alors<adv> (9) .<sent> (10) bise (11)

TABLE 5.3 – Exemple de traduction sans des étiquettes morphosyntaxiques

5.3.2.3 Désambiguïsation de formes polysémiques

Selon Bréal (1897) qui le premier a introduit le terme *polysémie*, il s'agit de l'association d'un mot à plus d'une signification (par exemple : *bureau* définit un lieu de travail, l'ensemble des personnes travaillant dans ce lieu, un meuble, etc.). Cette relation est très fréquente dans le langage humain, et naturellement, aussi présente dans le langage SMS. Systématiquement il combine des significations différentes mais liées sous le même lexème. Plus précisément, pour le langage SMS, une unité SMS portant du sens devient polysémique non seulement lorsque elle correspond à deux ou plusieurs significations mais aussi lorsque elle équivaut à deux ou plusieurs traductions différentes, dont le signifié et le signifiant sont différents, donc, non homonymiques.

Exemple de polysémie SMS :

pe → peu, peut-être, peux, peut

Avec la création du dictionnaire bilingue certaines entrées ont pu bénéficier d'une ou plusieurs traductions possibles (par exemple : ct → c'est, cette). En effet lors du fusionnement des deux dictionnaires monolingues le système à fait le choix de sélectionner l'occurrence la plus fréquente par l'application d'un seuil de fréquence afin de contraindre la création des multi-mots. Ainsi, une entrée à laquelle correspond plus d'un mot ne va être créée que si elle figure n fois Caseli et Nunes (2007). La sélection lexicale s'effectue après la recherche dans le dictionnaire bilingue (traduction de mots), avant l'application de règles de transfert structurel. Elle consiste à choisir parmi plusieurs traductions de la langue source, avec la même morphosyntaxe, la traduction la plus adaptée parmi eux dans la langue cible. La conception des règles s'avère une tâche assez minutieuse et coûteuse en terme de temps (avec la construction de grammaires par contraintes ou des règles de transfert) ainsi un nombre restreint à était appliqué sur le corpus d'évaluation.

Forme ambiguë : c → ce, ces, c'est ,se, ses, s'est, sais, sait

SMS brut	Cc! Euh en fait pr **c** soir **c** compliqué
SMS normalisé	Coucou! Euh en fait pour **ce** soir **c'est** compliqué

FIGURE 5.16 – Exemple de normalisation avec ambiguïté

5.3.2.4 La traduction

Les ressources linguistiques que nous avons induites ont été intégrées dans le système de traduction automatique. Afin de transformer la représentation intermédiaire d'un SMS en sa forme standard, la chaîne de traitement d'Apertium (figure 5.17) réalise, dans un premier temps, une analyse morphologique en utilisant le dictionnaire (*smsfra*) et un étiquetage morphosyntaxique grâce au modèle de probabilité (pour la langue française) fourni par le système. Par la suite, une phase de transfert s'effectue à l'aide du dictionnaire bilingue (*smsfra-fra*) pour

arriver ainsi à la génération morphologique en utilisant le dictionnaire français (*fra*). Finalement, l'ajout des corrections est effectué dans la phase de la post-génération (pas de ressources requises) [7]. Le système lit la phrase déjà analysée (exemple de traduction ci-dessous) et il opère en deux phases en effectuant la traduction mot à mot et le transfert (Caseli *et al.*, 2006). La traduction mot à mot consiste à chercher dans le dictionnaire bilingue la meilleure traduction possible, tandis que la phase de transfert fait appel aux règles de transfert appropriées qui sont implémentées pour le français.

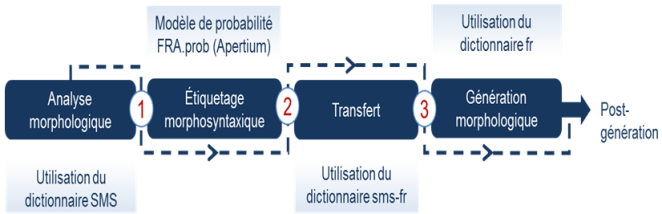

FIGURE 5.17 – Chaîne de traitement Apertium

Nous présentons, ci-après, un exemple de SMS et les étapes de traitement afin d'arriver au résultat du système de traduction :

Impeccc ! Je suis au rdv

Ressources linguistiques

Entrées dans le dictionnaire monolingue smsfra :

<e lm="**impec**" a="webform">**impec**<par n="académique__adj"/></e>
<e lm="**rdv**">**rdv**<par n="mois__n"/></e>

Entrées dans le dictionnaire bilingue smsfra-fra :

7. Pour plus de détails concernant la chaîne de traitement d'Apertium, voir Forcada *et al.* (2008)

<e> <l>**impec**<s n="adj"/></l><r>**impeccable**<s n="adj"/></r> </e>

<e> <l>**rdv**<s n="n"/><s n="m"/><s n="sp"/></l><r>**rendez-vous**<s n="n"/><s n="m"/><s n="sp"/></r> </e>

<u>La normalisation</u>

Représentation intermédiaire :

Impec ! Je suis au rdv

Analyse morphologique (ambiguïtés) :

Impec/Impec<adj><mf><sg>$ ^ !/ !<sent>$

Je/je<prn><tn><p1><mf><sg>

suis/être<vblex><pri><p1><sg>/être<vbser><pri><p1>
<sg>/suivre<vblex><pri><p1><sg>/suivre<vblex><pri><p2>
<sg>/suivre<vblex><imp><p2><sg>

au/à<pr>+le<det><def><m><sg>

rdv/rdv<n><m><sp>

Étiquetage morpho-syntaxique (désambiguïsation) :

Impec<adj><mf><sg>$ ^ !<sent>$

je<prn><tn><p1><mf><sg>

être<vblex><pri><p1><sg>

à<pr>+le<det><def><m><sg>

rdv<n><m><sp>

Transfert :

Impec<adj><mf><sg>/**impeccable**<adj><mf><sg>$ ^ !<sent>/ !<sent>$

je<prn><tn><p1><mf><sg>/**je**<prn><tn><p1><mf><sg>

être<vblex><pri><p1><sg>/**être**<vblex><pri><p1><sg>

à<pr>/à<pr> le<det><def><m><sg>/le<det><def><m><sg>
rdv<n><m><sp>/**rendez-vous**<n><m><sp>

Génération morphologique :
impeccable ! je suis à le rendez-vous

Post-génération :
impeccable ! je suis au rendez-vous

Comme résultat la sortie du système de normalisation produit un résultat plus lisible pour un autre système d'analyse textuelle, mais aussi des résultats partiellement corrects 5.4 ou erronés. La pertinence de ces résultats sera discutée au travers de différentes métriques dans la section 5.4.1 en comparaison avec les SMS transcrits manuellement.

SMS brut	jminkiété jvoulai juste savoir si t t avc lui
SMS normalisé	jminkiété je voulais juste savoir si t t avec lui
SMS brut	sa va cout 3e
SMS normalisé	ça va coûter troisième

Table 5.4 – Exemples erronés de normalisation

5.4 Évaluation

Après la phase d'entraînement du système de traduction automatique sur les deux tiers du corpus de SMS, nous avons procédé à la phase de test. La phase de test consiste à évaluer la performance de notre modèle hybride sur les 7 087 SMS bruts (un tiers du corpus) tirés du corpus alpes4science qui n'ont pas été utilisés pour la construction du dictionnaire bilingue (tableau 5.5).

Voici certaines caractéristiques concernant les SMS du tableau 5.5 :

n	Type	Taille
21 261 SMS	Transcrits manuellement	1 548 Kb
14 174 SMS	Apprentissage	1 044 Kb
7 087 SMS	Corpus de référence	504 Kb
7 087 SMS	Corpus de test (non transcrits)	460 Kb

TABLE 5.5 – Tableau récapitulatif des données SMS

— Le corpus de test de 7 087 SMS que nous utilisons comporte environ 113 070 formes.
— Tous les messages ont été transcrits en minuscules.
— Les émoticônes et les données personnelles ont été maintenues étiquetées.
— Des messages contenant uniquement des chiffres ont été supprimés.
— Les messages contenant des caractères non latins ou écrits dans une langue autre que la langue française ont été exclus.
— Les messages identiques et impossibles à traduire par le traducteur humain ont été aussi écartés.

Dans une deuxième phase, nous avons souhaité, évaluer également les performances du système sur un échantillon de 1 000 SMS du corpus 88milSMS issus du projet *sud4science LR* (Sud4science Languedoc Roussillon).

5.4.1 Les résultats d'évaluation

Les métriques que nous avons appliquées sont le WER, le BLEU et le NIST score [8], trois mesures couramment utilisées pour l'évaluation des transcriptions en CMO (Aw *et al.*, 2006, Han et Baldwin, 2011, Kaufmann et Kalita, 2010a, Beaufort *et al.*, 2010a, Sidarenka *et al.*, 2013). Le choix de ces métriques a été fait afin de pouvoir établir une comparaison avec d'autres recherches réalisées sur la normalisation, eux-même utilisant ces métriques.

8. Les résultats d'évaluation ont été produits à l'aide du logiciel MTEval toolkit : https ://github.com/odashi/mteval

Le Word Error Rate (WER - taux d'erreur de mots en français) est une mesure couramment utilisée pour calculer la performance dans le domaine de la reconnaissance automatique de la parole et de la traduction automatique. Elle est basée sur la mesure de distance de mots de Levenshtein et consiste à calculer le taux d'erreur entre une séquence de mots candidate et une séquence de mots de référence. Selon la définition de Theng (2009), il s'agit de la mesure de performance qui tient compte des erreurs de substitution (reconnaissance erronée d'un mot), des erreurs de suppression (absence de reconnaissance d'un mot) et des insertions (insertion de mots).

Le BLEU score (Papineni *et al.*, 2002) est un outil destiné à évaluer la précision du processus de traduction d'une langue en une autre. Cette évaluation nécessite préalablement un gold standard, autrement dit, un résultat de référence, comme par exemple, une traduction réalisée par un humain. Le gold standard sera par la suite comparé avec la traduction produite automatiquement afin d'établir un score entre *0* et *1*, où *1* signifie l'exactitude absolue entre les deux traductions et *0* que les deux traductions ne présentent aucune similarité (Kaufmann et Kalita, 2010a,b).

Nous utilisons également le NIST score, une métrique alternative dérivée du BLEU score qui diffère sur le degré informatif qu'un n-gramme particulier peut avoir. Si un n-gramme rare a été correctement trouvé, il lui sera attribué plus de poids (Doddington, 2002).

Technique	**BLEU score**	**NIST score**	**WER**
Approche de référence	0.60	11.78	0,26
Représentation intermédiaire (*RI*)	0.62	11.92	0,23
Traduction automatique (*TA*)	0.72	13.45	0,17
Modèle hybride (*RI+TA*)	**0.76**	**14.00**	**0,13**
Gold standard	1	16.38	0

TABLE 5.6 – Résultats d'évaluation

Les résultats de l'évaluation figurant dans le tableau 5.6 nous permettent d'observer un

écart de 0.16 (26.67%) en BLEU score entre l'approche de référence (*baseline*)[9] et la transcription obtenue par l'approche hybride. Par ailleurs, les résultats pour le NIST score nous confirment, tout comme l'écart de 2 points entre le gold standard et l'approche hybride, la bonne qualité de la production, en comparaison avec d'autres systèmes qui fournissent des informations complètes sur ce type de métriques (Kaufmann et Kalita, 2010a,b, Beaufort *et al.*, 2010a,b).

	Français		Anglais		
	Kobus et al.	Beaufort et al.	Aw et al.	Choudh. et al.	Raghun. et al.
Avant normalisation	-	0.47	0.57	0.57	0.54
Après normalisation	0.8	0.83	0.8	0.8	0.86

Table 5.7 – BLEU scores de l'état de l'art

En effet, la table 5.7 fournit les résultats d'approches fondées sur la normalisation des SMS figurant dans l'état d'art réalisées pour le français et l'anglais. Nous remarquons que les deux approches réalisées en français :

— Kobus *et al.* (2008) combinent une méthode de traduction automatique avec reconnaissance de la parole sur un corpus de test de 3 000 SMS.
  Ils privilégient l'utilisation du WER score pour évaluer leurs résultats de normalisation en rapportant une amélioration du score de 12.26 à 10.82. En ce qui concerne le BLEU score, nous sommes limités au score approximatif de 0.8 sans même avoir accès au score initial.

— Beaufort *et al.* (2010a,b) basés sur le concept de modèles entraînés sur corpus, ils utilisent une méthode hybride combinant la correction orthographique et la traduction automatique. Leur système obtient les résultats de la normalisation grâce à une méthode de validation croisée de k-blocs[10].

9. Évaluation effectuée entre les SMS bruts (sans aucune normalisation) et leurs transcriptions.
10. Il s'agit d'une estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage.

En divisant les 30 000 SMS du corpus en 10 blocs (3 000 SMS) le système est entraîné et testé 10 fois, en excluant chaque bloc à son tour du corpus d'entraînement, mais le seul à servir de corpus de test. Comme résultat à chaque jeux de test le corpus d'entraînement représente 27 000 SMS sur 3 000 milles SMS de test.
En effet, les scores obtenus sont assez encourageants avec un écart de 0.36 pour le BLEU score entre la phase avant et après normalisation. Cependant, nous ne considérons pas que ce résultat est comparable puisque la taille du corpus et la méthode appliquée ne reflète pas la nôtre.

Citons également trois évaluations réalisées pour la normalisation de SMS en anglais (Aw *et al.*, 2006, Choudhury *et al.*, 2007, Raghunathan et Krawczyk, 2009). Raghunathan et Krawczyk (2009) ont exploré deux approches de normalisation : une par substitution de mots à partir d'un dictionnaire et une autre basée sur la traduction automatique statistique. Les deux approches ont été évaluées sur trois corpus SMS en anglais. Un corpus a été comparé avec l'approche de Aw *et al.* (2006) (0,80 BLEU score) afin de montrer qu'ils obtiennent de meilleurs scores (0,86 BLEU score). Quant à Choudhury *et al.* (2007), ils sont à égalité de score avec Aw *et al.* (2006) en utilisant une approche qui opère une normalisation mot à mot couplée avec un modèle statistique de Markov caché.

En règle général, nous observons que les résultats que nous obtenons sont assez encourageants et proches des autres méthodes, bien que la comparaison semble arbitraire tant au niveau de la langue (français/anglais), qu'au niveau de la taille du corpus d'apprentissage/test.

Par ailleurs, Kaufmann et Kalita (2010a) démontre que le BLEU score semble ne pas être le meilleur choix pour évaluer les productions de normalisation de SMS. Son argument se fonde sur le constat que cette métrique n'est pas destinée à évaluer les produits d'une normalisation d'un texte bruité mais aussi sur le fait que la subjectivité des annotateurs humains peut entraîner des variations considérables. Ainsi, un meilleur BLEU score ne signifie pas une meilleure traduction. L'exemple du mot *miam* tiré du corpus alpes4science et annoté de *manger*, *repas*,

appétit est cité ci-dessous :

1. **Annotateur A**

   SMS Original : 1 bouteille de chartreuse, 1 de genepi, prend plutot du blanc ou/et du **miam**.

   SMS Transcrit : Une bouteille de chartreuse, une de genepi, prends plutôt du blanc ou/et à **manger**.

2. **Annotateur A**

   SMS Original : A partir de 11h. Tu avais appelé sogestra ? Bon **miam** Bisous

   SMS Transcrit : À partir de 11h. [/Est-ce que/] tu avais appelé sogestra ? Bon **repas** Bisous

3. **Annotateur A**

   SMS Original : Pas de cable usb chez sfr ni surcouf. Bisous bon **miam**

   SMS Transcrit : Pas de cable usb chez sfr ni surcouf. Bisous bon **appétit**

4. **Annotateur B**

   SMS Original : Midi, l'heure du **miam** !

   SMS Transcrit : Midi, l'heure de **manger** !

Nous remarquons que le mot miam est une interjection couramment utilisée dans les productions orales, de telle façon que le transcripteur pourrait l'ignorer lors de la transcription. En outre, nous soulignons le fait que l'annotateur A pour les exemples 2 et 3 a utilisé pour *bon miam* aléatoirement le mot *appétit* et *repas*. Étant donnée que le BLEU score est une métrique qui se base sur le calcul des similarités entre les *n-grammes*, ce type d'erreurs peut affecter considérablement le score final.

5.4.1.1 Types d'erreurs détectés

Dans l'optique d'une amélioration possible du système dans l'avenir nous fournissons une analyse fondée sur la sortie produite par le système hybride en comparaison avec le corpus de référence. Étant donné le grand volume de SMS, nous avons sélectionné de façon aléatoire 1 500 SMS afin de procéder par la suite à une classification manuelle de fautes/erreurs et de problèmes détectés. Ainsi, le tableau 5.8 nous indique les six catégories qui ont été établies afin d'apporter des améliorations possibles au résultat obtenu.

L'analyse effectuée nous montre donc que 49,7 % d'erreurs proviennent du fait que les mots, majoritairement des noms, ne figurent pas dans le dictionnaire bilingue *smsfra-fra*. L'amélioration proposée repose sur l'utilisation des dictionnaires plus élargis car la qualité des résultats du système de traduction dépend des ressources linguistiques fournies pour alimenter le système. L'alimentation du système avec des dictionnaires provenant d'autres projets de récolte de messages courts pourrait améliorer significativement les résultats.

Nous remarquons également que 20,1 % d'erreurs sont dues au manque d'accent, 10,1 % au manque d'accord avec le participe passé et 6,7 % à l'absence d'accord entre le sujet et le verbe. En effet, le système ne vise pas la correction orthographique. Il a comme objectif la normalisation lexicale du SMS dans le but de rendre le message court compréhensible par la machine en vue d'une analyse plus étendue dans la suite du traitement. Cependant, diverses solutions pourraient être proposées afin d'apporter une correction orthographique et conduire à une amélioration du résultat au travers de la construction de règles de transfert supplémentaires ou de l'application d'un correcteur orthographique sur les messages problématiques localisés.

Dans le tableau 5.8, la catégorie *analyse plus élargie* englobe les cas où l'identification du problème de normalisation est assez complexe. D'une part, car le type d'erreur est ambigu, nécessitant des informations liées au contexte de la situation langagière (utilisation de l'infinitif à la place du participe passé) et, d'autre part, parce qu'un nombre de SMS traités s'avèrent

incomplets, du fait du découpage de messages contentant plus de 160 caractères.

Catégorie d'erreur	**%**
Mots hors dictionnaire	49,7
Manque d'accent	20,1
Analyse plus élargie	13,4
Accord participe passé	10,1
Accord sujet verbe	6,7

TABLE 5.8 – Catégories d'erreurs

5.4.2 Évaluation du système sur le corpus 88milSMS

Pour notre expérimentation, dans le but d'évaluer notre système sur un corpus différent de celui d'alpes4science, nous avons fait appel à l'échantillon de 1 000 SMS bruts anonymisés et transcodés en français standardisé du corpus de 88milSMS [11] (Panckhurst *et al.*, 2014a,b, Patel *et al.*, 2013). Le corpus 88milSMS est le fruit du projet sud4science, lui-aussi faisant partie du projet international sms4science autour de la collecte de données SMS. Une initiative pendant laquelle environ 88 000 SMS authentiques en français ont été recueillis provenant de 424 personnes (Panckhurst *et al.*, 2013, Panckhurst et Moïse, 2012). La collecte de SMS a eu lieu dans la région Languedoc-Roussillon en 2011 sur une période de trois mois. Selon Panckhurst *et al.* (2013) les messages collectés contiennent en moyenne 55 caractères, sans espaces, 67 caractères, avec espaces et chaque message est d'une longueur moyenne de 13,75 mots.

Afin de pouvoir effectuer l'évaluation nous avons procédé à une harmonisation du texte qui consiste à transformer toutes les formes du corpus SMS brut en minuscules aussi bien qu'à éliminer tous les marqueurs spécifiques provenant du modèle de normalisation, dans le but d'avoir un résultat d'évaluation représentatif.

Nous avons utilisé les mêmes métriques que nous avons employées dans la partie 5.4.1,

11. Disponible après téléchargement : `http://88milsms.huma-num.fr/corpus.html` et sur le site d'Ortolang : `https://hdl.handle.net/11403/comere/cmr-88milsms`

Technique	**BLEU score**	**NIST score**	**WER score**
Approche de référence	0.50	8.96	0.37
Modèle hybride	**0.75**	**11.49**	**0.15**
Gold standard	1	13.83	0

TABLE 5.9 – Résultats d'évaluation - corpus 88milSMS

concernant l'évaluation menée sur l'échantillon de 7 087 SMS bruts du corpus alpes4science.

Les résultats sont assez encourageants, notamment en comparaison avec les résultats de l'évaluation du corpus alpes4science. Il faut souligner le fait que le système n'a pas été entraîné sur ce corpus. En effet, les scores pour le corpus de 88milSMS, en comparaison avec les résultats obtenus pour le corpus alpes4science (tableau 5.6), se montrent meilleurs. Nous remarquons que pour le BLEU score du corpus 88milSMS, il y a un écart de 0,25 points entre l'approche de référence et le modèle hybride quand pour le corpus d'alpes4science l'écart pour ces mêmes techniques est de 0,16 points. Idem, pour le score WER, l'écart est de 0,22 points et pour le corpus alpes4science il est de 0,13 points, ce qui signifie qu'il y a un plus grand taux de correction pour le corpus 88milSMS que pour le corpus alpes4science. Cependant, il ne faut pas négliger le fait que les deux corpus ne sont pas de la même taille. Lopez *et al.* (2014) ont testé, sur un échantillon de 100 SMS du même corpus, un modèle d'alignement statistique pour la normalisation de SMS. Cependant, les faibles scores de précision (0,59) et rappel (0,55) montrent la nécessité d'améliorer la méthode avec plusieurs heuristiques d'alignement. Tarrade (2017) a aussi réalisé un modèle pour la normalisation de SMS sur le corpus 88milSMS. Les résultats obtenus atteignent le 0,61 pour le BLEU score et le 0,25 pour le WER.

5.5 Conclusion

Dans ce chapitre nous avons décrit l'architecture d'un modèle hybride destiné à la normalisation de messages contenant du bruit, tels les SMS. L'approche est fondée sur l'analyse typologique et néographique à l'aide de grammaires locales, couplée avec un système de tra-

N° d'erreurs	SMS original	SMS normalisé	Nbr d'erreurs
6	Tkt tu ne me deranges pas ! Je sais pas du tout, je suis en robe mais bon j'met jamais de pantalon :/ jean tshirt assez joli ca le fait je pense ! C :est pas un pyjama hahaha	T'inquiètes pas tu ne me deranges pas ! je sais pas du tout, je suis en robe mais bon je met jamais de pantalon ***emoticon*** jean t-shirt assez joli ça le fait je pense ! C :est pas un pyjama hahaha	3
6	Cc ma chérie ça va ? Dis moi kes ki c passe avc le *taf* ? Un pblm ? Gros bisous <3	coucou ma chérie ça va ? dis moi qu'est ce qui c'est passe avec le *travail* ? un problème ? gros bisous ***emoticon***	1

Table 5.10 – Exemples de traduction du corpus 88milSMS

duction automatique.

Les résultats d'évaluation en terme de BLEU score et WER nous montrent que l'approche proposée améliore la qualité, la lisibilité et l'opérationnalité des SMS. Cependant, nous voulons pointer deux problèmes de normalisation : la désambiguïsation de formes polysémiques et la catégorie d'erreurs qui correspond aux mots absents du dictionnaire.

L'amélioration des résultats pourrait, comme nous l'avons constaté, dépendre des ressources linguistiques fournies (donc le corpus de départ). Des données avec des informations lexicales plus enrichies (par exemple des informations morphosyntaxiques) peuvent être facilement incorporées au système de traduction et générer des règles qui s'appliqueront pour augmenter les performances lors de la phase de normalisation.

Les résultats sont encourageants et suffisants pour procéder dans une étape suivante à l'application d'outil liés au traitement du langage comme les étiqueteurs morphosyntaxiques, la reconnaissance d'entités, la traduction automatiques, la lecture vocale des messages etc. et nous incitent à explorer de nouveaux modes d'hybridation. La conception de systèmes

additionnels pour le traitement des emprunts, la phonétisation, le calcul de distances d'édition et la mémoire de traduction pourraient aussi être ajoutés au système initial et faciliter la tâche de normalisation. Nous estimons que cette approche pourrait être adaptée pour s'appliquer à d'autres types de messages bruités courts (tweets, chats, forums etc.), mais à condition d'avoir des ressources linguistiques adéquates.

L'étape suivante consiste à exploiter les normalisations réalisées afin d'arriver à extraire des informations contenus dans les SMS. Plus précisément, le chapitre suivant est consacré à la présentation des résultats issus d'un analyseur morphosyntaxique et l'illustration d'un modèle dédié à l'extraction d'entités nommées de type nom, prénom, lieu que nous allons appliquer sur les messages déjà normalisés.

Chapitre 6

Autour de la reconnaissance d'entités nommées dans les SMS

Sommaire

6.1 Introduction 133

6.2 Les entités nommées 134

6.2.1 Une définition de l'entité nommée 138

6.2.2 Les catégories 140

6.2.3 Les difficultés de la catégorisation 143

6.2.4 Pourquoi extraire des entités nommées ? 147

6.3 La reconnaissance d'entités nommées dans les SMS 155

6.3.1 Une typologie d'entités nommées pour les SMS 161

6.4 Conclusion 166

6.1 Introduction

Ce chapitre s'intéresse à l'entité nommée du point de vue théorique, en visant sa définition et son parcours historique. A travers la définition nous découvrirons les catégorisations

existantes pour acquérir une typologie autour de l'annotation d'entités nommées au moyen de la tâche de reconnaissance d'entités nommées. En effet, la reconnaissance d'entités nommées est importante pour de nombreuses applications issues du TAL, notamment pour l'extraction d'informations. Plus précisément, ce chapitre se focalise sur les spécificités des entités nommées du langage SMS et, par extension, aux messages courts et bruités. D'une part, on se concentre sur les travaux réalisés et, d'autre part, sur les applications possibles et l'introduction d'une typologie pour l'extraction d'entités nommées de SMS.

Ce chapitre 6 présente ainsi la définition de l'entité nommée (partie 6.2.1) et expose ses catégorisations (partie 6.2.2). Les différents problèmes de catégorisation dans l'état de l'art de la reconnaissance d'entités (section 6.3) pour les textes courts sont présentés dans la partie 6.2.3. Les applications possibles de la tâche de reconnaissance d'entités nommées dans les SMS sont décrites dans la partie 6.2.4 et la première typologie pour les entités nommées issues des SMS et des messages électroniques en général se trouve dans la partie 6.3.1.

6.2 Les entités nommées

La reconnaissance d'entités nommées (REN) comme processus issu du traitement automatique du langage trouve son origine dans la tâche d'extraction d'information. En effet, l'extraction d'information vise à détecter des éléments pertinents d'information, comme par exemple, extraire des événements spécifiques. La REN est liée au domaine de la recherche d'informations et vise l'identification de données informationnelles d'un texte. Elle se concrétise au travers d'une suite de sept conférences *MUC* (1987–1998) (Ehrmann, 2008). Une des premières réalisations dans l'extraction d'information à travers la reconnaissance d'EN est celle de Rau (1991) grâce à un algorithme pour extraire automatiquement les noms des entreprises issus des documents traitant des nouvelles financières. Depuis plusieurs recherches ont vu le jour. Parmi les recherches centrées sur la REN nous trouvons Nadeau et Sekine (2007) sur

les caractéristiques de la REN et la classification d'entités nommées. (Chinchor *et al.*, 1998) travaillent sur le plan d'évaluation HUB-4 autour la reconnaissance des nouvelles de radio-diffusion. Le projet IREX (Satoshi et Hitoshi, 2000) porte sur l'extraction d'informations en japonais. La suite de conférences CONLL sur l'apprentissage de la langue naturelle notamment avec les recherches de Tjong Kim Sang (2002) et Tjong Kim Sang et De Meulder (2003a) sur la reconnaissance d'entités nommées indépendamment de la langue du texte. Le programme ACE (Doddington *et al.*, 2004) avait comme objectif principal de développer la technologie pour déduire automatiquement les entités, les relations entre ces entités directement exprimées et les événements auxquels ces entités participent. Une autre campagne d'évaluation de la reconnaissance d'entités nommées pour le portugais est HAREM (Freitas *et al.*, 2010, Santos *et al.*, 2006).

Une entité nommée (EN) représente un élément du langage qui fait référence à une entité unique du domaine du discours avec une valeur informationnelle prépondérante. Parmi les entités nommées nous trouvons les unités lexicales qui décrivent les noms de personnes, les organisations, les lieux, les dates, les quantités, les distances, les valeurs, etc. Tjong Kim Sang et De Meulder (2003b), Chinchor (1998), Meur *et al.* (2004).

Les EN font l'objet de recherches portant sur la linguistique théorique, la linguistique informatique, la linguistique de corpus, la terminologie, la lexicologie et la traduction. La reconnaissance d'entités nommés est une tâche importante pour l'extraction et le traitement de l'information, en effet, les entités reconnues donnent des indications sur le contenu d'un texte (Crucianu *et al.*, 1999) et s'avèrent ainsi fondamentales, par exemple, pour la conception et fonctionnement des moteurs de recherche, des systèmes de question-réponse, pour la fouille de données textuelles et la compréhension des textes. Par ailleurs, cette reconnaissance peut s'adapter au type de texte et s'appliquer à une vaste gamme de domaines.

Selon (Enjalbert, 2005), étant donné un texte en entrée, le processus de reconnaissance d'entités nommées se divise en deux phases : la première de *repérage*, la deuxième d'*étiquetage*.

Tout d'abord, il faut *repérer* toutes les formes linguistiques qui décrivent de manière univoque une entité nommée par leur pouvoir de sélectivité, puis leur attribuer une étiquette choisie dans une liste prédéfinie.

Par contre, pour (Ehrmann, 2008), étant donnée un texte en entrée, le processus de reconnaissance d'entités nommées se divise en trois étapes : 1) *identifier* les unités dans le texte, 2) *catégoriser*, selon le cas, les unités identifiées, et 3) *annoter* ces unités, souvent à l'aide de balises. Considérons l'exemple ci-après :

Exemple d'annotation d'entités nommées
Le Londres du XVIe siècle était déjà une ville démesurée. Cheapside était la grande rue. Saint-Paul, qui est un dôme, était une flèche. La peste était à Londres presque à demeure et chez elle, comme à Constantinople. Il est vrai qu'il n'y avait pas loin de Henri VIII à un sultan.
Le <location>**Londres**</location> du <date>**XVIe siècle**</date> était déjà une ville démesurée. <location>**Cheapside**</location> était la grande rue. <location>**Saint-Paul**</location>, qui est un dôme, était une flèche. La peste était à <location>**Londres**</location> presque à demeure et chez elle, comme à <location>**Constantinople**</location>. Il est vrai qu'il n'y avait pas loin de <person>**Henri VIII**</person> à un sultan.

TABLE 6.1 – Annotation d'entités nommées

Dans l'exemple [1], présenté au tableau 6.1, ont été identifiées les unités lexicales : *Londres*, *XVIe siècle*, *Cheapside*, etc. Par la suite, les unités lexicales identifiées, ont été catégorisées parmi une des catégories : *location*, *date*, ou *person*. Finalement, les unités lexicales catégorisées ont été annotées à l'aide d'une balise correspondant à la catégorie choisie : <location>...</location>, <date>...</date>, <person>...</person>. McDonald (1996) déclare que pour acquérir l'identification d'EN il est nécessaire de définir *la modélisation de classes*

1. Extrait tiré de l'œuvre de Hugo, *À propos de William Shakespeare vers 1564-1616*, disponible sur : `http://abu.cnam.fr/cgi-bin/donner_unformated?hugoshak1`

orthographiques et lexicales d'éléments qui sont analogues à ce que l'on trouve dans les autres types de phrases riches en structures syntaxiques, riches en contenu, dans la «périphérie» du langage comme par exemple les dates, les numéros, les citations, etc. Soutenant l'idée que cette grammaire doit *être sensible en contexte* et qu'elle doit incorporer *un modèle sémantique riche de noms et de leurs relations avec les individus.* Inspiré par cette perception d'EN, il formula sa fameuse dichotomie qui est fondée sur l'identification d'éléments indicateurs d'EN : la *preuve interne* et la *preuve externe.*

Preuve interne :
La preuve interne est dérivée de la séquence de mots qui compose l'entité elle-même, autrement dit, à l'intérieur de l'entité. Par exemple pour l'entité *Union des Républiques Socialistes Soviétiques* le mot *Union* fait partie à part entière de l'entité nommée. Les preuves jouent le rôle de marqueurs, des mots ou abréviations permettant de typer l'entité. Le marqueur typographique le plus représentatif, dans la majorité de cas, est la majuscule. Cependant, dans le cas du langage de SMS les règles ne sont pas respectées, par exemple, quand il s'agit de la lettre initiale après la fin de phrase.

Nous trouvons également grâce à Ehrmann (2008) et Hatmi (2014) d'autres indices parmi lesquels : les prénoms, les titres générationnels pour désigner les noms de personnes, les affixes de type classifiant pour les noms de personnes (par exemple le préfixe *Mac* dérive de la forme ancestrale du gaélique Macc qui signifie « fils », sert de préfixe à de nombreux noms de famille d'origine irlandaise et écossaise ou le suffixe -*(r)ena* en basque désigne "la maison de" se référant à une famille ou à sa résidence(Hanks, 2003)), les organisations et les lieux, les sigles ou les signes graphiques comme la ponctuation et les caractères spéciaux (le *et commercial*) et numériques (pour les noms d'organisations). Voici certains indices internes :

Ltd. International, INPG Entreprise *SA*, *Gmbh & Co. KG*
Roisy CDG

Brigitte Dubois, *Victor* Cousin, *Marion* Dumas

Gare de Lyon, *Châteaux* de Versailles, *Bois* de Boulogne

Place de la Concorde

George Bush *Jr.*

*Mac*Kenzie, Amigo*rena*

Preuve externe :

La preuve externe est l'ensemble de critères fournis par le contexte dans lequel l'entité apparaît. Il s'agit d'un contexte périphérique droit et gauche de l'entité dans une phrase donnée. Le contexte constitue ainsi le "satellite" de l'entité puisqu'il l'entoure en permettant, par exemple, l'observation du contexte immédiat d'une entité afin de confirmer facilement et constituer l'indice de critères pour la catégorisation. Pour l'exemple *René Décurey, le directeur général d'Air Côte d'Ivoire* le marqueur *directeur général* présente simultanément une preuve externe et constitue le contexte droit pour l'entité *René Décurey* en même temps qu'il sert de contexte gauche pour l'entité *Air Côte d'Ivoire*. Ci-après nous trouvons quelques exemples permettant l'observation de ces éléments :

le *groupe* Fincantieri

la *maison de retraite* Les Opalines

L'*entreprise* Gervais Dubé Inc.

René Décurey, le *directeur général* d'Air Côte d'Ivoire

Dr. Dubois

6.2.1 Une définition de l'entité nommée

Les conférences MUC[2](*Message Understanding Conferences*) ont été créées pour encourager le développement de nouvelles et meilleures méthodes d'extraction de l'information. À l'occasion de la MUC-7 en 1998, Chinchor (1998) donne la définition suivante d'entité nommée :

2. `http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/muc.htm`

Les entités nommées se définissent comme des noms propres[3] *et des quantités d'intérêt. Les noms de personne, d'organisation et de localisation sont marqués aussi bien que les dates, les heures, les pourcentages et les montants monétaires.*

Cette définition est discutée par Friburger (2002), qui remarque le fait que le terme *entité nommée* est employé par les informaticiens travaillant dans le domaine de l'extraction d'information pour regrouper tous les éléments du langage définis par référence : *les noms propres au sens classique, les noms propres dans un sens élargi mais aussi les expressions de temps et de quantités.* La raison de cette simplification est selon l'auteur due à la difficulté à différencier les *noms propres des autres noms*, par exemple, à établir une délimitation entre *l'ensemble des noms propres et l'ensemble des noms communs.*

Pour compléter la définition de Chinchor (1998) en prenant en compte la remarque de Friburger (2002), nous citons Meur *et al.* (2004), qui mentionne que bien qu'il n'existe pas de définition standard pour l'entité nommée, il est possible de la considérer comme une *unité lexicale particulière.* Ainsi, une entité nommée fait référence à une *entité du monde concret* qui correspond à des *domaines spécifiques* (humains, sociaux, politiques, économiques, géographiques, etc.) et qui a un *nom* (nom propre ou acronyme). Cette définition peut également se rapprocher de celle des éléments *discursifs monoréférentiels*, qui selon Vicente (2005), présentent des qualités similaires à celles des noms propres et suivent des *patrons syntaxiques déterminés.*

En ce qui concerne les entités du monde concret, qui sont référencées par les entités nom-

3. Le terme *nom propre* dérive du latin *nomen proprium* et du grec *kúrion onoma*, étymologiquement indique le nom à proprement parler, le *vrai nom*, celui qui est *le nom par excellence* (Gary-Prieur, 1991). Selon Wilmet (1995), les noms propres ont toujours leurs lettres initiales en majuscules, ils se réfèrent aux prénoms, aux noms de famille, aux noms de dynasties, aux noms de peuples, aux noms géographiques désignant des pays, des villes, etc. Pour sa part, Goosse (1986) spécifie qu'un nom propre n'a pas de signification véritable ou de définition, il se rattache à ce qu'il désigne par un lien qui n'est pas sémantique, mais par une convention qui lui est particulière.

mées, la conférence MUC-7 établit, par exemple, trois supra-catégories à prendre en compte : *les noms d'entités, les expressions temporelles et les expressions numériques*. Ces trois supra-catégories sont subdivisées en 1) organisations, personnes, emplacements ; 2) dates et heures et 3) valeurs monétaires et pourcentages.

Nous utiliserons désormais la notion d'entité nommée comme étant des unités lexicales qui suivent des patrons syntaxiques déterminés et qui font référence à des entités du monde concret, préalablement définies, dans des domaines spécifiques.

6.2.2 Les catégories

Les catégories que vous pouvons attribuer aux entités nommées correspondent à l'annotation de différentes étiquettes sémantiques que vous pouvons identifier dans un texte. A l'occasion des conférences MUC un certain nombre de classifications a vu le jour. L'identification des entités nommées se divise en trois classes d'expressions (Fourour, 2001, Poibeau, 2005, Daille *et al.*, 2000) :

ENAMEX	Les noms propres qui se référent aux noms de personne, de lieu ou d'organisation.
TIMEX	Les expressions temporelles divisées en dates et heures
NUMEX	Les expressions numériques qui incluent les expressions monétaires et les pourcentages

TABLE 6.2 – Trois catégories d'entités nommées selon MUC (1995)

La première classe *ENAMEX* regroupe tout les noms propres, quant aux deux autres, TIMEX et NUMEX englobent les entités numériques, les expressions de temps et numériques.

Poibeau (2005) nous retrouvons le graphe hiérarchique des types définies pour le système QALC(Ferret *et al.*, 2001) basé sur cette classification (figure 6.1).

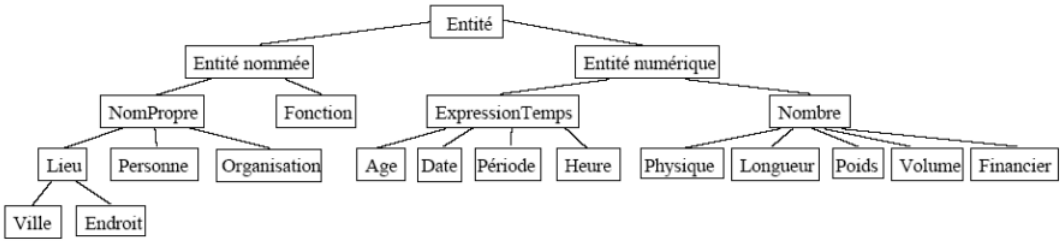

FIGURE 6.1 – Graphe hiérarchique de types définies pour le système QALC dans Poibeau (2005)

Une année plus tard Paik *et al.* (1996), dans l'implémentation du module *DR-LINK* pour la catégorisation automatique, d'entités nommées introduiront leur typologie répartie en neuf classes et trente catégories : Entités Géographiques (villes, ports, aéroports, îles, départements, provinces, etc.), Affiliations (religions, nationalités), Organisation (entreprises, types d'entreprises, institutions, organisations), Humain (personnes, fonctions), Document (documents), Équipement (logiciels, matériels, machines), Scientifique (maladies, drogues, médicaments), Temporelle (dates et heures) et Divers. En effet, comme Ehrmann (2008) et Daille *et al.* (2000) le remarquent, cette classification nous rappelle la typologie la plus connue et seule existante pour la traduction (Fourour, 2001) sur les noms propres que le linguiste germanophone Bauer (1985) a introduit. Citée par Daille *et al.* (2000) et Tran et Maurel (2006) cette typologie regroupe six classes principales, avec pour chacune, plusieurs catégories :

— *Anthroponymes* : noms de personnes individuelles et groupes (patronymes, prénoms, pseudonymes, gentilés, hypocoristes, etc.).
— *Toponymes* : les noms de lieux (pays, villes, microtoponymes, etc.).
— *Ergonymes* : les noms d'objets et de produits manufacturés (marques, entreprises, établissements d'enseignement et de recherche, etc.).

— *Praxonymes* : les noms de faits historiques, de maladies et d'événements culturels.
— *Phénonymes* : les noms de phénomènes météorologiques (ouragans, les zones de haute et de basse pressions, les astres et les comètes).
— *Zoonymes* : les noms d'animaux familiers.

Parmi les propositions nous pouvons citer la hiérarchie de l'entité nommée étendue de Sekine *et al.* (2002), antérieure aux conférences MUC, du projet IREX [4] (Sekine et Isahara, 1999) et du programme ACE [5](Doddington *et al.*, 2004), qui se limitent à huit catégories maximum. Ils ont conçu une hiérarchie qui peut couvrir un maximum d'EN qui figurent dans les journaux et les articles. En effet, il considèrent qu'une classification de 7 ou 8 types d'EN ne couvre pas les problèmes généraux car différents types d'EN sont nécessaires selon les besoins du domaine d'application. Ainsi, dans une tentative de couvrir le plus d'entités possible, ils proposent une classification de 150 types d'EN, tout en mentionnant le fait qu'ils ne visent pas la couverture de domaines spécifiques. Aussi, bien que ce nombre de type puisse paraître exhaustive, comme Ehrmann (2008) le déclare : *Il n'existe bien sûr aucune catégorisation idéale ni de solution pour y parvenir ; le mieux semble être de suivre la proposition de Sekine* et al.*, «We believe that there is no ultimate solution, so we seek rather empirical solution* [6]*», et de multiplier les sources d'inspiration.* La complexité de la vision de l'EN est illustrée à la figure

4. Projet IREX (Information Retrieval and Extraction Exercise). Il s'agit d'un projet de compétition pour la recherche et l'extraction de l'information. Le projet a démarré en mai 1998 et s'est achevé en septembre 1999 avec un atelier IREX tenu à Tokyo. Pour ce projet sept classes d'EN ont été définies, dont les classes de MUC plus la classe ARTIFACT pour annoter les noms de produits ou de prix (par exemple le Prix Nobel) et la classe OPTIONAL pour les cas où même l'humain a des difficultés à déterminer une classe sans ambiguïté : Organisation, Personne, Lieux, Artifact, Date, Heure, Monnaie et Pourcentage.

5. Programme ACE (Automatic Content Extraction). Le programme a démarré en 1999 avec pour objectif principal de développer la technologie pour déduire automatiquement les entités, les relations entre ces entités directement exprimées et les événements auxquels ces entités participent. Les entités se limitent à cinq types : Personne, Organisation, Installations (bâtiments), GSP (Entités Socio-Geo-Politiques) et Lieux.(Doddington *et al.*, 2000)

6. *Nous croyons qu'il n'existe pas de solution ultime, ainsi nous recherchons plutôt une solution empirique*

de Daille *et al.* (2000) (figure 6.2).

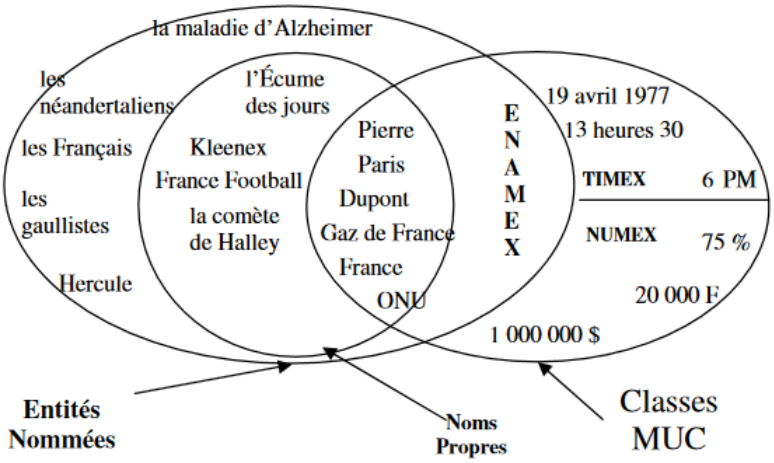

FIGURE 6.2 – La notion d'entité nommée dans Daille *et al.* (2000)

6.2.3 Les difficultés de la catégorisation

La tâche de catégorisation d'EN n'est ni toujours facile ni évidente, puisque nous repérons deux groupes de phénomènes nécessitant davantage d'analyse avant de procéder à la catégorisation : (a) la polysémie et (b) les frontières d'EN.

6.2.3.1 La polysémie

Comme tous les autres types d'unités linguistiques les ENs sont également concernées par les phénomènes de la polysémie et métonymie lexicale, cela signifie qu'elles peuvent correspondre à différentes classes sémantiques et, par conséquent, d'être ambiguës (Campedel et Hoogstoël, 2011). La polysémie se traduit par le fait qu'une unité lexicale est associée à plusieurs significations, autrement dit selon Gaudin (2000), à des unités de sens différents qui *portent le même nom* et que les locuteurs *ne trouvent pas ce fait fortuit*. Par exemple l'unité

face désigne (a) chacun des côtés d'une chose, la partie extérieure de quelque chose, (b) la partie antérieure de la tête humaine, le visage, mais aussi (c) le côté d'une monnaie portant l'effigie du souverain ou l'image personnifiant l'autorité au nom de laquelle la pièce est émise. Poibeau (2005) distingue cinq types d'ENs sujettes à se référer souvent à plusieurs classes sémantiques :

— Les **héméronymes** qui de manière métonymique[7] déclarent des événements produits à la même date. Dans l'exemple qui suit la date 14 juillet est une EN qui décrit à la fois une date et un événement historique en comparaison avec le 13 et 15 juillet qui ne peuvent être a priori que de dates.
Un feu d'artifice a été organisé à la Tour Eiffel vendredi, pour célébrer le 14 Juillet.

— Les **noms d'organisations** peuvent designer une institution, une communauté d'individus ou encore un bâtiment.
La NASA a lancé deux sondes spatiales.
La NASA a donné son feu vert à la construction de Dart.
Quel budget prévoir pour un voyage à laNASA ?
Combien gagne un employé de la NASA ?

— Les **toponymes** renvoient aux lieux, aux habitants et à l'institution qui les gouverne (pays ou commune) mais aussi aux noms de personnes.
Washington était un entrepreneur et il fut le premier président des États-Unis...
La conférence de Paris appelle Washington à maintenir son financement.
Selon Khartoum, Washington est "responsable" de l'impact sécuritaire des sanctions.
Venezuela : l'échec de Washington à Cancun.

7. La métonymie fait aussi part de la polysémie. Elle consiste à la transition d'une désignation d'une réalité par un nom qui se réfère à une autre réalité. Dans Ehrmann (2008) la métonymie est le fait d'*employer un mot attaché à une certaine entité pour en désigner une autre, la seconde étant liée à la première par un rapport fonctionnel ou structurel.*

Paris veut 2024 et rien d'autre.

— Les **anthroponymes** partagent les mêmes propriétés polysémiques que les toponymes tout en étant instables et nécessitant assez souvent un contexte afin de résoudre l'ambiguïté. Dans les exemples qui suivent l'EN *Charles-de-Gaulle* correspond au phénomène d'*homonymie*[8], un cas linguistique qui entraîne de l'ambiguïté qui est a priori connue, mais, comme l'expression linguistique est identique pour plusieurs référents, seul le contexte pourra désambiguïser (Nouvel, 2012).
Aux Rencontres d'Arles, Dior[9] expose 50 ans d'images de beauté.
Le Charles-de-Gaulle[10] sort tout juste d'un entretien intermédiaire de six mois.
Il a projeté de faire de somptueux travaux dans l'hôtel Lambert, dans l'ile Saint-Louis.
Un essai de chargement a été réalisé sur le pont Charles-de-Gaulle.
Quelques mouvements de personnel sont attendus au collège Charles-de-Gaulle.
Un Picasso est devenu lundi la toile la plus chère jamais vendue aux enchères.

— L'**antonomase**[11] est une figure de style métonymique lorsqu'un personnage historique substitue une circonlocution (une phrase qui déclare une qualité, action ou propriété transformant un nom propre en nom commun).
Mark Sloan est un Don Juan.

— Les **sigles** sont constitués d'une seule unité lexicale constituée de plus d'une majuscule et dont chaque lettre en majuscule réfère elle-même à une autre unité lexicale comme

8. On parle d'homonymie lorsque la relation entre plusieurs formes linguistiques a le même signifiant, graphique ou phonique, et des signifiés entièrement différents (Breckx, 1996).
9. La maison de haute-couture Christian Dior.
10. Il s'agit du porte-avions Charles-de-Gaulle
11. Terme issu de la forme anthonomasie, un emprunt au latin antonomasia, levant lui-même du vocabulaire rhétorique, et construit sur des racines grecques anciennes *ἀντονομασία*, antonomasía de même sens, composé de *ἀντί*, antí (« à la place de ») et *ὄνομα*, ónoma (« nom ») (Leroy, 2001).

par exemple : Organisation du traité de l'Atlantique Nord → OTAN. Les exemples qui suivent avec l'unité *PVI* illustrent le fait qu'un seul sigle peut correspondre à plusieurs entités nommées.

PVI [12] conçoit pour les villes des technologies de transport 100 % électriques.

Nous avons reçu en retour votre contrat signé de la présidence ainsi que le PVI [13].

PVI [14] est la revue trimestrielle de la Grande loge de France.

L'an 1 du nouveau Pentair, devenu PVI [15].

PVI [16], le taux d'alcoolisme parmi policiers est le double de la moyenne nationale.

Un autre constat de polysémie est mentionné dans Campedel et Hoogstoël (2011), selon lesquels une entité nommée peut appartenir à plusieurs fonctions, ce que Cruse (2004) appelle *facettes*. Par exemple, *Emmanuel Macron* est un homme d'État français, ancien haut fonctionnaire et banquier d'affaires et président de la République française.

6.2.3.2 Les frontières

Dans cette partie, nous voulons aborder le problème de délimitation d'une entité nommée, la prise en compte ou non des éléments entrant dans la constitution d'un syntagme dont la tête est une EN, autrement dit, pouvoir définir ou commence et ou finit une EN. A propos de ce sujet, Ehrmann (2008) se positionne sur l'étendue de l'entité nommée et le problème de l'annotation des entités. En effet, il n'existe pas une règle standard à suivre puisque chaque programme et chercheur applique ses propres directives. Dans les conférences MUC nous trouvons que seuls les *titres générationnels*(Jr, Sr) et les *suffixes propres aux organisations* (Inc., Co.) peuvent être annotés, en excluant ainsi les titres et les noms de rôle qui ne font pas partie de l'entité nommée et par conséquent ne doivent pas être annotés (Chef, Monsieur, Professeur,

12. Power Vehicle Innovation
13. Procès Verbal d'Installation
14. Points de Vue Initiatiques
15. Picardie Valves Industrie
16. Pour Votre Information

etc.). Pour Hatmi (2014) il s'agit d'un problème d'identification se heurtant à : (a) la *limite droite* de l'EN car les mots qui suivent le premier mot en majuscule ne sont pas capitalisés, par exemple, *La Fédération nationale de la Mutualité française* ; (b) la *coordination* qui unit deux ou plusieurs entités nommées en effaçant l'un des constituants communs, par exemple, *Brigitte et Emmanuel Macron instaurent peu à peu leurs habitudes au palais de l'Elysée* ; (c) l'*imbrication* lorsqu'une entité nommée est imbriqué dans une autre, par exemple, *l'Assemblée générale de l'ONU*.

Quant à la campagne ESTER (Meur *et al.*, 2004) les directives sont encore plus strictes puisque seul l'entité doit être annotée en excluant ainsi les déterminants, les articles, les adverbes, les étiquettes statutaires, etc., sauf si elles font partie intégrante de l'entité à annoter, par exemple, *Le Monde*.

6.2.4 Pourquoi extraire des entités nommées ?

Comme nous l'avons vu précédemment, la tâche de la reconnaissance d'entités nommées est le processus consistant à reconnaître des noms propres (entités) dans un texte et à leur attribuer des catégories pertinentes.

En terme général, la reconnaissance d'entités nommées fait partie des tâches les plus courantes du TAL. En effet, elle figure en première place dans de nombreuses applications traitant directement ou indirectement les noms propres et leurs propriétés. Elle constitue un élément clé des systèmes d'extraction d'informations, elle est essentielle pour la manipulation robuste de noms propres dans de nombreuses applications, le filtrage d'informations et la mise en relation d'informations (*information linking*). Ehrmann (2008) distingue deux types d'applications des EN, de natures différentes : (1) *indirecte*, interne comme partie d'un composant du traitement du langage qui bénéficie de cette information (le centre de l'application n'étant pas l'EN) et (2) *directe*, une chaîne de traitement avec une application directe particulière (le centre de

l'application est l'EN). Ci-dessous nous présentons certaines applications des EN :

6.2.4.1 Applications indirectes

La désambiguïsation lexicale

La désambiguïsation lexicale (*Word Sense Disambiguation* (WSD) en anglais) en TAL est une technique qui détecte le sens correct d'une unité lexicale ambiguë dans un contexte précis. Il parait que les mots fréquemment utilisés dans un langage présentent plusieurs sens en comparaison avec les mots moins fréquents (Pal et Saha, 2015). Le processus de restitution lexicale met en place une stratégie pour la détermination du sens d'une unité lexicale dans un contexte précis alors que cette unité a potentiellement plusieurs sens possibles. Mais quel est l'apport de la reconnaissance d'entités nommées lors de la désambiguïsation lexicale ? En effet, les informations sémantiques associées aux entités nommées peuvent contribuer à la désambiguïsation lexicale prenant la place d'un contexte pour l'unité ambiguë. Pour mieux visualiser ce constat nous citons deux phases :

Norwegian Airlines fait voler ses nouveaux avions vers les Etats-Unis.

Vincenzo Peruggia a volé la Joconde en 1911.

Pour les deux phrases l'unité verbale *voler* est une forme ambiguë qui définit l'action de : (1) effectuer un parcours déterminé par la voie des airs, dans l'espace et (2) s'emparer frauduleusement et quel que soit le procédé utilisé, de ce qui appartient à autrui, avec l'intention de le faire sien [17]. Dans la première phase, l'entité nommée *Norwegian Airlines* est une entité du type collectif <Organisation> ayant la place du sujet du verbe. L'entité *Vincenzo Peruggia*, dans la deuxième phrase, est un anthroponyme du type <Personne>, en combinaison avec l'entité *Joconde* qui est une entité ergonymique du type <Œuvre>. Ces arguments assignés

17. Définition selon : `http://www.cnrtl.fr/definition/voler`

aux verbe peuvent, en accord avec d'autres éléments dans les phrases, permettre de déterminer le sens approprié parmi les sens que le verbe voler peut avoir.

La coréférence

La résolution de la coréférence est une tâche assez importante pour de nombreuses applications du TAL qui impliquent la compréhension du langage naturel, comme par exemple le résumé automatique des documents, la tâche de question-réponse et l'extraction de l'information. Il s'agit de la tâche qui cherche toutes les expressions qui font référence à une même entité dans un texte. Autrement dit, on trouve des désignations différentes qui réfèrent au même individu, par exemple :

Paris est, avec sa banlieue, la capitale économique et commerciale de la France. Elle est également le chef-lieu de la région Île-de-France.

Les désignations attribués à *Paris* (capitale économique, capitale commerciale, elle, chef lieu de la région Île-de-France) sont des coréférences, car elle partagent le même référent. Ces désignations permettront de résoudre la coréférence avec partie interne du processus la reconnaissance de l'EN. Cependant, comme Désoyer *et al.* (2015) déclarent, ces systèmes sont encore aujourd'hui extrêmement rares pour la résolution de la coréférence en français, mise à part des systèmes à base de règles.

Fouille de données textuelles

La fouille de textes ou « l'extraction de connaissances » (en anglais text mining), définit un ensemble de traitements informatiques pour extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes produits. Les données textuelles qui contiennent des informations se montrent sous une grande diversité de formes (textes formels, tags, phrases

courts et bruités), qui peuvent ne pas être traitées par les méthodes classiques de la fouille de données.

Parmi certaines fonctions de la fouille de données textuelles nous trouvons que la reconnaissance d'entités nommées joue un rôle assez utile. Par exemple, la détection automatique des thèmes, qui vise à regrouper des textes en impliquant l'utilisation de méthodes de classification automatique sur des données textuelles en passant par les représentations vectorielles. Trouver donc les occurrences de parties spécifiques des textes est nécessaire, comme par exemple les entités nommées, afin de les illustrer séparément.

De plus, la détection d'EN peut être bénéfique pour la classification automatique des données textuelles à l'aide de modèles décisionnels, obtenus par apprentissage supervisé.

En ce qui concerne les messages courts postés sur des médias sociaux, chatbots [18] (agent conversationnel en français), relations service-client, la fouille de données se centre autour de l'identification de tendances à partir de ces messages et le classement thématique de textes afin de déterminer les fils de discussion d'intérêt. Par exemple, une association entre entités nommées des marques de produits et des mots peut offrir des indices sur l'appréciation globale et spécifiques des produits ou marques.

L'analyse de sentiments

L'analyse de sentiment (an anglais *opinion mining*) se réfère à la définition d'opinions, de sentiments et d'attitudes présentes dans un ensemble de données. Avec la grande diffusion des blogs et des réseaux sociaux, l'exploration des opinions et l'analyse du sentiment sont devenues

18. Le *chatbot* est un logiciel prenant la place d'une personne qui dialogue avec une autre personne en lui donnant l'impression de discuter elle-même avec une personne réelle. Historiquement, le premier chatbot est la machine ELIZA (Weizenbaum, 1966) réalisée 1966 par une équipe du laboratoire d'intelligence artificielle du MIT. L'objectif était la simulation un psychothérapeute rogérien en posant des questions sur la plupart des affirmations du patient.

un domaine d'intérêt pour de nombreuses recherches notamment en ce qui concerne les tweets.

La tâche d'analyse de sentiments consiste à rechercher, à la fois, les opinions ou les sentiments exprimés dans un document et à acquérir de nouvelles méthodes pour effectuer automatiquement cette analyse (Paroubek *et al.*, 2010). La détection d'entités nommées peut être étroitement liée avec l'analyse de sentiments, comme Pak (2012) dans ses travaux de thèse examine l'impact que les traits spécifiques aux entités nommées ont sur la classification des opinions minoritaires afin d'améliorer la classification de ce type d'opinion. En outre, Fraisse *et al.* (2013) soutiennent l'idée qu'*un système de fouille d'opinion doit pouvoir distinguer entre les indicateurs d'opinion exprimés de façon explicite, dans l'expression de l'opinion, et ceux qui sont liés à la présence des traits contextuels comme par exemple les EN*. Pour la meilleure performance de la classification des opinions minoritaires, il est nécessaire de réduire l'importance des traits contextuels (par exemple les entités nommés).

L'annotation en rôles sémantiques

L'annotation sémantique consiste à étiqueter les unités lexicales avec des liens qui associent une description sémantique. Le contenu d'un document peut être analysé de différentes manières afin d'être utile dans un but bien précis d'une application qui utilise les annotations descriptives issues de ces analyses. Dans le Web sémantique, quand nous nous intéressons à la structure logique du contenu d'un document, nous nous référons à l'annotation sémantique. La définition des rôles sémantiques dans un texte (agent, patient, objet, instrument, etc.) peut être influencée par les types des entités nommées reconnues (Nouvel, 2012).

6.2.4.2 Application directe

L'anonymisation

Nous avons déjà évoqué le fait que la reconnaissance d'entités nommées est étroitement liée

à la tâche d'anonymisation de données textuelles comportant des informations personnelles. L'anonymisation est cette procédure où tout élément permettant d'identifier un individu ou un groupe de personnes doit être supprimé ou substitué par une étiquette. Plusieurs activités de recherche ont comme objet l'anonymisation de données, comme par exemple, les données médicales et juridiques qui contiennent des noms, dates, lieux et autres éléments qui pourraient conduire à l'identification des personnes. Étant donné l'importance de cette tâche le traitement automatique doit être toujours accompagné d'une vérification manuelle.

En ce qui concerne l'anonymisation de données personnelles pour les SMS, nous avons déjà cité Patel *et al.* (2013) qui présentent deux méthodes d'anonymisation pour un corpus de SMS. Pour le langage SMS les entités qui doivent être anonymisées sont soumises à des variations lexicales et syntaxiques tout en contenant des diminutifs, des surnoms, etc. Par exemple : *Tu as eu mon mail bebenouche ?*. Le plus souvent l'anonymisation des adresses e-mail, des numéros de téléphone et des URL s'effectue à l'aide d'expressions régulières.

Systèmes de question-réponse

Les systèmes de réponse à des questions (Question answering (QA) en anglais) font partie des applications du TAL avec des systèmes qui répondent automatiquement aux questions posées par les humains en langue naturelle. Un exemple d'un tel système est un *chatbot*. La méthode entreprise par le système qui fournit une réponse à une question donnée peut nécessiter la détection des entités contenues dans la question posée, dans le but de parcourir des bases de connaissances et fournir la réponse adéquate (Nouvel, 2012). Selon le type de question (questions factuelles, questions booléennes, définitions, causes/conséquences, etc.) posée, une catégorisation sera utilisée afin de sélectionner la stratégie pour répondre à cette question. Exemple d'une question-réponse contentant une EN :

- Question : Qui a écrit Les Travailleurs de la mer ?

- Réponse : Victor Hugo a écrit Les Travailleurs de la mer.

Comme Ehrmann (2008) le signale, lors de la chaîne de traitement les entités nommées ont un rôle est assez important car elles aident, dans une première phase, à spécifier le type de la réponse attendue et, par la suite, à repérer la réponse adéquate. Cependant, une telle tâche en langage SMS devrait surmonter les divers spécificités du langage. Prenons par exemple la question suivante :

- Question : Dans quelle région se trouve le Mnt Snt-Michel ?

Le système doit dans un premier temps détecter le type de question (questions évaluatives/comparatives) et en fonction de ceci, le type de réponse attendue. Dans un deuxième temps, il cherche le *focus* de la question qui correspond à la propriété ou l'entité recherchée par la question, pour notre cas *Mnt Snt-Michel*. Au final, le système doit détecter le thème de la question (ici, c'est la *région*) dans le but de rechercher les documents susceptibles de répondre à la question. La difficulté du système pour réussir avec succès est de reconnaître, dès la deuxième étape, l'entité nommée *focus* dans sa description non standard.

La traduction automatique

Dans le domaine de la traduction automatique la reconnaissance d'entités nommées est sujet non seulement de la détection mais aussi de la traduction d'EN. Comme McNamee *et al.* (2011) l'affirment, les noms propres ne suivent pas toujours les règles d'orthographe. Les difficultés de la reconnaissance d'entités nommées s'identifient en premier lieu (*c.f.* 6.2.3). Notamment, avec les entités du type association qui ne sont pas *orthographiés de façon homogène* lors du processus de traduction nous nous apercevons que les pratiques sont assez variées (Grass et Maurel, 2008). De même, certaines entités assez connues peuvent ne pas avoir de traduction ou avoir une traduction vocalisée ou littérale (*Médecins sans frontières* = *Ärzte*

ohne Grenzen ou *Secours catholique* pour *Caritas France*[19], *New York* en français pour *New York* mais *Νέα Υόρκη* en grec). En effet la traduction d'entités dépend de la source et de la cible, ainsi, dans tous les cas, il faut toujours respecter les normes de chaque langue.

Nous pouvons donc nous rendre compte qu'une telle tâche appliquée au langage SMS nécessiterait une étude plus approfondie pour établir initialement une stratégie de détection et par la suite d'attribution en langue cible.

L'aide à la rédaction : saisie intuitive et autocomplétion

L'aide à la rédaction répond au processus de suggestion et correction d'un texte saisi à l'aide d'un éditeur de texte. Quand cette tâche est relative à un appareil de communication électronique nous l'appelons *saisie intuitive*, *prédictive* ou *T9* auparavant (acronyme de *text on 9 keys*, en français *texte sur 9 touches*). La saisie intuitive s'effectue grâce à un dictionnaire intégré à la mémoire de l'appareil où la plupart des mots communément utilisés sont répertoriés. L'écriture intuitive permet la composition plus rapide des messages en appuyant une seule fois sur chaque touche du clavier de l'appareil. Ce procédé avancé permet la reconnaissance automatique des mots qui correspondent aux combinaisons de graphies saisies. Ainsi, le procédé de l'écriture intuitive permet le remplacement ou la complétion d'un texte avec des mots saisis partiellement erronés ou de manière incomplète.

L'*autocomplétion* (*word completion* en anglais), également dérivé de l'aide à la rédaction dans le monde électronique, est la tâche qui devine des mots ou même des expressions saisies par l'utilisateur d'un appareil de communication électronique en saisissant partiellement les lettres nécessaires. Derrière l'autocomplétion se trouve un dictionnaire comprenant des expressions et des phrases fréquemment utilisées, qui permet d'exploiter le contexte du mot saisi afin de déterminer ce qui est attendu.

19. Exemples tirés de : Grass et Maurel (2008),

La reconnaissance d'entités nommées est effectuée lorsque la saisie intervient au moyen d'appareil de communication électronique. En effet, même si l'aide à la correction est faite pour limiter le temps de saisie et les fautes, divers éléments peuvent être saisis de façon erronée. Par exemple, l'entité patronymique *Decroix* saisie par un utilisateur en tant que *2croix* nécessiterait la définition d'une stratégie d'autocomplétion différente.

6.3 La reconnaissance d'entités nommées dans les SMS

Les trois approches dominantes dans le domaine de la reconnaissance d'entités nommées se divisent en celles qui sont centrées sur (a) les données, les techniques d'apprentissage (statistiques) à l'aide de données disponibles grâce aux campagnes ACE05 [20], Enron (Minkov *et al.*, 2005) et CoNLL03 (Tjong Kim Sang et De Meulder, 2003b), (b) les symboliques à base de règles, par exemple, les transducteurs à nombre fini d'états et (c) les hybrides combinant les techniques des méthodes orientées connaissances et celles orientées données.

La plupart des recherches sur la reconnaissance d'entités nommées reposent sur des corpus de textes standard qui utilisent un langage formel, par exemple, des textes journalistiques ou scientifiques (Bunescu et Mooney, 2004, Martineau *et al.*, 2007, Friburger, 2002, Ehrmann, 2008, Fourour, 2002, Poibeau, 1999, McCallum et Li, 2003, Etzioni *et al.*, 2005). Ces types de textes sont écrits pour un public assez large, et leurs auteurs ont été vigilants lors de la préparation du texte. Dans la thèse de Friburger (2002), en ce qui concerne la reconnaissance des entités nommées, nous trouvons un état de l'art sur les trois principales méthodes employées par ces systèmes. Les tableaux 6.3, 6.4 et 6.5 contiennent des tables récapitulant les travaux cités par Friburger pour les systèmes à base de règles, d'apprentissage et hybrides pour le français et l'anglais. Plusieurs systèmes de REN sont présentés en détail dans Poibeau (2005), Friburger (2002), Fourour (2004), de ce fait, nous n'exploitons pas davantage ces travaux mais

20. http ://projects.ldc.upenn.edu/ace/

nous nous intéressons aux travaux effectués sur la reconnaissance d'entités nommées à base de données dégradées et courtes, comme les SMS.

Les systèmes à base de règles :
En anglais
FUNES (Figuring-out Unknown Nouns from English) (Coates-Stephens, 1992) FASTUS (Hobbs *et al.*, 1997) obtient une F-mesure de 94% à MUC 6 pour l'extraction des entités nommées. PNF (Proper Name Facility) (McDonald, 1996) LaSIE (LArge Scale Information Extraction) (Wakao *et al.*, 1996). Les résultats de rappel sont de 91%, 90%, 88% pour les organisations, personnes, lieux et la précision est resp. de 91%, 95%, 89% sur des textes du Wall Street Journal pour MUC 6. Nominator (Wacholder *et al.*, 1997) extrait 91% des noms propres avec une précision 92%, seuls 79% des noms propres trouvés sont catégorisés par le système. NetOwl Extractor (système commercial de Isoquest Inc.) (Krupka et Hausman, 2005) obtient une F-mesure de 96,42% pour MUC 6 et seulement 91,32% pour MUC 7.
En français
Exoseme (Landau *et al.*, 1993) est un système de filtrage avec un module sur les entités nommées en français. Ce système obtient 90% de rappel et en catégorise correctement 85% sur des dépêches AFP. ThingFinder (Trouilleux, 1998)

TABLE 6.3 – État de l'art des systèmes à base de règles

Comme nous l'avons déjà évoqué, le langage SMS reflète également d'espaces communicationnels comme les messages publiés sur les réseaux sociaux Twitter et Facebook ou autres, le courrier électronique, les forums de discussion et les chats. Les points caractéristiques de ce discours sont, selon Marcoccia (2016), l'utilisation de l'orthographe non standard, les spécificités morphologiques autour de la formation des mots, l'apparition de nouveaux termes au niveau lexical et l'écart avec la syntaxe standard qui résulte du discours *fragmenté* et *télégraphique*. En effet, à cause des propriétés propres, à ces discours, la reconnaissance et la catégorisation des entités nommées utilisant des techniques triviales est inefficace, bien plus, si l'entité nommée est elle-même affectée par des phénomènes graphiques ou phonétiques. Ainsi,

Les systèmes à apprentissage : :
En anglais
Alembic (Aberdeen *et al.*, 1995). BBN IdentiFinder (Miller *et al.*, 1999). MENE (Maximum Entropy Named Entity) (Borthwick *et al.*, 1998). (Collins et Singer, 1999) obtiennent 91% de rappel et 83% de précision sur le New York Times. Answer extraction (Abney *et al.*, 2000)

TABLE 6.4 – État de l'art des systèmes à apprentissage - suite

Les systèmes hybrides : :
En anglais
LTG system (Language Technology Group) (Mikheev *et al.*, 1998) est le meilleur à MUC 7, nous lui réservons une section afin d'expliquer comment il procède pour extraire les entités nommées. (Lin *et al.*, 1998) obtient une F-mesure de 86,37%
En français
(Senellart, 1998) en français et en anglais. SemTex (Poibeau, 1999) en français et anglais, il annonce 80% de rappel sur le journal Le Monde. (Fourour, 2002) obtient, avec Nemesis , 90% de rappel et 95% de précision.

TABLE 6.5 – État de l'art sur les systèmes systèmes hybrides - suite

autant les preuves internes qu'externes ne sont pas exploitables (Hatmi, 2014). L'idée qui parait plus évidente, dans ce cas là, est de passer par la normalisation pour améliorer l'efficacité de REN. Effectivement, Wan *et al.* (2011) mentionnent que la F-mesure de l'étiqueteur Stanford REN (Finkel *et al.*, 2005), entrainée sur l'ensemble des données de CoNLL03(Tjong Kim Sang et De Meulder, 2003b) atteint des performances de pointe sur cette tâche, présente une diminution de 90,8% (Ratinov et Roth, 2009) à 45,8% lorsqu'il s'applique à des messages tweets.

Un certain nombre de travaux qui concernent notamment sur la REN pour l'anglais sont

fondés sur les entités nommées issues de messages publiés sur ces réseaux. Ritter *et al.* (2011) ont développé un système qui exploite des champs aléatoires conditionnels [21] (Conditional Random Fields ou CRFs). Afin de segmenter des entités nommées et puis à l'aide d'une approche supervisée réussir la classification d'EN. En utilisant les CRFs en combinaison avec un classificateur basé sur l'algorithme des k plus proches voisins (k-nearest neighbor ou KNN, k-NN), Liu *et al.* (2011) proposent un modèle sous forme d'apprentissage semi-supervisé. Le classificateur effectue un pré-étiquetage pour recueillir des preuves globales à travers les tweets, tandis que le modèle CRF effectue un étiquetage séquentiel pour capturer des informations fines codées dans un tweet. Une année plus tard, Liu *et al.* (2012) proposera un nouveau modèle graphique pour mener simultanément la REN et la normalisation d'entités nommées (NEN) sur plusieurs tweets. Leur modèle introduit une variable aléatoire binaire pour chaque paire de mots avec le même lemme sur des tweets similaires, dont la valeur indique si les deux mots apparentés mentionnent la même entité. Après une évaluation sur l'ensemble des données annotées manuellement, ils obtiennent une amélioration de la *baseline* qui gère ces deux tâches séparément, en augmentant le F-mesure de 80,2% à 83,6% pour le REN et la précision de 79,4% à 82,6% pour la NEN, respectivement. Dans la plus récente recherche que nous identifions Yamada123 *et al.* (2015) faisant partie des travaux de Baldwin *et al.* (2015) pour le WNUT2015 (Workshop on Noisy User-generated Text 2015) sur la reconnaissance d'entités nommées sur Twitter qui a obtenu les meilleurs scores parmi les 8 groupes de travail. Leur méthode améliore les performances de la reconnaissance d'entités nommées sur Twitter en utilisant la liaison d'entités qui est une méthode pour détecter les mentions d'entités dans le texte et les rattacher à des entrées correspondantes dans des bases de connaissances telles que Wikipedia. La méthode est basée sur l'apprentissage par machine supervisé et utilise les connaissances obtenues à partir de plusieurs bases de connaissances libres.

21. Selon la définition de Hebert *et al.* (2012), il s'agit d'un *processus stochastique qui modélise les dépendances entre un ensemble d'observations discrètes réalisées sur une séquence discrète (une séquence de mots) et un ensemble d'étiquettes (analyse morphosyntaxique)*. Un des plus grands avantages est que les CRFs permettent d'éviter le problème d'une estimation biaisée rencontrée avec les modèles conditionnels de Markov Cachés.

D'autres recherches se focalisent sur la REN dans les messages e-mail ou les blogs. Jansche et Abney (2002) explorent une approche biphasée simple composée a) d'une grammaire fabriquée à la main pour la proposition de candidats et b) d'un classificateur pour l'extraction d'informations à partir de la messagerie vocale. Minkov *et al.* (2005) réalisent une étude expérimentale sur la reconnaissance des noms de personnes dans les courriers électroniques au travers de deux méthodes basées sur l'apprentissage automatique en utilisant les CRFs et une méthode à base de perception pour l'apprentissage du modèle de Markov Caché (MMC). Les expériences pour ce système montrent que la performance de la F-mesure de baseline peut être considérablement améliorée, d'une part en introduisant des fonctionnalités d'apprentissage spécifiques au courrier électronique et, d'autre part, par une nouvelle méthode de rappel, qui exploite le fait que dans les textes de courrier électronique, les noms sont généralement répétés plusieurs fois dans le corpus. Wan *et al.* (2011) aborde la tâche de la reconnaissance d'entités nommées dans les commentaires dans des nouvelles en chinois sur le Web. Trois schémas sont exploités pour recueillir des entités utiles à partir de nouvelles sur le web. Les informations d'entités s'incorporent dans un algorithme basé sur les CRFs pour la reconnaissance des entités nommées dans les commentaires.

Les REN dans des SMS comparativement à la REN dans des tweets, où la disposition d'entités dans les phrases peut être beaucoup plus structurée et les entités étiquetées par des symboles (#, @), est une tâche légèrement plus complexe. Le système de REN pour les SMS en suédois de Ek *et al.* (2011) se concentre à l'extraction d'entités du type lieux, noms, dates, temps et numéros de téléphone, à l'aide d'expressions régulières couplées avec des classificateurs qui utilisent la régression logistique. Leur système atteint une F-mesure de 86% pour les correspondances strictes et 89% pour les matches partiels. Polifroni *et al.* (2010) ont développé un ensemble de données et un classificateur pour reconnaître les entités nommées dans la parole, issues de la lecture automatique de SMS. Au travers de données réelles et un grand nombre d'entités nommées, ils ont construit un corpus de données d'entrainement et de test. Ces données ont été utilisées pour développer le classificateur dans le but d'identifier

des noms et des lieux en utilisant le principe d'entropie maximale, dont la sortie était une estimation de probabilité pour chaque mot de chacune des classes. Les résultats de performance du modèle atteint 87.2% de F-mesure pour les noms et 88.4% pour les lieux.

Cependant, nous remarquons que les travaux réalisés sur les SMS et les tweets en français sont relativement faibles. En effet, l'extraction d'entités temporelles dans les messages SMS de Weiser *et al.* (2011) est une seule recherche traitant la REN en français dans les SMS. Le système traite les messages sans passer par la normalisation en se basant sur la typologie et les grammaires locales temporelles de Kevers (2011) initialement développées pour la langue française standard qui ont été, par la suite, adaptées pour les besoins des SMS. Le rappel comme résultat atteint 65% tandis que la précision 79%.

Un autre traitement étroitement lié à la REN est l'anonymisation d'entités nommées dans les SMS. Comme nous l'avons évoqué dans le chapitre 3.2.3.1, la phase d'anonymisation de données sensibles (nom, prénom, adresse postale, codes, numéros de téléphone, etc.) est primordiale pour la collecte, le traitement et la diffusion de corpus de SMS. Une recherche centrée sur la détection d'EN contenues dans des SMS est celle de Patel *et al.* (2013) qui présente deux méthodes d'anonymisation d'un corpus de SMS dont une, nommée *Seek&Hide*, basée sur des règles heuristiques et l'utilisation de dictionnaires ; et une autre fondée sur des méthodes d'apprentissage supervisées à l'aide d'arbres de décision. La combinaison des deux approches présente séparément chaque fonction de chaque approche toujours dans le but de réduire le temps d'étiquetage manuel. La proposition est de croiser les résultats obtenus par les deux approches. Cependant, à l'égard des résultats obtenus (F-mesure de 41%) le niveau de la détection de la classe "à anonymiser" est faible.

Zenasni *et al.* (2016) ont développé une approche pour la reconnaissance d'entités spatiales (toponymes) issues de SMS. L'approche se fonde sur l'identification et l'extraction de nouvelles variantes d'entités spatiales, correspondant à des formulations existantes, en calculant la similarité (String matching (Maedche et Staab, 2002)) entre le dictionnaire d'entités spatiales

et les mots issus d'un corpus de SMS. Cependant le problème d'ambiguïté remonte en surface lorsque par exemple il faut traiter une entité comme *Orange* en tant que toponyme et *Orange* en tant que société de télécommunications.

Raymond et Fayolle (2010) utilisent comme texte bruité les transcriptions automatiques de la parole issues des documents multimédia contenant de l'oral. En effet, les transcriptions de la parole sont partiellement similaires au langage SMS, notamment concernant la phonétisation d'unités lexicales. Les trois méthodes qu'ils emploient sont basées sur des algorithmes d'apprentissage automatique : les machines à vecteurs de support, les transducteurs à états finis et les champs conditionnels aléatoires. Les trois systèmes obtiennent tous des résultats inférieurs à 60% en *Slot Eroor Rate* (SER) sur les données d'évaluation de la campagne d'évaluation ESTER 2 [22].

6.3.1 Une typologie d'entités nommées pour les SMS

Dans la partie 6.2.2, nous avons présenté les catégories issues de campagnes fondées sur la reconnaissance d'entités nommées. Nous considérons que chaque tâche de détection d'EN a des spécificités et des besoins bien particuliers selon le type de texte à annoter. En effet, il n'existe pas une typologie universelle couvrant tous les besoins des chercheurs, malgré la typologie exhaustive de Sekine *et al.* (2002).

L'annotation d'entités nommées issues de SMS doit répondre à un certain nombre de besoins, elle ne nécessite pas d'être exhaustive et fine mais elle doit être pertinente, applicable, et exploitable par des systèmes, comme nous l'avons vu en partie 6.2.4.

En se basant sur les typologies existantes, nous voulons présenter une typologie pour les SMS (tableau 6.6 et 6.7). Pour la définition de la typologie nous prenons en compte les entités qui doivent être anonymisées dans un texte SMS (noms, adresses postales, codes, etc., pour

22. `http://www.afcp-parole.org/camp_eval_systemes_transcription/`

plus d'informations *cf.* 3.2.3.2). Le choix de catégories est fait en fonction d'entités nommées que nous rencontrons le plus souvent dans les SMS et en lien avec les applications possibles (internes et externes).

En effet, les entités les plus fréquentes dans les SMS sont les noms de personnes. Ainsi, nous proposons la catégorie *anthroponymique* qui inclut les noms des personnes, les surnoms, qui sont également assez fréquents dans les SMS, les éthnonymes et les noms de célébrités (chanteurs, acteurs, personnes de la showbiz etc.). Dans un deuxième temps, nous nous intéressons aux entités *toponymiques*. La catégorie *toponymique* inclut les entités nommées géographiques permettant une géolocalisation. Ainsi, nous trouvons le pays, la ville, la région, la voie, une adresse complète et les espaces publics qui correspondent à des endroits particuliers urbains, des espaces de passage et de rassemblement qui sont dédiés à l'usage de tous. La catégorie *organisations* inclut les anthroponymes collectifs qui décrivent des entités, telles que les entreprises, les filiales, les équipes et les organisations éducatives. Les entités *temporelles* sont divisées en cinq types permettant l'identification d'une expression de temps (date, heure, jour, mois, temps). En association avec d'autres entités, les entités temporelles peuvent servir pour l'identification d'évènements, par exemple un rendez-vous ou une fête (14 juillet). Les entités du type *ergonymique* correspondent à des entités inanimés qui désignent une oeuvre, un objet, une marque qui fait référence à un produit et un produit d'une réalisation humaine, par exemple un film ou une série télévisée. Les ergonymes pourraient permettre l'observation globale de l'appréciation d'un produit en lien avec la tâche de l'analyse de sentiments dans un message. Les *pragmonymiques*[23] désignent un événement, par exemple une manifestation, une fête etc. Les *identifiants numériques* permettent l'extraction d'informations pour les numéros de téléphone, codes, url, etc. et pourraient servir pour la tâche de l'anonymisation automatique. Enfin, la catégorie *valeurs* inclut les valeurs physiques et monétaires.

23. Les termes ergonymique et pragmonymique (noms propres d'événements provenant de la fusion de praxonymes et phénonymes) sont deux termes empruntés à la typologie primaire de *Prolex* (Tran et Maurel, 2006) elle-même inspirée par la classification linguistique de Bauer (1985), reprise par Grass (2000).

La typologie que nous avons introduite met en évidence les relations entre les entités nommées que nous pouvons trouver dans les SMS. Elle pourra permettre le développement d'outils de traitement automatique des langues comme le traitement des coréférences, la recherche d'information, la traduction automatique et d'autres.

Les anthroponymes

Personne	Anne, Pauline Decroix, Antoine Martineau
Surnom	titine, cloclo, cece
Ethnonymes	français, grenoblois, niçois
Célébrités	Charlize Theron, Neymar, Marion Cotillard

Les toponymes

Pays	France, Allemagne, Angleterre
Ville	Lyon, San Francisco, Rome
Région	Bretagne, Normandie, Guadeloupe
Voie	rue Paul Cocat, avenue Général de Gaulle, rue Descartes
Adresse	15 rue Ampère 77420 Champs sur Marne
Espace public	gare de Lyon, jardin du Luxembourg, Bastille

Les organisations

Association	Médecins sans frontières, Solidarités International
Entreprise	Microsoft, Sony, Symantec
Équipes	Équipe de France de Handball, les Bleus
Éducatives	Université Paris-Est, UGA, Université Paris-Sorbonne

Les temporelles

Date	19 septembre 1984, Lundi 26 juin 1989
Heure	05h35, 17h45, 6h20, 20 :12
Temps	deux heures, matin, soir, lendemain, vingt minutes
Mois	juin, juillet, août
Jour	lundi, mardi, mercredi, jeudi, vendredi, samedi, dimanche

Table 6.6 – Typologie d'entités nommées SMS

Les ergonymes

Objet	Mac, iPhone, PSP
Œuvre	Guernica, Statue de la Liberté, Les Quatre Saisons
Marque de produits	Coca-cola, Stabilo, Apple
Film–Série–TV	Le Trône de fer, Tfou, Rendez-vous en terre inconnue

Les pragmonymes

Fête	Noël, Saint Valentin, 1er mai
Manifestation	Salon d'Agriculture, Coupe du monde de la FIFA
Météorologie	canicule, inondations, orages
Catastrophe	Fukushima, Tchernobyl, 11 septembre

Les identifiants numériques

Téléphone	+33 672030789, 0672030789
E-mail	bluemix@emm.ibmmail.com
Url	`http://www.ibm.com/privacy/fr/fr/`
Code	774204, 5047Ajk69
Mentions	@UNICEF, #India, #YouthDay

Les valeurs

Valeur physique	40°C, 5.5PH, 50 kilos
Valeur monétaire	£24.00, 50€, 5 euros

TABLE 6.7 – Typologie d'entités nommées SMS - suite

Nous voulons introduire la classe emoji [24] qui peut dans certains cas prendre la place d'une entité nommée. Les emojis sont des idéogrammes qui peuvent être considérés comme l'évolution naturelle des émoticônes. Il existe une grande variété de types d'emojis pouvant designer les expressions faciales, les animaux, les objets ou les lieux. Nous les trouvons dans les plateformes comme Twitter, Instagram et WhatsApp mais aussi dans les applications pour l'écriture de SMS, sous des formes diverses selon la compagnie de fabrication du téléphone portable. Ils sont utilisés pour communiquer des choses mais aussi le plus souvent, des sentiments dans un contexte visuel et bref. Pour inclure cette classe dans la typologie, nous nous sommes inspirés

24. Il s'agit d'un terme japonais pour désigner les émoticônes. Signifiant à l'origine pictogramme, le mot emoji signifie littéralement « image » (e) + « lettre » (moji) source : `https://fr.wikipedia.org/wiki/Emoji`

du travail de Barbieri *et al.* (2016) où l'exemple d'un message Twitter contient un emoji (figure 6.3). En effet, il s'agit du drapeau des États-Unis qui dans ce cas remplace l'entité nommée toponymique *Amérique*, du type <PAYS>.

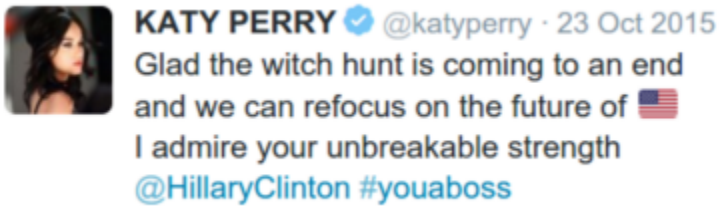

FIGURE 6.3 – Exemple de message tweet contenant un emoji sur Barbieri *et al.* (2016)

Nous introduisons ainsi une typologie d'entités nommées pour les emoji contenant la catégorie *pays*, *symboles*, *voyages et endroits*, *activités* et *événements*. Notre typologie n'est ni exhaustive, ni fine, ni très analytique. Les emojis peuvent être ambiguës, cependant, en combinaison avec d'autres éléments du texte ils peuvent contribuer à l'extraction d'informations dans des textes courts.

Les emoji

Voyages & endroits	: cinéma, : statue de Liberté , : park, campagne
Activités	: volleyball, : jeux vidéo, : basketball, : ping pong
Pays	: Grèce, :France , :Russie
Événements	: anniversaire, : mariage, : Noël

FIGURE 6.4 – Typologie d'entités nommées pour les emoji

6.4 Conclusion

Ce chapitre a été centré sur l'entité nommée, étroitement liée avec l'extraction d'information et la reconnaissance des entités nommées. Plus précisément, le chapitre décrit d'un point de vue théorique la définition d'entité nommée, ses catégories et leurs problèmes de catégorisation. Dans un deuxième temps, nous avons présenté un état de l'art sur les travaux réalisés sur les documents formels et bruités. La troisième section du chapitre présente un certain nombre d'applications qui implique la tâche de la reconnaissance d'entités nommées. La dernière partie concerne la proposition d'une typologie pour la reconnaissance d'entités nommées issues des SMS.

Le travail de thèse se poursuit dans le chapitre 7 par un troisième volet autour de l'extraction automatique d'information contenue dans les SMS. Plus précisément, notre hypothèse étant que la normalisation morphosyntaxique des SMS permettrait d'améliorer les performances des méthodes traditionnelles pour l'identification d'entités nommées. Après avoir défini une classification d'entité nommées adaptées aux SMS, nous nous sommes basés sur une série d'expérimentations afin d'observer la qualité d'annotation de trois systèmes issus de l'état de l'art pour l'identification d'entités nommées en français standard : OpeNER (fondé sur l'apprentissage automatique), CasEN (fondé sur l'ingénierie des grammaires) et mSX (fondé sur l'ingénierie des grammaires et l'apprentissage automatique).

Chapitre 7

Apport de la normalisation à la reconnaissance d'entités nommées

Sommaire

7.1 Introduction 167

7.2 Systèmes d'extraction d'entités nommées appliquées aux SMS . . 168

7.2.1 Présentation du corpus 168

7.2.2 Système Baseline 172

7.3 Évaluation 176

7.3.1 Mesures de performance 176

7.3.2 Évaluation et analyse des résultats 177

7.4 Conclusion 184

7.1 Introduction

Après avoir défini les méthodes utilisées dans le chapitre précédent, nous voulons évaluer la performance des systèmes de reconnaissance d'entités nommées dans une série de tests.

La reconnaissance d'entités nommées dans des textes informels et bruités se retrouve dans plusieurs applications, notamment, pour les courriers électroniques, pour l'ajout et la planification semi-automatique de réunions et d'événements, l'analyse des réseaux sociaux, etc. Cependant, le traitement automatique de ces textes est plus compliqué que celui des textes formels. De ce fait, les application REN existantes peuvent nécessiter des modifications pour mieux fonctionner sur ce type de textes.

Notre objectif, en lien direct avec le modèle hybride pour la normalisation de SMS, est de prouver que les performances des systèmes présentent des améliorations significatives lorsqu'ils sont appliqués sur des SMS automatiquement normalisés.

7.2 Systèmes d'extraction d'entités nommées appliquées aux SMS

7.2.1 Présentation du corpus

Pour notre expérimentation, qui permet d'évaluer les méthodes d'identification automatique, nous nous basons sur deux partitions du corpus de SMS alpes4science. Il s'agit de deux sous-corpus chacun divisé en types de messages. Les deux sous-corpus sont composés de 200 SMS chacun de *SMS bruts*, de *SMS transcrits manuellement* par les annotateurs et un troisième corpus résultat de l'approche hybride que nous avons présenté dans le chapitre 5, contentant les SMS normalisés automatiquement. Eux-mêmes comportent trois versions de ces messages : la version originale, la version avec tous les caractères en minuscules et la version avec les caractères initiaux de chaque unité en majuscule (tableau 7.1). La variation des corpus utilisés devrait mettre en évidence les points faibles de chaque système et la place importante que la casse prend dans les performances de ces systèmes.

Toutes les versions du corpus ont été produites automatiquement. Nous disposons égale-

	SMS bruts
$V_{original}$	C le 17 q t viens à paris ?
$V_{miniscule}$	c le 17 q t viens à paris ?
$V_{initial\ maj.}$	C Le 17 Q T Viens A Paris ?
	SMS normalisés automatiquement
$V_{original}$	c'est le 17 que tu viens à paris ?
$V_{miniscule}$	c'est le 17 que tu viens à paris ?
$V_{initial\ maj.}$	C'Est Le 17 Que Tu Viens A Paris ?
	SMS transcrits manuellement
$V_{original}$	C'est le 17 que tu viens à Paris ?
$V_{miniscule}$	c'est le 17 que tu viens à paris ?
$V_{initial\ maj.}$	C'Est Le 17 Que Tu Viens A Paris ?

TABLE 7.1 – Les types du corpus d'expérimentation

ment de la version *référence* de SMS qui comporte le corpus annoté avec les entités nommées que nous cherchons à détecter. Le corpus de référence a été annoté manuellement et sera le point de repère pour l'évaluation de chaque production automatique par les logiciels que nous décrivons par la suite.

Les entités nommées ont été annotées avec les catégories anthroponymiques (Personne), et toponymiques selon la typologie que nous avons proposée dans la partie 6.6 et 6.7. Ci-dessus nous trouvons un exemple d'un SMS annoté avec les étiquettes <Pers> et <Loc>. Pour des raisons de confidentialité tous les noms et prénoms contenus dans le message ont été changés et choisis aléatoirement :

Désole <Pers>**Marie**</Pers> ! Je suis en tournage a <Loc>**paris**</Loc> ! Tu as trouve un compatriote pour aujourd'hui ? ? ? T'es célib ? C'est toi <Pers>**Anne Moss**</Pers> ? ? ;-)

L'évaluation a été réalisée sur neuf fichiers. Les évaluations réalisées sur le corpus mesurent la détection d'entités nommées strictes (entité et type), autrement dit, aucune erreur de bornes et de typage n'a été acceptée.

Caractéristiques du corpus *référence* :

Le corpus de référence a été annoté manuellement avec les entités nommées *<pers>* et *<loc>*. Il est constitué de 200 SMS, soit 7 802 tokens. Le tableau 7.2 résume en chiffres les entités nommées étiquetées dans les corpus. Au total, nous avons identifié 60 entités, dont 34 correspondent aux entités nommées du type personne et le reste aux toponymes.

Caractéristiques	*n*
Taille	16,6 Ko
Tokens	7 802
Formes	3 084
Entités	
<pers>	34
<loc>	26
Total	60

TABLE 7.2 – Caractéristiques du corpus de référence

Caractéristiques du corpus *baseline* :

Nous appelons corpus *baseline* le corpus de SMS brut équivalent aux SMS de référence. Ce corpus est important pour la phase d'évaluation car il joue un rôle de point de repère pour la qualification de la performance des systèmes.

En effet, nous allons pouvoir juger, en comparaison avec les résultats des performances sur les autres corpus s'il y a une amélioration au niveau de la reconnaissance entre les SMS bruts et les SMS normalisés automatiquement.

7.2.1.1 Normalisation de la casse pour le corpus normalisé

Lors de la saisie d'un message, il est courant que l'utilisateur ne respecte pas les normes de casse après l'utilisation de la ponctuation. De même, les entités nommées contenues dans les SMS sont souvent entièrement écrites en majuscules ou en minuscules. Lors de la phase de la normalisation du corpus par l'approche hybride du corpus décrite au chapitre 5 aucune intervention n'a été faite sur la casse des messages. Cependant, pour la performance optimale de systèmes autour de la reconnaissance d'entités nommés la casse joue un rôle très important et la mise en forme du corpus est une tâche primordiale.

La méthode que nous avons utilisée consiste dans un premier temps à la mise en forme de caractères majuscules en début de phrase et après le point d'interrogation, le point d'exclamation et les points de suspension [1]. La deuxième partie consiste à normaliser la casse des entités nommées contenues dans les SMS. Cette tâche s'avère un grand défi puisqu'elle nécessite la détection préalable de ces entités afin de leur restituer la casse adéquate. Notre méthode se fonde sur l'idée d'identifier uniquement les entités nommées qui ne présentent pas d'ambiguïtés avec les noms communs, par exemple : *émilie/Émilie*, *hélène/Hélène*, et pas, *victoire/Victoire*, *aimée/Aimée*, *merci/Merci* etc. En se basant entièrement sur des ressources lexicales capables de couvrir un grand nombre d'entités nommées anthroponymiques (personne) couplées avec des dictionnaires de mots communs, par défaut d'Unitex, nous avons transformé en majuscule la lettre initiale d'entités nommées identifiées. La figure 7.1 illustre un extrait de la concordance en mode *merge* d'Unitex avec la normalisation de casse pour les entités du type personne.

1. Pour précision le corpus normalisé ne contient aucune émoticône car il pourrait générer de l'ambiguïté entre la ponctuation et ces caractères

petit drame entre julie et emilie Emilie dans la soirée mais au
a eu un petit drame entre julie Julie et emilie dans la soirée m
vous 2 j'ai pas le numéro de margotte Margotte!!). Ben avar
'accord ! Elle est ou miss noemie Noemie ? Truc super bon!!

FIGURE 7.1 – Extrait de concordance en mode "merge" de normalisation de casse sur Unitex

7.2.2 Système Baseline

7.2.2.1 API

L'API (Application Programming Interface en anglais) qui en français se traduit par *Interface de Programmation Applicative* est une solution informatique qui donne accès à des applications pour communiquer entre elles et s'échanger mutuellement des services. Il consiste en un ensemble de fonctions qui, à l'aide d'un langage de programmation, garantissent l'accès aux services d'une application. L'API peut faciliter le développement d'un programme informatique en fournissant tous les éléments constitutifs, qui sont ensuite mis en place par le programmeur. Une API peut être pour un système basé sur le Web, un système d'exploitation, un système de base de données, un matériel informatique ou une bibliothèque de logiciels.

L'avantage d'une API est la possibilité de pouvoir utiliser un programme sans devoir passer par le fonctionnement complexe d'une application. Selon Webopedia [2], il s'agit d'un ensemble de fonctions, de protocoles et d'outils pour la construction de logiciels. Une API spécifie comment les composants logiciels doivent interagir. De même, les APIs sont utilisées lors de la programmation des composants de l'interface utilisateur graphique (GUI). Il existe plusieurs types d'API pour les systèmes d'exploitation, les applications ou les sites Web.

2. `http://www.webopedia.com/TERM/A/API.html`

Nero

Le Nero [3] (Named Entities Recognition–Online) est une API développée par Christian Raymond à IRISA/INSA à Rennes. Il s'agit d'un service pour la reconnaissance d'entités nommées en français, capable de détecter dans un texte des entités du type : personne, fonction, organisation, lieu, production humaine, temps et montant.

Nero est hérité du travail de Raymond et Fayolle (2010) pour la reconnaissance d'entités nommées de la parole transcrite automatiquement. La particularité de la transcription automatique de la parole est que les documents transcrits automatiquement ne sont que peu structurés et certains mots transcrits sont erronés avec un *taux d'erreur de mots qui peut varier de 5% à plus de 50% selon le document et les conditions de transcriptions*, selon Raymond et Fayolle. Le système est capable de détecter les entités dans de documents bruités, comme les transcriptions automatiques. Il est basé sur deux méthodes d'apprentissage automatique, couramment utilisées pour la reconnaissance d'entités nommées. Il s'agit d'une approche fondée sur les champs conditionnels aléatoires et sur une combinaison de trois transducteurs à états finis en exploitant, parmi les caractéristiques textuelles, les mots eux-mêmes, avec des informations supplémentaires (connaissance préalable de leur classe ou leur importance, et/ou information morpho-syntaxique).

7.2.2.2 Logiciels libres

Le terme a fait son apparition en 1980, il est dû au fondateur du *projet GNU* (General Public License) Richard Stallman, donnant aux utilisateurs la liberté et le contrôle dans l'utilisation de leurs ordinateurs et de leurs appareils informatiques, en développant de manière collaborative et en fournissant des logiciels basés sur les droits de liberté [4]. Par définition un logiciel est dit libre lorsque l'utilisation, l'étude, la modification et la duplication en vue de sa

3. `https://allgo.inria.fr/`
4. Source :`https://fr.wikipedia.org/wiki/Logiciel_libre`

diffusion sont permises, techniquement et légalement. Selon *GNU*, le terme décrit des logiciels qui *respectent la liberté des utilisateurs* faisant référence à la liberté et pas forcement à la gratuité. Les trois logiciels libres que nous utilisons sont les *CasEN*, l'*OpeNER* et le *mXS*.

CasEN

CasEN[5] est une cascade utilisant des ressources lexicales et de transducteurs, des descriptions locales de motifs, qui opèrent sur le texte en entrée afin de procéder à des insertions des remplacements ou des suppressions. *CasEN* fonctionne à l'aide du module *CasSys* sur la plateforme Unitex[6] (Paumier, 2006) afin de construire les transducteurs qui, dans l'interface d'Unitex, se présentent sous la forme de grammaires locales, représentées par des graphes. *CasEN* utilise le principe des cascades pour reconnaître des entités nommées (Friburger et Maurel, 2004, Maurel *et al.*, 2011).

La répartition des graphes de cascades est divisée en cinq catégories (Maurel *et al.*, 2011) :

1. les graphes de reconnaissance qui étiquettent une catégorie d'entités nommées ;
2. les graphes outils qui réalisent différentes reconnaissances ou transformations utilisées par la suite ;
3. les graphes listes qui sont des sous-graphes contentant des listes de mots polylexicaux ;
4. les graphes masques, également des sous-graphes, qui décrivent des listes qui contiennent des expressions régulières ou des descriptions morphologiques, et
5. les graphes étiqueteurs qui ajoutent des informations aux éléments d'une entité nommée.

CasEN a été réalisée dans le cadre des projet ANR Variling[7], FEDER Région Centre

5. `http://tln.li.univ-tours.fr/Tln_CasEN.html`
6. `http://unitex.univ-mlv.fr/`
7. `http://tln.li.univ-tours.fr/Tln_Variling.html`

Entités nommées et nommables [8], Ortolang [9] et Istex [10].

OpeNER

L'*OpeNER* [11] [12] est un projet libre avec un code source disponible gratuitement et prêt à être utilisé, financé par la Commission Européenne dans le cadre du 7e Programme-Cadre. L'OpeNER a comme objectif de fournir un ensemble d'outils prêts à l'emploi pour réaliser des tâches de traitement du langage naturel, aisément adaptable aux besoins académiques, de recherche, des petites et moyennes entreprises. Plus précisément, l'*OpeNER* est capable de détecter et désambiguïser des entités, d'effectuer une analyse de sentiment et une détection d'opinion dans les textes. Le projet est disponible en néerlandais, anglais, allemand, français, espagnol et italien, offrant la possibilité d'ajouter des langues grâce aux composants individuels et aux lignes directrices qui sont fournies. L'outil d'analyse de la langue consiste à installer, à améliorer et à configurer facilement des composants pour la détection de la langue d'un texte, la tokenisation, la détermination de la polarité des textes (analyse de sentiments), la détection des sujets inclus dans le texte et la détection d'entités nommées dans les textes.

L'outil reconnaît une variété de types d'entités nommées comme les noms de personne, les lieux, les organisation etc . *OpeNER* envisagera, dans l'avenir, de reconnaître et de classer les types d'entités nommées reliés au domaine touristique comme les restaurants, les hôtels et peut-être les monuments, les théâtres, etc.

mXS

Le *mXS* [13] est un système qui vise à chercher de manière exhaustive des modèles séquentiels hiérarchiques pour la reconnaissance automatique des entités nommées. Il a été implémenté

8. `http://tln.li.univ-tours.fr/Tln_Feder.html`
9. `https://www.ortolang.fr/`
10. `http://www.istex.fr/`
11. Il s'agit de l'acronyme *Open Polarity Enhanced Name Entity Recognition*
12. `http://www.opener-project.eu/`
13. `http://damien.nouvels.net/fr/mxs`

par Nouvel *et al.* (2014) pour le français, tout en étant capable de traiter des textes, comme la parole spontanée, sans information préalable sur le type de texte. L'approche du système se base sur la technique de l'apprentissage automatique divisée en deux étapes. La première phase se centre sur l'extraction des règles d'annotation et l'estimation des paramètres d'un modèle numérique à partir de ces règles issues de données annotées et la deuxième étape réalise une annotation à partir de ces paramètres sur la base d'une prédiction. Les textes en entrées sont prétraités au niveau de la tokenisation, de la lemmatisation et de l'étiquetage morphosyntaxique grâce à *TreeTagger*[14] et des ressources linguistiques, indispensables pour cette opération. Des dictionnaires électroniques et des transducteurs contribuent aux annotations des entités nommées, qui sont présentes lors du paramétrage. L'enrichissement des données est réalisé à l'aide de connaissances (linguistiques, lexicales).

7.3 Évaluation

7.3.1 Mesures de performance

Galibert *et al.* (2010), dans le cadre des campagnes d'évaluation concernant les entités nommées du programme Quæro, définissent les entités nommées étendues dans la perspective d'une constitution de base de connaissances à partir de textes. Par la suite, ils présentent parmi les autres mesures d'évaluation (*Rappel*, *Précision*, *F-mesure*) la mesure d'évaluation *Slot Error Rate* (SER) mise en œuvre pour évaluer les sorties des systèmes dans l'évaluation Quæro 2010. Dans le même concept de l'évaluation d'entités nommées annotées automatiquement, Galibert *et al.* (2011) définissent une méthodologie d'évaluation basée sur les métriques traditionnelles (Rappel, Précision, F-mesure) utilisées pour les entités nommées (Van Rijsbergen, 1979) et la métrique SER avec une énumération d'erreurs.

14. Outil pour l'annotation du texte avec des informations morphosyntaxiques et lemmatiques développé par Helmut Schmid : `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

Rappel : mesure la quantité des informations trouvées. Il se calcule par le nombre de documents correctement retrouvés au regard de l'ensemble des documents pertinents.

$$Rappel = \frac{Vrais\ Positifs}{Vrais\ Positifs + Faux\ Ngatifs} \tag{7.1}$$

Précision : mesure la qualité des informations trouvées. Elle se calcule par le nombre de documents correctement retrouvés parmi tous les documents retrouvés par un système.

$$Prcision = \frac{Vrais\ Positifs}{Vrais\ Positifs + Faux\ Positifs} \tag{7.2}$$

F-mesure : moyenne harmonique, combinaison pondérée du rappel et de la précision. La valeur du coefficient β permet d'une part d'équilibrer le Rappel et la Précision (β = 1), et d'une autre part d'accorder plus d'importance à l'une des deux mesures : le *Rappel* (β > 1) ou la *Précision* (β < 1).

$$F - mesure = \frac{(1 + b^2)\ *\ Prsision\ *\ Rappel}{b^2\ *\ Prsision\ +\ Rappel} \tag{7.3}$$

7.3.2 Évaluation et analyse des résultats

Étant donné les particularités de chaque système de reconnaissance, notamment en sortie, le développement d'un outil d'évaluation a été conçu. L'évaluation pour chaque système se réalise automatiquement au moyen d'une série de scripts adaptés à la sortie de chaque système, qui

préparent les corpus dans les différentes versions que nous avons mentionnées (partie 7.2.1), lancent les systèmes et calculent automatiquement les métriques. Ainsi, pour chaque version du corpus est calculé la performance de chaque système pour le Rappel, Précision et F-score. Pour l'évaluation nous calculons les scores sur chaque système et version de corpus pour les entités nommées *<pers>...</pers>*, *<loc>...</loc>* et pour l'ensemble des entités reconnues.

En tant qu'erreur dans les annotations nous considérons les entités qui sont partiellement détectées, par exemple, *Pierre Danton* au lieu de *Jean Pierre Danton*. De même, les entités nommées détectées incorrectement, par exemple, l'entité toponymique *Allènes* reconnue en tant que personne. Mais aussi, des mots incorrectement reconnus en tant qu'entités nommées, par exemple, le mot *bon* en tant que personne.

Le graphique 7.2 nous donne un premier aperçu de la performance globale (entités reconnues et fautes) pour les quatre systèmes par rapport à deux types d'entités reconnues sur les trois corpus, sans ses variations. L'analyse détaillée des résultats pour chaque système et corpus est présentée dans la partie 7.3.2.1.

Dans un premier temps, nous remarquons que *CasEN* reconnaît des entités nommées contenant un grand nombre d'erreurs, à l'exception du corpus normalisé, fait qui nous laisse penser à un faible taux de précision. Un deuxième constat est que *mSX* semble présenter les meilleures performances en relation avec le nombre des entités reconnues et les erreurs produites en comparaison avec les deux autres systèmes. Nous signalons, également, le fait que pour les systèmes *mSX* et *CasEn* le corpus normalisé obtient moins d'erreurs par rapport aux deux autres types de corpus. Finalement, seule une analyse détaillée pour chaque type d'entité nommée pourra clarifier la précision et le rappel produit par les systèmes pour chaque corpus saisi avec l'appui des métriques d'évaluations que nous avons définies.

Le système API *Nero* que nous avons décrit dans la partie 7.2.2.1 ne figure pas parmi les résultats de notre évaluation car le service en ligne n'est pas parvenu à traiter les textes du corpus brut. En effet, le corpus brut contenant un certain nombre de mots inconnus a rendu

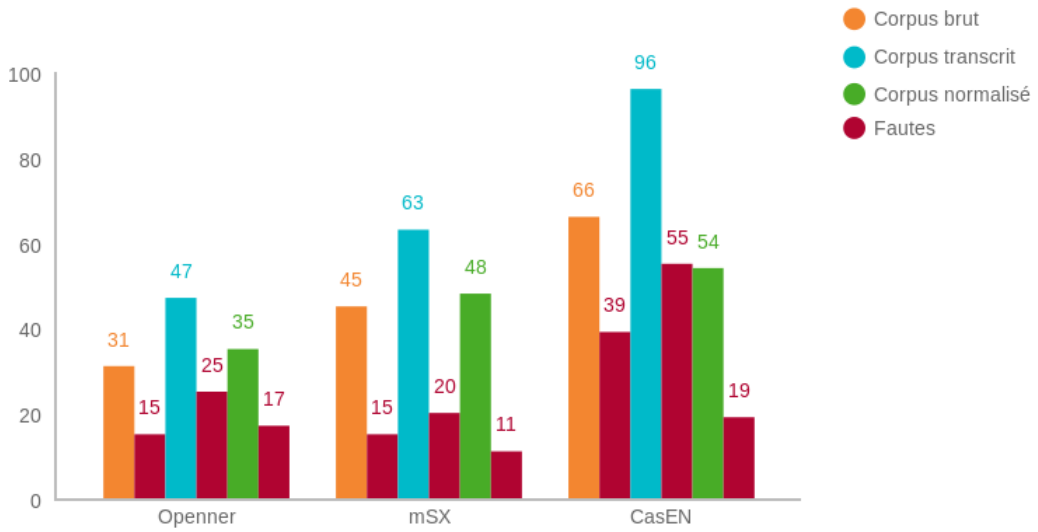

FIGURE 7.2 – Graphique générale de performance de systèmes sur l'ensemble d'entités reconnues

impossible la détection d'entités nommées, malgré le fait que le système a été implémenté dans le but de reconnaître les entités issues de textes produits par des transcriptions automatiques contentant du bruit, donc présentant des similitudes avec ceux du corpus de SMS.

Les tests effectués sur les deux autres types de corpus nous confirme que le système parvient à la détection d'entités nommées avec certaines restrictions. Plus précisément, des problèmes d'encodage ont été notés, dégradant la performance de résultats lors du traitement de la détection. Par exemple, l'entité nommée *Agnès* écrite correctement a été reconnue en tant qu'organisation et la même entité sans l'accent grave a été correctement reconnue en tant que personne (figure 7.3). De même, nous supposons que, suite au problème d'encodage et par extension à la multiplication du nombre de mots inconnus dans le corpus, la taille du corpus devrait être minimisé de 200 à environ 30 SMS pour chaque traitement du logiciel Nero (figure 7.4).

test.txt 135 Bytes

```
Agnès  tu as des dispo cette semaine pour un repas Thai chez moi ?
Agnes  tu as des dispo cette semaine pour un repas Thai chez moi ?
```

test_nero.txt 216 Bytes

```
<org> Agnès </org> tu <org> as des dispo </org> <time>cette semaine </time> pour un
<pers> Agnes </pers> tu as  <org> des dispo </org> <time>cette semaine </time> pour
```

irisa_ne.log 126 Bytes

```
Warning the word is not know par the Named Entity tagger:dispo
Warning the word is not know par the Named Entity tagger:dispo
```

allgo.log 46 Bytes

```
test.txt
  NERO process time : 0.490 seconds
```

test.json 726 Bytes

```
{
"general_info":{
"src":"test.txt",
"text":{
```

FIGURE 7.3 – Capture d'écran de message d'erreur du service *Nero* pour l'entité nommée *Agnès*

7.3.2.1 Analyse des résultats

Après la série d'expérimentations, nous sommes en mesure de fournir les résultats obtenus pour chaque système appliqué sur chaque type de corpus utilisé à ces fins. Les résultats issus des trois systèmes (*CasEN*, *OpeNER*, *mSX*) pour la détection d'entités nommées seront présentés dans la suite.

Le protocole d'évaluation s'appuie sur trois corpus de départ : 1) brut, 2) normalisé au-

FIGURE 7.4 – Capture d'écran de message d'erreur du service *Nero*

tomatiquement, et 3) transcrit manuellement, ainsi que sur un corpus de référence constitué par 200 SMS avec des balises délimitant des entités du type personne (34 au total) et lieu (24 au total). Afin de prendre en compte la performance des systèmes dans différentes situations, trois versions de chaque corpus de départ (*cf.* tableau 7.1) ont été considérées pour constituer les corpus de test : a) sans changements, b) en minuscules et c) en majuscules en mode titre. Cela nous a permis alors de constituer neuf corpus à évaluer : 1a, 2a, 3a, 1b, ... 3c.

Une fois les entités identifiées par chacun des trois systèmes, nous avons créé des scripts pour harmoniser le format des résultats, pour permettre leur comparaison au corpus de référence, ainsi que pour calculer la précision, le rappel et la F-mesure (score global et par type d'entité). Les résultats globaux en pourcentage sont résumés dans le tableau 7.3.

Les trois systèmes évalués sont implémentés selon des approches différentes : apprentissage automatique (OpeNER), ingénierie des grammaires (CasEN) et hybride (mSX). Cela nous permet d'avoir un large aperçu des résultats. Le corpus transcrit manuellement constitue la version en langage standard de référence, la comparaison des performances est alors effectuée entre les résultats issus du corpus brut et ceux issus du corpus normalisé (le meilleur score des deux est en caractères gras).

Nous remarquons que tous les systèmes ont une pauvre performance lorsque les messages

Corpus	Traitement	OpeNER			CasEN			mSX		
		R	P	F	R	P	F	R	P	F
Brut	$V_{miniscule}$	0	0	0	33.3	1.7	2.8	66.6	3.4	6.2
	$V_{initial\ maj.}$	1.0	6.8	0	16.1	53.4	24.4	29.6	63.7	40.2
	$V_{original}$	**51.6**	27.5	35.6	40.9	46.5	43.4	66.6	51.7	58.0
Normalisé Autom.	$V_{miniscule}$	0	0	0	20.0	1.7	2.6	66.6	3.4	6.2
	$V_{initial\ maj.}$	1.2	8.6	2.0	15.8	51.7	24.0	27.1	60.3	37.2
	$V_{original}$	51.4	**31.0**	**38.4**	**64.8**	**60.3**	**62.2**	**77.0**	**63.7**	**69.6**
Transcrit Manuel.	$V_{miniscule}$	0	0	0	16.6	1.7	2.0	66.6	3.4	6.2
	$V_{initial\ maj.}$	1.4	10.5	1.6	13.4	59.6	21.6	28.6	60.3	38.6
	$V_{original}$	46.8	38.5	42.2	42.7	71.9	53.4	68.2	74.1	70.8

TABLE 7.3 – Résultats globaux de trois systèmes

en entrée sont complètement constitués par des lettres minuscules. Dans le cas où les messages sont en majuscules en mode titre, tous les systèmes ont une faible précision et un rappel élevé. Ce constat souligne l'importance du traitement de la casse dans une tâche d'extraction d'entités nommées : tandis qu'une absence de casse empêche leur identification (faible F-mesure), sa surutilisation produit un surplus d'entités reconnues dont une grande partie ne sont pas correctes (faible précision et rappel élevé).

Le corpus normalisé automatiquement est obtenu avec une version améliorée du modèle de normalisation des SMS que nous avons développé durant la deuxième partie du travail de thèse. Le modèle de normalisation a été complété par l'ajout d'un module de normalisation de casse. L'approche utilisée est fondée sur des heuristiques (prise en compte de symboles de ponctuation, des émoticônes, etc) ainsi que sur l'utilisation des connaissances linguistiques (traitement de mots hors vocabulaire, utilisation de dictionnaires de noms propres et analyse du contexte des mots dans la phrase) représentées sous la forme de grammaires locales. La performance des 3 systèmes (calculée comme le taux de F-mesure) est toujours supérieure lorsque l'identification des entités nommées s'effectue sur des SMS normalisés en comparaison aux SMS bruts : OpeNER (35.6 vs 38.4), CasEN (43.4 vs 62.2) et mSX (58.0 vs 69.6).

Les écarts de pourcentages concernant les résultats de F-mesure obtenus sur la transcription manuelle (tableau 7.4) sont améliorés dans tous les cas où les SMS ont été normalisés. Par exemple, nous remarquons que pour mSX l'écart est seulement de 1.70% et pour CasEN, supérieur à celui de la transcription automatique. Cela indique que CasEN obtient de meilleurs résultats sur le corpus normalisé qu'avec le corpus transcrit manuellement.

Corpus	OpeNER			CasEN			mSX		
	F	Ftrans	Écart %	F	Ftrans	Écart %	F	Ftrans	Écart %
Brut	35.6	42.2	-15.64	43.4	53.4	-18.72	58.0	70.8	-18.08
Normalisé	38.4	42.2	**-9.00**	62.2	53.4	**+16.48**	69.6	70.8	**-1.70**

TABLE 7.4 – Écart des résultats de F-mesure obtenus sur la transcription manuelle

Ce paradoxe s'explique lorsque nous observons les trois versions du corpus (tableau 7.5), par exemple l'entité *aurelie* du corpus brut n'a pas été transcrite correctement par l'annotateur. Cependant, sur le corpus normalisé la première lettre est corrigée grâce à l'étape de normalisation de case (7.2.1.1), ceci a permis aux trois systèmes d'étiqueter correctement cette entité. Dans le deuxième exemple, les annotateurs ont traduit la forme anglaise *Let's* par *Allons*. Nous remarquons, également, que CasEN a attribué l'étiquette lieu pour *Allons* et personne pour *Let's*. Dans le corpus normalisé aucune modification n'a été effectuée et le système n'a ajouté aucune étiquette.

Corpus transcrit	Corpus brut	Corpus normalisé
Tu serais le dieu d'**aurelie** ;) bisous	T serais le dieu d'**aurelie** ;) bizous	tu serais le dieu d'<**pers**>**Aurelie**<**pers**> ;) bisous
<**loc**>**Allons**<**loc**> sur skype ! (CasEN)	<**pers**>**Let**<**pers**>'s skype ! (CasEN)	let's skype ! (CasEN)

TABLE 7.5 – Exemples d'annotation par les trois systèmes

7.4 Conclusion

Dans ce chapitre nous avons mis en œuvre la tâche d'extraction d'entités nommées. Nous avons confirmé notre hypothèse qui repose sur le fait que la normalisation morphosyntaxique des SMS permettrait d'améliorer les performances des méthodes traditionnelles pour l'identification d'entités nommées.

Les observations réalisées nous ont permis, d'une part, d'améliorer notre modèle hybride de normalisation grâce à la conception d'un nouveau module de normalisation de casse, et d'autre part, de confirmer l'adéquation d'un processus de normalisation morphosyntaxique des SMS afin d'augmenter les performances (calculée comme le taux de F-mesure) dans l'identification des entités nommées.

Nous concluons que l'étape de normalisation est indispensable pour effectuer la reconnaissance d'entités nommées dans des SMS et par extension des messages courts, présentant des caractéristiques similaires. Nous constatons, également, que les scores du corpus normalisé nous permettent de réaliser que les résultats sont proches au maximum que nous pouvons obtenir. Par conséquent, nous pouvons dire que le système de normalisation ne nécessite pas d'être amélioré afin de produire un meilleur résultat destiné à la reconnaissance d'entités nommées.

Conclusion et Perspectives

Notre travail de thèse est consacré à l'étude de la communication par SMS du point de vue du traitement automatique du langage naturel. Le point de départ est un constat simple : la plupart des messages courts présentent des différences significatives en comparaison avec le langage standard.

D'une part, la différence est mise en évidence par la morphologie particulière des mots : fusionnement, formes abrégées imprévisibles, suppression de caractères, manque de ponctuation, etc. D'autre part, par les règles de syntaxe et de grammaire qui ne sont pas respectées lorsque l'émetteur considère que cela ne nuirait pas à l'intelligibilité du message. Ces écarts avec le langage standard constituent des défis importants à surmonter lors de la conception d'approches pour le traitement automatique des SMS. Pour mieux comprendre les particularités de la communication par SMS, le travail de recherche a débuté par l'analyse d'un corpus de SMS dans le but de mettre en exergue les phénomènes linguistiques qui peuvent constituer une entrave à leur traitement automatique. Nous avons notamment conclu que l'étiquetage morphosyntaxique se révèle une étape fondamentale afin de pouvoir traiter des données textuelles de type SMS. En effet, comme pour le traitement du langage standard, le rôle de l'étiquetage morphosyntaxique est d'une haute importance pour envisager d'effectuer des tâches telles que la reconnaissance d'entités nommées, la traduction automatique, les systèmes de questions-réponses, etc.

À partir des constats de l'étape précédente, la deuxième partie du travail de recherche a consisté à étudier les méthodes probabilistes et celles fondées sur l'ingénierie des grammaires pour la normalisation morphosyntaxique des SMS. Notre but est d'appliquer cette normalisation pour pouvoir effectuer une analyse morphosyntaxique standard sans avoir à mettre en place un traitement spécifique. À partir de l'état de l'art et des expériences menées, nous avons pu conclure, que les méthodes probabilistes n'étaient pas adaptées pour normaliser des messages courts, ceci est reflété par un rappel haut mais une précision faible lors des évaluations. À la fois, les méthodes basées sur l'ingénierie des grammaires avaient une haute précision mais un faible rappel. La question de cette deuxième étape qui a permis d'atteindre notre but initial a été : Comment concevoir une méthode de normalisation offrant des résultats proches de ceux d'une version morphosyntaxique standard ? Nous avons répondu en proposant un modèle hybride pour la normalisation des SMS fondé sur la construction de réseaux de transition récursifs et l'utilisation des techniques issues du domaine de la traduction automatique. Les résultats montrent que l'approche proposée permet de rapprocher les messages de leur version en langue standard, ce qui permet par exemple, comme constaté dans la suite des travaux, d'améliorer les résultats d'une tâche d'identification des entités nommées dans des SMS.

Le travail de thèse s'est poursuivi par un troisième volet autour de l'extraction automatique d'information contenue dans les SMS. Plus précisément, nous nous intéressons à la tâche d'extraction d'entités nommées, notre hypothèse étant que la normalisation morphosyntaxique des SMS permettrait d'améliorer les performances des méthodes traditionnelles pour l'identification d'entités nommées. Après avoir défini une classification des entités nommées adaptées aux SMS, nous nous sommes basée sur une série d'expérimentations afin d'observer la qualité d'annotation de trois systèmes issus de l'état de l'art pour l'identification d'entités nommées en français standard : OpeNER (fondé sur l'apprentissage automatique), CasEN (fondé sur l'ingénierie des grammaires) et mSX (fondé sur l'ingénierie des grammaires et l'apprentissage automatique). Les observations réalisées nous ont permis deux choses : d'une part, d'améliorer notre modèle hybride de normalisation grâce à la conception d'un nouveau module de nor-

malisation de casse. D'autre part, de confirmer l'adéquation d'un processus de normalisation morphosyntaxique des SMS afin d'augmenter les performances (calculée comme le taux de F-mesure) dans l'identification des entités nommées.

Perspectives

Le corpus de SMS est un matériel authentique potentiellement précieux à analyser, au sein de plusieurs disciplines, afin de permettre aux chercheurs d'en tirer des conclusions. Le corpus du projet offre un certain nombre de perspectives liées à la recherche axée sur corpus. En effet, une étude variationnelle plus approfondie, basée sur les recherches de Cougnon (2015), pourrait offrir un panorama plus large sur les aspects sociolinguistiques des participants. De plus, une nouvelle collecte dans la région pourrait servir d'un élément de comparaison, par rapport à ce corpus, pour effectuer une analyse sur la variation diachronique des SMS et donner une réponse à la question : Est-ce la façon d'écrire les SMS a changé au fil des années ? Par ailleurs, la comparaison du corpus avec d'autres corpus similaires, tels que le corpus de communication médiée par ordinateur (tweets, chats, forums, mails etc.), contribuerait à apporter des réponses sur des points communs à observer (graphiques, orthographiques, syntaxiques etc.). Une autre proposition serait de collecter un corpus de conversations. Ces données permettraient, à la fois d'observer les habitudes communicationnelles de l'émetteur et du récepteur selon différentes situations de communication.

Le système hybride, pour la normalisation des SMS, nous fournit des résultats assez satisfaisants. Cependant, nous considérons que de nouveaux modes d'hybridation pourront être explorés sous forme de modules supplémentaires au système. L'utilisation des systèmes additionnels pour le traitement des emprunts, la phonétisation, le calcul de distances d'édition et la mémoire de traduction pourront aussi être ajoutés au système initial, faciliter la tâche de normalisation et augmenter encore plus les scores des résultats obtenus. Le système hybride

offre la possibilité d'être appliqué à d'autres corpus similaires (SMS, tweets, mails, forums, chats etc.) pour servir d'aide à plusieurs tâches, telle que la transcription manuelle. Son rôle de rapprochement d'un SMS à la langue standard remplit bien sa tâche pour l'extraction d'information et plus précisément celle de la reconnaissance des entités nommées. Ainsi, son application à d'autres corpus à normaliser pourrait servir comme matériel pour d'autres applications du TAL, par exemple les systèmes d'anonymisation, les chatbots, la traduction automatique, la synthèse vocale, etc. Le système a été conçu pour la normalisation de SMS en français. Cependant, il serait possible de l'adapter aux particularités de chaque langue et de l'alimenter avec des ressources linguistiques et des règles pertinentes.

Bibliographie

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P. et Vilain, M. (1995). Mitre : description of the alembic system used for muc-6. *Dans : Proceedings of the 6th conference on Message understanding*, pages 141–155. Association for Computational Linguistics.

Abney, S., Collins, M. et Singhal, A. (2000). Answer extraction. *Dans : Proceedings of the sixth conference on Applied natural language processing*, pages 296–301. Association for Computational Linguistics.

AbuSeileek, A. F. et Qatawneh, K. (2013). Effects of synchronous and asynchronous computer-mediated communication (cmc) oral conversations on english language learners' discourse functions. *Computers & Education*, 62:181–190.

Amaghlobeli, N. (2012). Linguistic Features of Typographic Emoticons in SMS Discourse. *Theory and Practice in Language Studies*, 2(2):348–354.

Anis, J. (1998). *Texte et ordinateur : l'écriture réinventée ?* De Boeck Supérieur.

Anis, J. (1999). *Internet communication et langue française.* Hermès science publications.

Anis, J. (2000). L'écrit des conversations électroniques de l'internet. *Le français aujourd'hui*, 129:59–69.

BIBLIOGRAPHIE

Anis, J. (2003). Communication électronique scripturale et formes langagières. *Actes des Quatrièmes rencontres Réseaux humains/Réseaux technologiques*, 31.

Anis, J., de Fornel, M. et Fraenkel, B. (2004). La communication électronique : Approches linguistiques et anthropologiques. *Dans : Colloque international, EHESS.*

Antoniadis, G., Chabert, G. et Zampa, V. (2011). Alpes4science : Constitution d'un corpus de sms réels en france métropolitaine. *Dans : TEXTOS conference : dimensions culturelles, linguistiques et pragmatiques.*

Auer, P. (1984). *Bilingual conversation.* Amsterdam : John Benjamins Publishing.

Aw, A., Zhang, M., Xiao, J. et Su, J. (2006). A phrase-based statistical model for SMS text normalization. *Dans : Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 33–40. 2006 Association for Computational Linguistics.

Baker, M. (1998). Réexplorer la langue de la traduction : une approche par corpus. *Meta : Journal des traducteurs/Meta : Translators' Journal*, 43(4):480–485.

Baldwin, T., Kim, Y.-B., De Marneffe, M. C., Ritter, A., Han, B. et Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text : Twitter lexical normalization and named entity recognition. *ACL-IJCNLP*, 126:2015.

Bangalore, S., Murdock, V. et Riccardi, G. (2002). Bootstrapping Bilingual Data Using Consensus Translation For A Multilingual Instant Messaging System. *Coling-2002.*

Barasa, S. et Mous, M. (2009). The oral and written interface on SMS : Technology Mediated Communication in Kenya. *Dans : Low-educated adult second language and literacy acquisition*, pages 234–242. Proceedings of the Low-Educated Second LanguageandLiteracy Acquisition (LESLLA) Symposium.

Barbieri, F., Ronzano, F. et Saggion, H. (2016). What does this emoji mean ? a vector space skip-gram model for twitter emojis. *Dans : Language Resources and Evaluation conference.*

Barbizet, J. et Lenoir, G. (1968). Étude dynamique d'échantillons verbaux chez des sujets normaux et chez des malades atteints de lésions cérébrales . Étude du « débit verbal ». *L'année psychologique*, 68, no 2:431–449.

Bauer, G. (1985). *Namenkunde des Deutschen.* Germanistische Lehrbuchsammlung. P. Lang.

Beaufort, R., Roekhaut, S., Cougnon, L.-A. et Fairon, C. (2010a). A hybrid rule/model-based finite-state framework for normalizing sms messages. *Dans : Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 770–779. Association for Computational Linguistics.

Beaufort, R., Roekhaut, S., Cougnon, L.-A. et Fairon, C. (2010b). Une approche hybride traduction/correction pour la normalisation des SMS Richard. *Dans : TALN 2010.*

Bevilacqua, S., Juan, I. et Fernández, R. (2012). La communication médiée par téléphone (CMT). *Synergies Argentine no 1 - 2012*, pages 117–126.

Borthwick, A., Sterling, J., Agichtein, E. et Grishman, R. (1998). Exploiting diverse knowledge sources via maximum entropy in named entity recognition. *Dans : Proc. of the Sixth Workshop on Very Large Corpora*, volume 182.

Bourdieu, P., Chamboredon, J., Passeron, J. et Krais, B. (2005). *Le métier de sociologue : préalables épistémologiques.* Textes de Sciences Sociales Series. Mouton de Gruyter.

Bowker, L. et Pearson, J. (2002). *Working with specialized language : a practical guide to using corpora.* Routledge.

Brasart, C. (2013). Corpus et alternance codique : que peut nous apprendre une approche comparative ? *Corela. Cognition, représentation, langage*, (HS-13).

Bréal, M. (1897). *Essai de sémantique : Science des Significations.* Hachette.

Breckx, M. (1996). *Grammaire française.* De Boeck.

Brody, S. et Diakopoulos, N. (2011). Coooooooooooooollllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs. *Dans : Proceedings Empirical Methods in Natural Language Processing (EMNLP)*, pages 562–570.

Bulot, T. et Blanchet, P. (2011). Dynamiques de la langue française au xxième siècle : une introduction à la sociolinguistique.

Bunescu, R. et Mooney, R. J. (2004). Relational markov networks for collective information extraction. *Dans : ICML-2004 Workshop on Statistical Relational Learning.*

Campedel, M. et Hoogstoël, P. (2011). *Sémantique et multimodalité en analyse de l'information.* Collection Recherche d'Information et Web. Hermès science.

Carroll, J. B. (1964). Language and thought. *Reading Improvement*, 2(1):80.

Caseli, H. et Nunes, M. (2007). Automatic induction of bilingual lexicons for machine translation. *Int J Transl*, 19:29–43.

Caseli, H. M., Nunes, M. D. G. V. et Forcada, M. L. (2006). Automatic induction of bilingual resources from aligned parallel corpora : Application to shallow-transfer machine translation. *Machine Translation*, 20(4):227–245.

Chabert, G., Zampa, V., de l'oralité alpine, C., Laboratoire de linguistique et didactique des langues étrangères et maternelles (Saint-Martin-d'Hères, I. et Centre de traitement automatique du langage (Louvain-la Neuve, B. (2012). *Des SMS alpins.* Éd. de la Bibliothèque départementale des Hautes-Alpes.

Chinchor, N., Robinson, P. et Brown, E. (1998). Hub-4 named entity task definition version 4.8.

Chinchor, N. A. (1998). Overview of muc-7/met-2. Rapport technique, Science Applications International Corp San Diego CA.

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S. et Basu, A. (2007). Investigation and modeling of the structure of texting language. *International Journal of Document Analysis and Recognition (IJDAR)*, 10(3-4):157–174.

Coates-Stephens, S. (1992). The analysis and acquisition of proper names for the understanding of free text. *Computers and the Humanities*, 26(5):441–456.

Collins, M. et Singer, Y. (1999). Unsupervised models for named entity classification. *Dans : Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, pages 100–110.

Collot, M. et Belmore, N. (1996). A new variety of english. *Computer-mediated communication : Linguistic, social, and cross-cultural perspectives*, 39:13.

Condamines, A. (2005). Linguistique de corpus et terminologie. *Langages*, 39(157):36–47.

Condamines, A. (2006). Modes de construction du sens en corpus spécialisé. *Cahiers de grammaire*, 30:75–88.

Cougnon, L.-A. (2010). Orthographe et langue dans les sms. *Ela. Études de linguistique appliquée*, (4):397–410.

Cougnon, L.-A. (2015). *Langage et sms : Une étude internationale des pratiques actuelles.* Louvain-la-Neuve : UCL Presses universitaires de Louvain, DL 2015, cop. 2015, presses un édition.

Cougnon, L.-A. et Beaufort, R. (2009). *SSLD : a French SMS to standard language dictionary.* S. Granger & M. Paquot (Eds.) eLexicography in the 21st century : New applications, new challenges. Proceedings of eLEX2009. Cahiers du Cental 7. Louvain-la-Neuve : Presses universitaires de Louvain.

Cougnon, L.-A. et François, T. (2011). Etudier l'écrit SMS - Un objectif du projet sms4science. *Dans : Linguistik online 48.*

BIBLIOGRAPHIE

Cougnon, L.-A. et Ledegen, G. (2008). c'est écrire comme je parle. une étude comparatiste de variétés de français dans l'écrit sms. *Actes du Congrès annuel de l'AFLS*.

Cougnon, L.-A., Roekhaut, S. et Beaufort, R. (2013). Typologies de variation graphique dans l'écrit sms. *L'orthographe en quatre temps*, pages 129–148.

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue française*, (87):11–22.

Covington, M. A. et McFall, J. D. (2010). Cutting the gordian knot : The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

Crucianu, M., Cubaud, P., Raphaël, F.-S., Marin, F. et Cnam (1999). Ingénierie de la fouille et de la visualisation de données massives (rcp216) — cours cnam rcp216.

Cruse, D. (2004). *Meaning in Language : An Introduction to Semantics and Pragmatics*. Oxford linguistics. Oxford University Press.

Daille, B., Fourour, N. et Morin, E. (2000). Catégorisation des noms propres : une étude en corpus. *Cahiers de Grammaire*, 25(25):115–129.

Danesi, M. (2008). *Dictionary of Media and Communications*. M. E. Sharpe Incorporated.

David, C. (2001). Language and the internet. *Cambridge, CUP*.

De Singly, F. (2012). *Le questionnaire : L'enquête et ses méthodes*. Sociologie. Armand Colin.

De Singly, F. (2016). *Le questionnaire : L'enquête et ses méthodes*. Sociologie. Armand Colin.

December, J. (1996). What is computer-mediated communication ?

Dejond, A. (2006). *Cyberlangage*. Lannoo Uitgeverij.

Désoyer, A., Landragin, F. et Tellier, I. (2015). Apprentissage automatique d'un modèle de résolution de la coréférence à partir de données orales transcrites du français : le système croc. *Dans : Vingt-deuxième Conférence sur le Traitement Automatique des Langues Naturelles*, pages 439–445.

Dessus, P., Lemaire, B. et Baillé, J. (1997). Etudes expérimentales sur l'enseignement à distance. *Sciences et techniques éducatives*, 4(2):137–164.

Di Sciullo, A.-M., Muysken, P. et Singh, R. (1986). Government and code-mixing. *Journal of linguistics*, 22(1):1–24.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Dans : Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. et Weischedel, R. M. (2000). Entity Detection and Tracking - Phase 1 ACE Pilot Study Task Definition. Rapport technique, ACE.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. et Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. *Dans : Language Resources and Evaluation Conference*, volume 2, pages 837–840.

Doehler, S. P. (2013). Hallo ! voulez vous luncher avec moi hüt ? le" code switching" dans la communication par sms. *Linguistik online*, 48(4).

Downes, E. J. et McMillan, S. J. (2000). Defining interactivity : A qualitative identification of key dimensions. *New media & society*, 2(2):157–179.

Dresner, E. et Herring, S. C. (2010). Functions of the nonverbal in cmc : Emoticons and illocutionary force. *Communication theory*, 20(3):249–268.

Ehrmann, M. (2008). *Les entités nommées , de la linguistique au TAL : Statut théorique et méthodes de désambiguisation*. Thèse de doctorat.

Ek, T., Kirkegaard, C., Jonsson, H. et Nugues, P. (2011). Named entity recognition for short text messages. *Procedia-Social and Behavioral Sciences*, 27:178–187.

Enjalbert, P. (2005). *L'extraction d'information*, pages 309–334. Traité IC2, série Cognition et traitement de l'information. Hermès Sciences, Lavoisier.

Espla-Gomis, M., Sánchez-Cartagena, V. M. et Pérez-Ortiz, J. A. (2011). Enlarging monolingual dictionaries for machine translation with active learning and non-expert users. *Dans : RANLP*, pages 339–346.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S. et Yates, A. (2005). Unsupervised named-entity extraction from the web : An experimental study. *Artificial intelligence*, 165(1):91–134.

Fairon, C., Klein, J.-R. et Paumier, S. (2006). *Le langage SMS : étude d'un corpus informatisé à partir de l'enquête "Faites don de vos SMS à la science"*. Cahiers du Cental (Louvain-la-Neuve), ISSN 1783-2845. UCL Presses universitaires de Louvain, DL 2006.

Fergadiotis, G., Wright, H. H. et Green, S. B. (2015). Psychometric evaluation of lexical diversity indices : assessing length effects. *Journal of Speech, Language, and Hearing Research*, 58(3):840–852.

Fergadiotis, G., Wright, H. H. et West, T. M. (2013). Measuring Lexical Diversity in Narrative Discourse of People With Aphasia. *American journal of speech-language pathology / American Speech-Language-Hearing Association*, 22(2).

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Monceaux, L., Robba, I. et Vilnat, A. (2001). Finding an answer based on the recognition of the question focus. *Dans : TREC*.

Finkel, J. R., Grenager, T. et Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *Dans : Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Forcada, M., Bonev, B., Rojas, S., Ortiz, J., Sánchez, G., Martínez, F., Armentano-Oller, C., Montava, M. et Tyers, F. (2008). Documentation of the open-source shallow-transfer machine translation platform Apertium.

Forcada, M. L., Tyers, F. M. et Ramírez-Sánchez, G. (2009). The Apertium machine translation platform : Five years on. *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 3–10.

Fourour, N. (2001). Identification et catégorisation automatiques des anthroponymes du Français. *Récital*, pages 2–5.

Fourour, N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. *Dans : Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'02)*, pages 265–274.

Fourour, N. (2004). *Identification et catégorisation automatique des entités nommées dans les textes français*. Thèse de doctorat.

Fraisse, A., Paroubek, P. et Francopoulo, G. (2013). L'apport des entités nommées pour la classification des opinions minoritaires. *TALN-RÉCITAL 2013*, page 588.

Frehner, C. (2008). *Email, SMS, MMS : The linguistic creativity of asynchronous discourse in the new media age*, volume 58. Peter Lang.

Freitas, C., Mota, C., Santos, D., Oliveira, H. G. et Carvalho, P. (2010). Second harem : Advancing the state of the art of named entity recognition in portuguese. *Dans : Language Resources and Evaluation Conference*.

Friburger, N. (2002). *Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques.* Thèse de doctorat.

Friburger, N. et Maurel, D. (2004). Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313(1):93–104.

Gadde, P., Subramaniam, L. et Tanveer A., F. (2011). Adapting a WSJ trained Part-of-Speech tagger to Noisy Text : Preliminary Results. *Dans : 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data (MOCR/AND '11)*. ACM.

Gadet, F. (1996). Une distinction bien fragile : écrit/oral. *Revue Tranel (Travaux neuchâtelois de linguistique)*, 25:13–27.

Galibert, O., Quintard, L., Rosset, S., Zweigenbaum, P., Nédellec, C., Aubin, S., Gillard, L., Raysz, J.-P., Pois, D., Tannier, X. *et al.* (2010). Named and specific entity detection in varied data : The quæro named entity baseline evaluation. *Dans : Language Resources and Evaluation conference.*

Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P. et Quintard, L. (2011). Structured and extended named entity evaluation in automatic speech transcriptions. *Dans : IJCNLP*, pages 518–526.

Gary-Prieur, M.-N. (1991). Le nom propre constitue-t-il une catégorie linguistique ? *Langue française*, (92):4–25.

Gaudin, F. (2000). Kleiber, Georges (1999) : Problèmes de sémantique. la polysémie en questions, villeneuve d'ascq, presses universitaires du septentrion, coll. ğ sens et structures ğ, 220 p. *Meta : Journal des traducteurs/Meta : Translators' Journal*, 45(2):370–376.

Gayraud, F. (2001). *Le développement de la différenciation oral : écrit vu à travers le lexique.* Thèse de doctorat, Université Lumière Lyon 2.

Goosse, A. (1986). *Le bon usage : grammaire Française.* Duculot.

GRASS, T. (2000). Typologie et traductibilité des noms propres de l'allemand vers le français. *TAL. Traitement automatique des langues*, 41(3):643–669.

GRASS, T. et MAUREL, D. (2008). Les noms propres d'association et d'organisation : traduction et traitement automatique. *Les Nouveaux Cahiers d'Allemand*, (2):161–174.

GROSS, M. (1993). Local grammars and their representation by finite automata. Hoey M. *Dans : Data, Description, Discourse. Papers on the English Language in honour of John McH Sinclair*. Harper-Collins, 26–38.

GUMPERZ, J. J. (1982). *Discourse strategies*, volume 1. Cambridge University Press.

HALTÉ, P. (2013). *Les marques modales dans les chats : étude sémiotique et pragmatique des émoticônes et des interjections dans un corpus de conversation synchrones en ligne*. Thèse de doctorat, Université de Lorraine.

HAN, B. et BALDWIN, T. (2011). Lexical Normalisation of Short Text Messages : Makn Sens a # twitter. *Dans : Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 368–378. Association for Computational Linguistics.

HANK, T. et HELSLOOT, N. (1995). *Michel Pêcheux : Automatic Discourse Analysis*. Rodopi, utrecht st édition.

HANKS, P. (2003). *Dictionary of American family names*. Oxford University Press.

HÅRD SEGERSTAD, Y. (2005). Language in sms—a socio-linguistic view. *Dans : The inside text*, pages 33–51. Springer.

HATMI, M. (2014). *Reconnaissance des entités nommées dans des documents multimodaux*. Thèse de doctorat.

HEAPS, H. S. (1978). *Information retrieval : Computational and theoretical aspects*. Academic Press, Inc.

Hebert, D., Paquet, T. et Nicolas, S. (2012). Champs aléatoires conditionnels et fonctions de caractéristique à quantification multi-échelle application à l'extraction de structures dans des journaux d'archive. *Dans : RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle)*, pages 978–2.

Heiden, S., Magué, J.-P. et Pincemin, B. (2010). Txm : Une plateforme logicielle open-source pour la textométrie-conception et développement. *Dans : 10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*, pages 1021–1032. Edizioni Universitarie di Lettere Economia Diritto.

Helsloot, N. et Hak, T. (2000). La contribution de Michel Pêcheux à l'analyse de discours. *Langage et société*, No 91:5–33.

Herring, S. C. (1996). *Computer-mediated communication : Linguistic, social, and cross-cultural perspectives*, volume 39. Amsterdam : John Benjamins.

Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M. et Tyson, M. (1997). Fastus : A cascaded finite-state transducer for extracting information from natural-language text. *Finite-state language processing*, pages 383–406.

Hugo, V. (2013). *A Propos de William Shakespeare*. Library of Alexandria. Library of Alexandria.

Jansche, M. et Abney, S. P. (2002). Information extraction from voicemail transcripts. *Dans : Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 320–327. Association for Computational Linguistics.

Jose, G. et Raj, N. S. (2014). Lexical normalization model for noisy sms text. *Dans : Computational Systems and Communications (ICCSC), 2014 First International Conference on*, pages 57–62.

Kalman, Y. M. et Gergle, D. (2014). Letter repetitions in computer-mediated communication : A unique link between spoken and online language. *Computers in Human Behavior*, 34:187–193.

Kaufmann, M. et Kalita, J. (2010a). Syntactic normalization of Twitter messages. *Dans : International conference on natural language processing, Kharagpur, India.*

Kaufmann, M. et Kalita, J. (2010b). Syntactic normalization of twitter messages. *Dans : International conference on natural language processing, Kharagpur, India.*

Kevers, L. (2011). *Accès sémantique aux bases de données documentaires. Techniques symboliques de traitement automatique du langage pour l'indexation thématique et l'extraction d'informations temporelles.* Thèse de doctorat.

Kiousis, S. (2002). Interactivity : a concept explication. *New media & society*, 4(3):355–383.

Kobus, C., Yvon, F. et Damnati, G. (2008). Normalizing SMS : are two metaphors better than one ? *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 441–448.

Kress, G. et Van Leeuwen, T. (2001). *Multimodal Discourse : The Modes and Media of Contemporary Communication.* An Arnold Publication Series. Arnold.

Krupka, G. et Hausman, K. (2005). Description of netowl extractor system as used in muc-7. *Dans : Proc. 7th Message Understanding Conf.*

Landau, M., Sillion, F. et Vichot, F. (1993). Exoseme : A thematic document filtering system. *Intelligence Artificielle*, 93.

Laursen, D. (2005). Please reply ! the replying norm in adolescent sms communication. *Dans : The inside text*, pages 53–73. Springer.

Laviosa, S. (2002). *Corpus-based translation studies : theory, findings, applications*, volume 17. Rodopi.

Leimdorfer, F. et Salem, A. (1995). Usages de la lexicométrie en analyse de discours. *Cahiers des sciences humaines*, 31(7):131–143.

Leroy, S. (2001). *Entre identification et catégorisation, l'antonomase du nom propre en français*. Thèse de doctorat.

Lin, D. *et al.* (1998). An information-theoretic definition of similarity. *Dans : Icml*, volume 98, pages 296–304.

Ling, R. (2005). The sociolinguistics of sms : An analysis of sms use by a random sample of norwegians. *Dans : Mobile communications*, pages 335–349. Springer.

Liu, X., Zhang, S., Wei, F. et Zhou, M. (2011). Recognizing named entities in tweets. *Dans : Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 359–367. Association for Computational Linguistics.

Liu, X., Zhou, M., Wei, F., Fu, Z. et Zhou, X. (2012). Joint inference of named entity recognition and normalization for tweets. *Dans : Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, pages 526–535. Association for Computational Linguistics.

Lopez, C., Bestandji, R., Roche, M. et Panckhurst, R. (2014). Towards electronic sms dictionary construction : An alignment-based approach. *Dans : LREC : Language Resources and Evaluation Conference*, pages 2833–2838.

Lopez, C., Roche, M. et Panckhurst, R. (2015). Classification des items inconnus de 88milSMS : aide à l'identification automatique de la créativité scripturale. *TRANEL. Travaux Neuchâtelois de Linguistique*, 63:71–86.

Lopez, C., Roche, M. et Panckhurst, R. (2016). Non-standard texts : from theoretical positions to Natural Language Processing normalisation'. PLIN-Day. Poster.

MAEDCHE, A. et STAAB, S. (2002). Measuring similarity between ontologies. *Dans : International Conference on Knowledge Engineering and Knowledge Management*, pages 251–263. Springer.

MAHMOUD, A. et AUTER, P. J. (2009). The interactive nature of computer-mediated communication. *American Communication Journal*, 11(4):1–36.

MAINGUENEAU, D. (2012). *Analyser les textes de communication.* Armand Colin.

MANGENOT, F. (2009). Du minitel aux sms, la communication éléctronique et ses usages pedagogiques. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (60):97–110.

MARCOCCIA, M. (2000a). La communication écrite médiatisée par ordinateur : faire du face à face avec de l ' écrit. *Journée d'étude de l'ATALA "Le traitement automatique des nouvelles formes de communication écrite (e-mails, forums, chats, SMS, etc.)"*, pages 1–4.

MARCOCCIA, M. (2000b). Les smileys : une représentation iconique des émotions dans la communication médiatisée par ordinateur. *Les émotions dans les interactions communicatives*, pages 249–263.

MARCOCCIA, M. (2016). *Analyser la communication numérique écrite.* I.COM. Armand Colin.

MARKETT, C., SÁNCHEZ, I. A., WEBER, S. et TANGNEY, B. (2006). Using short message service to encourage interactivity in the classroom. *Computers & Education*, 46(3):280–293.

MARTINEAU, C., TOLONE, E. et VOYATZI, S. (2007). Les entités nommées : usage et degrés de précision et de désambiguïsation. *Dans : 26ème Colloque international sur le Lexique et la Grammaire (LGC'07)*, pages pages–105.

MARTINON, P. (1927). Comment on parle en français, larousse, paris. *MartinonComment on parle en français1927.*

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol, I. et Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues*, 52(1):69–96.

McCallum, A. et Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Dans : Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

McDonald, D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *Corpus processing for lexical acquisition*, pages 21–39.

McEnery, A. M. et Wilson, A. (2001). *Corpus linguistics : an introduction.* Edinburgh University Press.

McMillan, S. J. (2002). Exploring models of interactivity from multiple research traditions : Users, documents, and systems. *Handbook of new media*, 2:205–229.

McMillan, S. J. et Hwang, J.-S. (2002). Measures of perceived interactivity : An exploration of the role of direction of communication, user control, and time in shaping perceptions of interactivity. *Journal of advertising*, 31(3):29–42.

McNamee, P., Mayfield, J. C. et Piatko, C. D. (2011). Processing named entities in text. *Johns Hopkins APL Technical Digest*, 30(1):31–40.

Mellet, S. (2002). Corpus et recherches linguistiques. *Corpus*, 1(1):1–6.

Meur, C., Galliano, S. et Geoffrois, E. (2004). Conventions d'annotations en entités nommées-ester. *Rapport technique de la campagne Ester.*

Mikheev, A., Grover, C. et Moens, M. (1998). Description of the ltg system used for muc-7. *Dans : Proceedings of 7th Message Understanding Conference (MUC-7)*, pages 1–12. Fairfax, VA.

Miller, D., Schwartz, R., Weischedel, R. et Stone, R. (1999). Named entity extraction from broadcast news. *Dans : Proceedings of the DARPA Broadcast News Workshop*, pages 37–40.

Minkov, E., Wang, R. C. et Cohen, W. W. (2005). Extracting personal names from email : Applying named entity recognition to informal text. *Dans : Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 443–450. Association for Computational Linguistics.

Monceaux, A. (1995). Le dictionnaire des mots simples anglais : mots nouveaux et variantes orthographiques. *Sèrie Informes IGM*, pages 95–15.

Mondada, L. (2000). Les effets théoriques des pratiques de transcription. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (42):131–146.

Moore, D. L. et Tarnai, J. (2002). Evaluating nonresponse error in mail surveys.

Morel, E. (2016). *Le bricolage plurilingue dans la communication par texto : interprétations d'une pratique entre affiliation locale et aspiration globale.* Thèse de doctorat, Université de Neuchâtel, Neuchătel, Switzerland.

Morel, E. et Doehler, S. P. (2013). Les 'textos' plurilingues : l'alternance codique comme ressource d'affiliation à une communauté globalisée. *Revue française de linguistique appliquée*, 18(2):29–43.

Muller, C. (1969). La statistique lexicale. *Langue française*, 2(LE LEXIQUE):30–43.

Muller, C. (1977). *Principes et méthodes de statistique lexicale.* Classiques Hachette.

Muysken, P. (2000). *Bilingual speech : A typology of code-mixing*, volume 11. Cambridge University Press.

Nadeau, D. et Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nouvel, D. (2012). *Reconnaissance des entités nommées par exploration de règles d'annotation-Interpréter les marqueurs d'annotation comme instructions de structuration locale*. Thèse de doctorat.

Nouvel, D., Antoine, J.-Y. et Friburger, N. (2014). Pattern mining for named entity recognition. *LNCS/LNAI Series*, 8387i (post-proceedings LTC 2011).

Och, F. J. et Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Paik, W., Liddy, E. D., Yu, E. et McKenna, M. (1996). Categorizing and standardizing proper nouns for efficient information retrieval. *Corpus processing for lexical acquisition*, pages 61–73.

Pak, A. (2012). *Automatic, adaptive, and applicative sentiment analysis*. Thèse de doctorat.

Pal, A. R. et Saha, D. (2015). Word sense disambiguation : A survey. *arXiv preprint arXiv :1508.01346*.

Palmer, D. D. (2000). Tokenisation and Sentence Segmentation. *Handbook of Natural Language Processing*, pages 11–35.

Panckhurst, R. (1997). La communication médiatisée par ordinateur ou la communication médiée par ordinateur ? *Terminologies nouvelles*, (17):56–58.

Panckhurst, R. (1998). Marques typiques et ratages en communication médiée par ordinateur. *Dans : Actes du colloque CIDE 98, INPT, Rabat*. Europia Productions, 31–43.

Panckhurst, R. (1999). La Communication médiée par ordinateur : un discours autre ? *Dans : L'autre en discours*, Bres, J., Delamotte-Legrand, R., Madray, F., Siblot, P., Dyalang-Praxiling. Montpellier : Service des publications de l'Université Paul-Valéry Montpellier 3, 307-331.

Panckhurst, R. (2006). Le discours électronique médié : bilan et perspectives. *Lire, Écrire, Communiquer et Apprendre avec Internet*, pages 345–366.

Panckhurst, R. (2007). Discours électronique médié : quelle évolution depuis une décennie ? *Dans : La langue du cyberespace : de la diversité aux normes, l'Harmattan*, Gerbault Jeannine, pages p. 121–136, 2007.

Panckhurst, R. (2009). Short message service (sms) : typologie et problématiques futures. *Arnavielle T. (coord.), Polyphonies, pour Michelle Lanvin*, pages 33–52.

Panckhurst, R. (2017). Discours numérique médié (DNM) : le cas des SMS. Conférence invitée, Université Grenoble Alpes.

Panckhurst, R., Détrie, C., Lopez, C., Claudine, M. et Bertrand, V. (2013). Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS. *Epistémè, Cambridge University Press (CUP), Communication électronique et écritures numériques*, pages 107–138.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M. et Verine, B. (2014a). 88milSMS. A corpus of authentic text messages in french. Produit par l'Université Paul-Valéry Montpellier III et le CNRS, en collaboration avec l'Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M. et Verine, B. (2014b). Transcodage d'un échantillon de 1000 SMS, extraits du corpus « 88milSMS ».

Panckhurst, R. et Moïse, C. (2012). French text messages : From sms data collection to preliminary analysis. *Lingvisticae Investigationes*, 35(2):289–317.

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *Dans : Proceedings of the 40th annual meeting*

on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

Paroubek, P., Pak, A. et Mostefa, D. (2010). Annotations for opinion mining evaluation in the industrial context of the doxa project. *Dans : Language Resources and Evaluation Conference*.

Patel, N., Accorsi, P., Inkpen, D., Lopez, C. et Roche, M. (2013). Approaches of anonymisation of an sms corpus. *Dans : International Conference on Intelligent Text Processing and Computational Linguistics*, pages 77–88. Springer.

Paumier, S. (2003). *De la reconnaissance des formes linguistiques à l'analyse syntaxique*. Thèse de doctorat, Université de Marne-la-Vallée. Thèse de doctorat dirigée par Gross, Maurice et Laporte, Eric Informatique linguistique Université de Marne-la-Vallée 2003.

Paumier, S. (2006). Unitex 1.2 - Manuel d'utilisation.

Pecheux, M. (1969). *Analyse automatique du discours*. Paris, Dunod.

Péraya, D. (1994). Formation à distance et communication médiatisée. *Recherches en communication*, 1(1):147–167.

Pfaff, C. W. (1979). Constraints on language mixing : intrasentential code-switching and borrowing in spanish/english. *Language*, pages 291–318.

Poibeau, T. (1999). Evaluation des systèmes d'extraction d'information : Une expérience sur le français. *Langues*, 2(2):110–118.

Poibeau, T. (2005). Sur le statut référentiel des entités nommées. *Dans : Conférence Traitement Automatique des Langues 2005*, pages 173–183. Association pour le Traitement Automatique des LAngues/LIMSI.

Polifroni, J., Kiss, I. et Adler, M. (2010). Bootstrapping named entity extraction for the creation of mobile services. *Dans : Language Resources and Evaluation Conference*.

Poplack, S. (2004). *Code-switching. Sociolinguistic. An International Handbook of the Science of Language*. Berlin : Walter de Gruyter.

Prochasson, E., Morin, E. et Viard-Gaudin, C. (2009). Vers la reconnaissance de mini-messages manuscrits. *arXiv preprint arXiv :0909.3028*.

Raghunathan, K. et Krawczyk, S. (2009). CS224N : Investigating SMS Text Normalization using Statistical Machine Translation.

Ramirez-Sánchez, G., Sánchez-Martinez, F., Ortiz-Rojas, S., Pérez-Ortiz, J. A. et Forcada, M. L. (2006). Opentrad Apertium open-source machine translation system : an opportunity for business and research. *Proceedings of the Twenty-Eighth International Conference on Translating and the Computer*.

Ratinov, L. et Roth, D. (2009). Design challenges and misconceptions in named entity recognition. *Dans : Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Rau, L. F. (1991). Extracting company names from text. *Dans : Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on*, volume 1, pages 29–32. IEEE.

Raymond, C. et Fayolle, J. (2010). Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. *Dans : Conférence Traitement automatique des langues naturelles, TALN'10*.

Richards, B. (1987). Type/Token Ratios : What do they really tell us ? *Journal of Child Language*, 14(May):201–209.

Ritter, A., Clark, S., Etzioni, O. *et al.* (2011). Named entity recognition in tweets : an experimental study. *Dans : Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Romiszowski, A. et Mason, R. (1996). Computer-mediated communication. *Handbook of research for educational communications and technology*, 2:397–431.

Santos, D., Seco, N., Cardoso, N. et Vilela, R. (2006). Harem : An advanced ner evaluation contest for portuguese. *Dans : quot ; In Nicoletta Calzolari ; Khalid Choukri ; Aldo Gangemi ; Bente Maegaard ; Joseph Mariani ; Jan Odjik ; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006).*

Satoshi, S. et Hitoshi, I. (2000). Irex : Ir and ie evaluation project in japanese. *Dans : Proceedings of the 2nd International Conference on Language Resources & Evaluation.*

Schmid, H. (1994). Treetagger. *TC project at the Institute for Computational Linguistics of the University of Stuttgart.*

Schmid, H. (2007). *Tokenizing.* Mouton de Gruyter, Berlin., corpus lin édition.

Schultz, T. (1999). Interactive options in online journalism : A content analysis of 100 us newspapers. *Journal of Computer-Mediated Communication*, 5(1):JCMC513.

Scott, M. (1996). Wordsmith tools.

Sekine, S. et Isahara, H. (1999). Irex project overview. *Dans : Proceedings of the IREX Workshop*, pages 7–12.

Sekine, S., Sudo, K. et Nobata, C. (2002). Extended named entity hierarchy. *Dans : The Third International Conference on Language Resources and Evaluation (LREC). Iles Canaries, Espagne.*

Senellart, J. (1998). Reconnaissance automatique des entrées du lexique-grammaire des phrases figées. *Travaux de linguistique*, (37):189–190.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(July 1928):379–423.

Sidarenka, U., Scheffler, T. et Stede, M. (2013). Rule-based normalization of german twitter messages. *Dans : Proccedings of the GSCL Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation.*

Sinclair, J. (1994). Corpus typology. *EAGLES DOCUMENT EAG–CSG/IR–T1*, 1.

Singer, E., Van Hoewyk, J. et Maher, M. P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64(2):171–188.

Smith, G. (2008). Does gender influence online survey participation ? : A record-linkage analysis of university faculty online survey response behavior. *ERIC Document Reproduction Service No. ED 501717.*

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M. et Richards, C. (2001). Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.

Stark, E. (2011). La morphosyntaxe dans les SMS suisses francophones : Le marquage de l'accord sujet – verbe conjugué.

Stark, E. (2015). 'de l'oral dans l'écrit' ?–le profil variationnel des sms (textos) et leur valeur pour la recherche linguistique. *Dans : Les variations diasystématiques et leurs interdépendances dans les langues romanes. Actes du Colloque DIA II à Copenhague (19–21 nov. 2012)*, pages 395–405.

Tagg, C. (2012). *Discourse of Text Messaging : Analysis of SMS Communication.* A&C Black.

Tarrade, L. (2017). Normalisation des messages issus de la communication électronique médiée. *Sciences de l'Homme et Société.*

Templin, M. (1957). Certain language skills in children : Their development and interrelationships (monograph series no. 26). *Minneapolis : University of Minnesota, The Institute of Child Welfare.*

THENG, Y. (2009). *Handbook of Research on Digital Libraries : Design, Development, and Impact : Design, Development, and Impact*. Information Science Reference.

TJONG KIM SANG, E. F. (2002). Introduction to the conll-2002 shared task : Language-independent named entity recognition. *Dans : Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4. Association for Computational Linguistics.

TJONG KIM SANG, E. F. et DE MEULDER, F. (2003a). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *Dans : Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

TJONG KIM SANG, E. F. et DE MEULDER, F. (2003b). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *Dans : Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

TORRES-MORENO, J.-M. (2011). *Résumé automatique de documents : une approche statistique*. Lavoisier.

TOUTANOVA, K. et MOORE, R. C. (2002). Pronunciation Modeling for Improved Spelling Correction. *Dans : Proceedings of the 40th Annual Meeting on Associ- ation for Computational Linguistics*, pages 144–151.

TRAN, M. et MAUREL, D. (2006). Prolexbase : Un dictionnaire relationnel multilingue de noms propres. *Traitement automatique des langues*, 47(3):115–139.

TROUILLEUX, F. (1998). Thingfinder prototype english version 2.0.

TSAKONA, V. (2009). Linguistic creativity, secondary orality, and political discourse : The modern greek myth of the" eloquent orator". *Journal of Modern Greek Studies*, 27(1):81–106.

Tyers, F. et Sánchez-Martínez, F. (2010). Free/open-source resources in the Apertium platform for machine translation research and development. *The Prague Bulletin of of Mathematical Linguistics No. 93*, (93):67–76.

Van Rijsbergen, C. (1979). Information retrieval. dept. of computer science, university of glasgow.

Véronis, J. et Guimier de Neef, E. (2006). Le traitement des nouvelles formes de communication écrite. *Compréhension automatique des langues et interaction*, pages 227–248.

Vicente, M. R. (2005). La glose comme outil de désambiguïsation référentielle des noms propres purs. *Corela - Cognition, représentation, langage*, (HS-2).

Wacholder, N., Ravin, Y. et Choi, M. (1997). Disambiguation of proper names in text. *Dans : Proceedings of the fifth conference on Applied natural language processing*, pages 202–208. Association for Computational Linguistics.

Wakao, T., Gaizauskas, R. et Wilks, Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. *Dans : Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 418–423. Association for Computational Linguistics.

Wan, X., Zong, L., Huang, X., Ma, T., Jia, H., Wu, Y. et Xiao, J. (2011). Named entity recognition in chinese news comments on the web. *Dans : IJCNLP*, pages 856–864.

Weiser, S., Cougnon, L.-A. et Watrin, P. (2011). Temporal expressions extraction in sms messages. *Dans : RANLP Workshop on Information Extraction and Knowledge Acquisition*, pages 41–44.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Williams, G. (2014). *French discourse analysis : The method of post-structuralism*. Routledge.

Wilmet, M. (1995). Le nom propre en linguistique et en littérature. *Communication à la séance mensuelle du*, vol. 13.

Xanthos, A. (2013). L'évaluation (de l'évaluation)+ de la diversité lexicale. *Mélanges offerts en hommage à Remi Jolivet*, 36:231–252.

Yamada123, I., Takeda, H. et Takefuji, Y. (2015). Enhancing named entity recognition in twitter messages using entity linking. *ACL-IJCNLP 2015*, page 136.

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing. *Pragmatics and beyond New Series*, pages 29–46.

Yvon, F. (2010). Rewriting the orthography of SMS messages. *Natural Language Engineering*, 16(02):133.

Zenasni, S., Kergosien, E., Roche, M. et Teisseire, M. (2016). Découverte de nouvelles entités et relations spatiales à partir d'un corpus de sms. *Dans : Actes de la conférence TALN'16*. AFCP.

Annexe A

Corpus alpes4science

Extrait du corpus brut anonymisé

coucou. Je suis déjà en week End ! si vous avez besoin de moi ces jours ci ia pas de problème.

Bonne soirée

Coucou. T'as eu mes messages ? T'aime. Bisous doux . À demain mon poussin ..

Sa valeur vient justement d'un ajax :) Je peux mettre une variable de session dans une page php vers laquelle Ajax envoie des données ? (J'aime les

J'ai pas compris ta sonnerietu veux que je fasse koi ?.

Si ! C bn,jsui priz. :-)

Elle habite ou ?

Ah, bisou.Alors qu'il va se marier demain.C'est maaaaal.

En fait jsais pas tp pcq david guetta m'a invitè å ibiza et lady gaga å miami als villard...

Ca m'étonne pas bon weekend !

Coucou ma belle. J'espère que ca va. Plein de bises. Je pense fort à toi.

Non pas un film d'horreur, encore une comédie romantique pourrie. Constituer un corpus en

gros.

Vs éte la ce soir ?

Yeah !

Ok. Bon courage bisous

Encore au taf ? ?

Hi hi hi ... j'adore !

J'amène quoi alors ? ? ps :mes cheveux sont bouclé ;-)

Tabac de chez moi : fermeture exceptionnelle .... dis à ***PRENOM_621_4*** de passer à celui du tram ...

Tu avais beaucoup de travail, j'allais pas te déranger. Lundi, ca va etre dur pour moi. Mais après pourquoi pas ! Ça va toi ?

Si arrete de me faire stresser, je vais conduire la et apres jy vais..

Ah ouais ? ? Éh ben, contente de l'apprendre ! Heureusment que j'vais pas souvent dessus et que je divulgue pas des secrets x) Cmt Ça s'fait que tu c

Salut ! Bon ben ca a pas fait a la chocolaterie ! Tant pis !

les filles proposent happy au subway a18h Ça te dit ? !

Génial. Bonne nuit

TEL_10

Je ne te les ai pas donné en revenant de berlioz hier ?

Il est pas là

Hein ? Ça va pas lol j'vais m'faire remarquer à mort. Encore en amphi j'aurai pas dit ms là on est ds une salle de colloque ac des tables et des cha

Ok ben je peux rien te promettre... Courage travail bien bisous et bonne soirée

Dsl j en pouvais plus d' être dans la voiture dans la rue... Bisous je t aime

pk ?

tu veux les quitter ? ... tes 12 m2

Han ***PRENOM_277_6*** a dit merdique !

ouais faut que je fasse un peu de lessive sinon c la merde trou de balle ! j'espère que y'aura un

transport a 17h parce que chui fatiguée et j'ai pas envie d'attendre. les cours c'était très long et j'ai rien compris haha...a toute bisous !

Laisse Moi 10min de plus ! Ou monte ! :)

Salut ***PRENOM_874_5*** !!! Cmt va ? ? Alors tes cours Ça s'passe ? Pas trop dur le M1 ? Ça t'plait ? Tu continues ATBCD ? Et le bad ? Raconte moi touuuut ! GrOs BizOuuuux !!!

J'ai fini mes devoirs :)

Cherchons.com est tjrs vivant apparemment. Je n'ai pas eu de pbs ici, pourtant je faisais des install reseau

Ben oui puisque c'est toi la neige. Amour

Oh mais quelle poufiasse !T'es trop une menteuse !

J'ai des trucs a te faire signer.

Revelation : J'Veux un mec qui possede un magasin de bonbons !

Oui lol ! Moi aussi jveux des bonbeks xD J'vais l'agresser le gamin mdr

Ca va bien passé la réunion. Je dirais petit ballon ! ! Bisous je t aime

coucou Ça va bien ? !je viens de voir que il y à une interprétation de moulin rouge ce soir à l amphidice aussi !kes ki te tente ?en mem tp le truc dim

PRENOM_569_5 a des soucis de doigts ;_)

Non c bon. Ca va. Bon rentrage ! Bisous ject aime

Par contre g pas de pochette.

Bah j'vais rentrer chez moi.. Et puis voilà.. J'vais tenter la guérison..

Salut, l'étude du notaire m'a appelé mais je n'étais pas dispo et je n'ai plus le numéro... Et est-ce que ***PRENOM_112_4*** t'a rappelée pour sfo ? Bises ***PRENOM_1054_2*** ***SURNOM_3*** ?On joue à Pyramide ?

Salut. Je repond à ton mail. Désolé mais je n'ai pas de commande babywalz à passer pour le moment. C sur que c embetant pour toi d'avoir autant de

Tant pis lol

Je suis seule jusqu a 12h. Cool.

Jviens chez toi à quelle h ?

Mise a jour ok. Rapide. Je vais manger une choucroute ! Pour 2 !lol. Ca va ? Moi oui. Je me repose et après je pâtisse ! Bisous je t aime

bin rien sa va les cours Mistigrif le sport et toi ?

T'inquiètes pas ma p'tite chérie,je pense à toi. Pas de répit entre 2 coups de balai,le repassage et les courses etc... Vivement demain que je boss

pàf les homme sont parti courir. Laché de gnou !

Bisous. Tu dors ? Sinon tu peux tel

TEL_10

c'est parce que je t'aime, mais amicalement !

Oué il vient avec des potes mais j'pense que Ça marchera pas T'faÇon

Quick aappartient aux francais depuis 4 ans...

Nan jcroi pa kil y en ai

La séance que ***PRENOM_989_3*** voulais ratraper est en décembre le mardi 14. Donc selon moi nous aurons dance mardi. =)

Ok. Tenez moi au courant pour la mer.

Je vais chercher le colis. Je le ramène ce soir. Si il rentre pas dans le top case, je l ouvrirai ! LOL. Bisous je t aime

Je peux mettre des points en moins pour le motif Votre site fait planter mon navigateur ?

Bjr ma petite chérie ! J'ai super bien dormi, un bon petit dej

Devine avec qui je suis et qui me parle de toi...

J'ai un beug sur ma boite mail je t'envoie ca via face book

Je suis arrivé

Coucou. Alors ? Tu te reposes ?

Exact ! Lol

Jc pas si t'as vu tn horoscope ds le 20minutes ms il est top génial als arrete d'te plaindre ! ;-)

Et bé.Tu veux du thé ?

oui mais pas tout de suite vers 9h 30 ou je dois faire un truc avant. Je te tien au courant d ke g fini

Ça va ?

Merci ***PRENOM_502_8*** d avoir tenu la porte :D

Coucou. Tu es chez toi ? Bises

Ouai ! ! Je m'attendai pa a ca ! ! !

y ?

ok ! A bientôt

Apel moi kan t sortie du cardiologue

Ouai jsui passer le prendre Tinkiete jvai bombarde tsai

Et c'était cool ? Et tu n'as pas pu faire d la moto si tu devais la réparer ? Ou si tu as qd meme pu en faire ?

ok pas de pb, et la chance t'as vu la source ? Hahaha. alors bien l'exam de ce matin ?

Salut papa ! Come vai ? T ds le coup ce soir ?

On est à l'arret d tram taillier tu viens toujours chez ***PRENOM_583_5*** ?

Salut. T'as vu le directeur ?

:D Je suis toujours sage !

On te les a récupérés oui ;) bonne soirée=)

Bonne journée mon ange

ben oué on sort mercredi ou jeudi soir ...

niarrive !

Ok dommage on aurait ou s appeller, bon âprem et bon ap. Bisous

pardon, oui ?

Tout s'explique !

Extrait du corpus transcrit manuellement

Coucou. Je suis déjà en week-end ! Si vous avez besoin de moi ces jours ci [/il n'/]y a pas de problème. Bonne soirée

Coucou. Tu as eu mes messages ? [/Je /]t'aime. Bisous doux . À demain mon poussin ..

Sa valeur vient justement d'un ajax :) Je peux mettre une variable de session dans une page php vers laquelle Ajax envoie des données ? (J'aime les

J[/e n/]'ai pas compris ta sonnerie tu veux que je fasse quoi ?.

Si ! C'est bon, je suis prise. :-)

Elle habite où ?

Ah, bisou. Alors qu'il va se marier demain. C'est mal.

Ça m'étonne pas bon weekend !

Coucou ma belle. J'espère que ça va. Plein de bises. Je pense fort à toi.

Non pas un film d'horreur, encore une comédie romantique pourrie. Constituer un corpus en gros.

Vous êtes là ce soir ?

Yeah !

OK. Bon courage bisous

Encore au travail ? ?

Hi hi hi ... J'adore !

J'amène quoi alors ? ? Ps :mes cheveux sont bouclés ;-)

Tabac de chez moi : fermeture exceptionnelle .... Dis à ***PRENOM_621_4*** de passer à celui du tramway ...

Tu avais beaucoup de travail, j[/e n/]'allais pas te déranger. Lundi, ça va être dur pour moi. Mais après pourquoi pas ! Ça va toi ?

Si arrête de me faire stresser, je vais conduire là et après j'y vais..

Ah ouais ? ? Eh ben, [/je suis /]contente de l'apprendre ! Heureusment que j[/e ne/] vais pas souvent dessus et que je [/ne/] divulgue pas des secrets x) comment Ça se fait que tu sais

Salut ! Bon ben ça [/n'/]a pas fait à la chocolaterie ! Tant pis !

Les filles proposent happy au subway à 18h [/est-ce que/] ça te dit ? !

Génial. Bonne nuit

TEL_10

Je ne te les ai pas donné en revenant de Berlioz hier ?

Il [/n'/]est pas là

Hein ? Ça [/ne/] va pas LOL je vais me faire remarquer à mort. Encore en amphithéâtre je [/n'/]aurai pas dit mais là on est dans une salle de colloque avec des tables et des chaises

OK ben je [/ne/] peux rien te promettre... Courage travaille bien bisous et bonne soirée

Désolé je [/n'/]en pouvais plus d' être dans la voiture dans la rue... Bisous je t'aime

Pourquoi ?

Tu veux les quitter ? ... tes 12 mètres carrés

Han ***PRENOM_277_6*** a dit merdique !

Ouais [/il/] faut que je fasse un peu de lessive sinon c'est la merde trou de balle ! J'espère qu[/'il/] y aura un transport à 17h parce que je suis fatiguée et j[/e n/]'ai pas envie d'attendre. Les cours c'était très long et j[/e n/]'ai rien compri...

Laisse-moi 10 minutes de plus ! Ou monte ! :)

Salut ***PRENOM_874_5*** ! ! ! Comment va[/s-tu/] ? ? Alors tes cours Ça se passe ? [/Ce n'est/] pas trop dur le master 1 ? Ça te plait ? [/Est-ce que/] tu continues _ATBCD_ ? Et le badminton ? Raconte moi tout ! GrOs bisous ! ! !

J'ai fini mes devoirs :)

IMPOSSIBLE

Ben oui puisque c'est toi la neige. Amour

Oh mais quelle poufiasse ! Tu es trop une menteuse !

J'ai des trucs a te faire signer.

Révélation : Je veux un mec qui possède un magasin de bonbons !

Oui LOL ! Moi aussi je veux des bonbons xD je vais l'agresser le gamin mdr

Ça va [/ça s'est/] bien passé la réunion. Je dirais petit ballon ! ! Bisous je t'aime

Coucou Ça va bien ? ! Je viens de voir qu' il y a une interprétation de moulin rouge ce soir à l'amphidice aussi ! Qu'est-ce qui te tente ? En même temps le truc dimanche

PRENOM_569_5 a des soucis de doigts ; _)

Non c'est bon. Ça va. Bon retour ! Bisous je t'aime

Par contre je [/n'ai/] pas de pochette.

Bah je vais rentrer chez moi.. Et puis voilà.. Je vais tenter la guérison..

Salut, l'étude du notaire m'a appelé mais je n'étais pas disponible et je n'ai plus le numéro... Et est-ce que ***PRENOM_112_4*** t'a rappelée pour _sfo_ ? Bises ***PRENOM_1054_2***

SURNOM_3 ? On joue à Pyramide ?

Salut. Je réponds à ton mail. Désolé mais je n'ai pas de commande babywalz à passer pour le moment. C'est sûr que c'est embêtant pour toi d'avoir autant de

Tant pis LOL

Je suis seule jusqu'à midi. Cool.

Je viens chez toi à quelle heure ?

Mise à jour OK. Rapide. Je vais manger une choucroute ! Pour deux ! LOL. Ça va ? Moi oui. Je me repose et après je pâtisse ! Bisous je t'aime

Bin rien ça va les cours Mistigrif le sport et toi ?

Paf les hommes sont partis courir. Laché de gnous !

Bisous. Tu dors ? Sinon tu peux téléphoner

TEL_10

C'est parce que je t'aime, mais amicalement !

Ouais il vient avec des potes mais je pense que ça [/ne/] marchera pas [/de /] toute façon

Quick appartient aux francais depuis 4 ans...

Non je [/ne/] crois pas qu'il y en ai

La séance que ***PRENOM_989_3*** voulais rattraper est en décembre le mardi 14. Donc selon moi nous aurons dance mardi. =)

OK. Tenez moi au courant pour la mer.

Je vais chercher le colis. Je le ramène ce soir. Si il [/ne/] rentre pas dans le top case, je l'ouvrirai ! LOL. Bisous je t'aime

Je peux mettre des points en moins pour le motif Votre site fait planter mon navigateur ?

Bonjour ma petite chérie ! J'ai super bien dormi, un bon petit déjeuner

Devine avec qui je suis et qui me parle de toi...

J'ai un bug sur ma boite mail je t'envoie ça via facebook

Je suis arrivé

Coucou. Alors ? Tu te reposes ?

Exactement ! LOL

Je [/ne/] sais pas si tu as vu ton horoscope dans le 20 minutes mais il est top génial alors arrête de te plaindre ! ;-)

Et bien. [/Est-ce que/] tu veux du thé ?

Oui mais pas tout de suite vers 9h 30 ou je dois faire un truc avant. Je te tiens au courant dès que j'ai fini

Ça va ?

Merci ***PRENOM_502_8*** d'avoir tenu la porte :D

Coucou. Tu es chez toi ? Bises

Ouais ! ! Je [/ne/] m'attendai pas à ça ! ! !

IMPOSSIBLE

OK ! À bientôt

Appelle moi quand tu es sortie du cardiologue

Ouais je suis passé le prendre [/ne/] t'inquiète [/pas/] je vais bombarder tu sais

Et c'était cool ? Et tu n'as pas pu faire de la moto si tu devais la réparer ? Ou si tu as quand même pu en faire ?

OK pas de problème, et la chance tu as vu la source ? Hahaha. Alors [/c'était/] bien l'examen de ce matin ?

Salut papa ! Ça va ? [/Est-ce que/] tu es dans le coup ce soir ?

On est à l'arrêt de tramway taillier [/est-ce que/] tu viens toujours chez ***PRENOM_583_5*** ?

Salut. Tu as vu le directeur ?

:D Je suis toujours sage !

On te les a récupérés oui ;) bonne soirée=)

Bonne journée mon ange

Ben ouais on sort mercredi ou jeudi soir ...

J'arrive !

OK dommage on aurait pu s'appeller, bon après-midi et bon appétit. Bisous

Pardon, oui ?

Tout s'explique !

Extrait du corpus aligné

```
<tok>
  <v lang="fr">Moi</v>
  <lang="sms">mwa</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">ça</v>
  <lang="sms">ca</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">va</v>
  <lang="sms">va</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">merci</v>
  <lang="sms">merci</v>
</tok>
<tok>
  <v lang="fr">N'oublie</v>
  <lang="sms">N'oublie</v>
</tok>
<punct value=" "/>
```

```
<tok>
  <v lang="fr">pas</v>
  <lang="sms">pas</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">de</v>
  <lang="sms">de</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">donner</v>
  <lang="sms">donner</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">les</v>
  <lang="sms">les</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">abricots</v>
  <lang="sms">abricots</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">secs</v>
  <lang="sms">secs</v>
</tok>
<punct value=". "/>
<tok>
  <v lang="fr">Je t'aime</v>
  <lang="sms">Jtm</v>
</tok>
<tok>
  <v lang="fr">C'est</v>
  <lang="sms">C'est</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">ce</v>
  <lang="sms">ce</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">que</v>
  <lang="sms">que</v>
</tok>
<punct value=" "/>
```

```
<tok>
  <v lang="fr">je</v>
  <lang="sms">je</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">me</v>
  <lang="sms">me</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">dis</v>
  <lang="sms">dis</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">aussi</v>
  <lang="sms">aussi</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">:)</v>
  <lang="sms">:)</v>
</tok>
<tok>
  <v lang="fr">Je</v>
  <lang="sms">je</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">ne</v>
  <lang="sms">ne</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">sais</v>
  <lang="sms">sais</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">vraiment</v>
  <lang="sms">vraiment</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">pas</v>
  <lang="sms">pas</v>
</tok>
<punct value=" "/>
```

```
<tok>
  <v lang="fr">***PRENOM_1069_4***</v>
  <lang="sms">***PRENOM_1069_4***</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">je [/ne/] pense</v>
  <lang="sms">jpense</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">pas</v>
  <lang="sms">pas</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">trop</v>
  <lang="sms">trop</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">non</v>
  <lang="sms">Nan</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">plus</v>
  <lang="sms">plus</v>
</tok>
<tok>
  <v lang="fr">J'ai</v>
  <lang="sms">j'ai</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">des</v>
  <lang="sms">des</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">cons</v>
  <lang="sms">cons</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">à</v>
  <lang="sms">à</v>
</tok>
<punct value=" "/>
```

```
<tok>
  <v lang="fr">longueur</v>
  <lang="sms">longueur</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">de</v>
  <lang="sms">de</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">journée</v>
  <lang="sms">journée</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">dans</v>
  <lang="sms">dans</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">mon</v>
  <lang="sms">mon</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">bureau</v>
  <lang="sms">bureau</v>
</tok>
<punct value="... "/>
<tok>
  <v lang="fr">Deux</v>
  <lang="sms">2</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">de</v>
  <lang="sms">de</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">plus</v>
  <lang="sms">+</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">ou</v>
  <lang="sms">ou</v>
</tok>
```

```
<punct value=" "/>
<tok>
  <v lang="fr">de</v>
  <lang="sms">de</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">moins</v>
  <lang="sms">-</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">ça</v>
  <lang="sms">ça</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">[/ne/] change</v>
  <lang="sms">change</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">pas</v>
  <lang="sms">pas</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">grand</v>
  <lang="sms">grand</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">chose</v>
  <lang="sms">chose</v>
</tok>
<tok>
  <v lang="fr">Jeudi</v>
  <lang="sms">Jeudi</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">m'arrangerait</v>
  <lang="sms">m'arrangerait</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">plus</v>
  <lang="sms">plus</v>
</tok>
```

```
<punct value=" "/>
<tok>
  <v lang="fr">en</v>
  <lang="sms">en</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">fait</v>
  <lang="sms">fait</v>
</tok>
<tok>
  <v lang="fr">^^</v>
  <lang="sms">^^</v>
</tok>
<tok>
  <v lang="fr">Je</v>
  <lang="sms">Je</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">vais</v>
  <lang="sms">vais</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">en</v>
  <lang="sms">en</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">cours</v>
  <lang="sms">cours</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">je</v>
  <lang="sms">je</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">t'appelle</v>
  <lang="sms">tapl</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">en</v>
  <lang="sms">en</v>
</tok>
<punct value=" "/>
```

```
<tok>
  <v lang="fr">sortant</v>
  <lang="sms">sortant</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">vers</v>
  <lang="sms">vers</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">10h30</v>
  <lang="sms">10h30</v>
</tok>
<punct value=". "/>
<tok>
  <v lang="fr">Bise</v>
  <lang="sms">Bizz</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">courage</v>
  <lang="sms">courage</v>
</tok>
<punct value="."/>
<tok>
  <v lang="fr">À</v>
  <lang="sms">A</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">toi</v>
  <lang="sms">toi</v>
</tok>
<punct value=" "/>
<tok>
  <v lang="fr">aussi.</v>
  <lang="sms">aussi.</v>
</tok>
```

Annexe B

Questionnaire du projet

Le questionnaire dans Chabert et al. (2012)

Étape 1 :

Quel est votre âge ? :

Si l'âge est inférieur à 18 :

❏ j'ai moins de 18 ans, je certifie avoir l'autorisation de mes parents pour remplir ce questionnaire.

Étape 2 :

Entrez votre numéro de téléphone :

Étape 3 :

Sexe : ❏ féminin

❏ masculin

Étape 4 :

Langue maternelle :

Étape 5 :

Quelle(s) autre(s) langue(s) parlez-vous couramment ?

Étape 6 :

Langue régionale :

Étape 7 :

Niveau d'étude :

Dans quel(s) domaine(s) travaillez-vous ?

❏ Sans emploi

❏ Administration des affaires

❏ Agriculture ; Art

❏ Sciences

❏ Commerce, affaire et économie

❏ Communications

❏ Droit

❏ Education (enseignants)

❏ Etudes (étudiants)

❏ Gouvernement et service publique (élus, fonctionnaires)

❏ Santé

❏ Foresterie

❏ Génie

❏ Journalier

❏ Personnel de bureau

❏ Sports

❏ Technologies (incluant technologies de l'information)

❏ Vente

❏ Autre

Si « Autre », précisez :

Si vous êtes collégien ou lycéen : Quelle est votre classe ? Etes vous en filière ❏ technique / ❏ générale ?

Étape 8 :

Code postal de votre domicile :

Si vous n'êtes pas originaire de ce département : Depuis combien de temps y vivez-vous ?

Quel est votre département/pays d'origine ?

Étape 9 :

Depuis combien de temps utilisez-vous les SMS ?

Avez-vous un abonnement avec SMS illimités ?

Combien de SMS envoyez-vous, en moyenne, par semaine ?

Combien de SMS recevez-vous, en moyenne, par semaine ?

Étape 10 :

Classez les destinataires de vos SMS par ordre de fréquence

	jamais	rarement	occasionnellement	parfois	souvent	fréquemment
Famille						
Amis						
Collègues						
Concours/jeux						

Étape 11 :

Codes et abréviations : Comprenez-vous facilement les SMS que vous recevez s'ils contiennent des codes ou abréviations ?

Étape 12 :

Utilisez-vous le dictionnaire d'aide à la rédaction (T9), c'est-à-dire la complétion automatique de votre téléphone comme aide à la rédaction de SMS ?

Étape 13 :

Quel type de clavier utilisez-vous ?

❏Alpha-numérique

❏Azerty

Étape 14 :

Quand vous rédigez des SMS, vous utilisez :

❏ le langage SMS

❏ la langue « normale »

❏ un mélange des deux

Vos SMS sont écrits en :

❏ français

❏ langue régionale

❏ autre

❏ un mélange de langues

Votre façon de rédiger vos SMS est-elle différente selon le destinataire ? Si oui expliquez.

Dans quelle situation utilisez-vous les SMS en priorité ?

Étape 15 :

De qui recevez-vous les plus de SMS ?

	jamais	rarement	occasionnellement	parfois	souvent	fréquemment
Famille						
Amis						
Collègues						
Concours/jeux						

La majorité des SMS que vous recevez sont rédigés :

❏ en langage SMS

❏ en langue « normale »

❏ en mélange des deux

Étape 16 :

Quels autres modes ou systèmes de communications sur Internet utilisez-vous ?

❏ Le courrier électronique

❏ les systèmes de clavardage (chat) ou de messagerie instantanée dans Internet (MSN Messenger, Yahoo Messenger, etc.)

❏ les forums dans Internet

❏ les réseaux sociaux (Facebook, etc.)

Étape 17 :

Commentaires Si vous le souhaitez, vous pouvez faire un commentaire sur la manière dont vous écrivez des SMS.

ANNEXE B. QUESTIONNAIRE DU PROJET

Annexe C

Données lexicometriques du corpus alpes4science

Taux de % grammaticaux

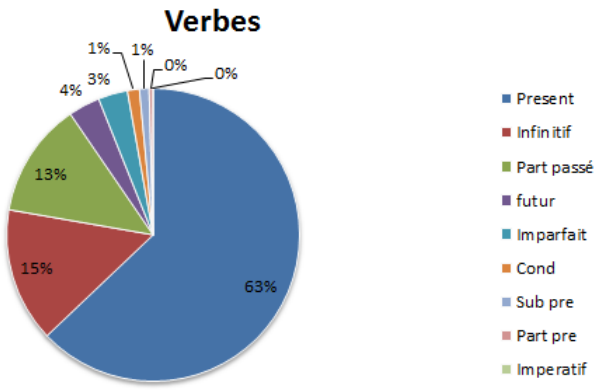

FIGURE C.1 – Taux en (%) des formes verbales

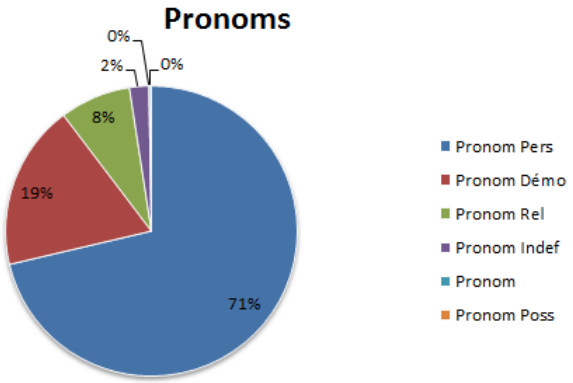

FIGURE C.2 – Taux en (%) des formes pronominales

Tokens plus fréquents du corpus

Rank	Freq	Mot	Rank	Freq	Mot	Rank	Freq	Mot	Rank	Freq	Mot
1	11453	je	26	2143	bien	51	1170	as	76	761	aussi
2	6400	à	27	2051	suis	52	1168	vous	77	756	faire
3	6331	tu	28	2027	mais	53	1141	demain	78	753	elle
4	6309	est	29	2008	va	54	1098	plus	79	747	sur
5	6253	de	30	1971	a	55	1084	es	80	736	après
6	5240	pas	31	1886	les	56	1079	avec	81	724	s
7	4335	que	32	1848	te	57	1070	lol	82	707	qui
8	4120	ça	33	1837	ok	58	1055	bonne	83	693	ouais
9	3912	et	34	1824	moi	59	1052	des	84	680	être
10	3766	j	35	1575	toi	60	1027	coucou	85	666	ton
11	3742	le	36	1478	si	61	1011	mon	86	659	ou
12	3535	la	37	1476	au	62	991	non	87	653	veux
13	3367	c	38	1467	y	63	969	m	88	623	sais
14	3366	ce	39	1458	me	64	911	vais	89	607	même
15	2980	pour	40	1450	d	65	895	merci	90	570	ben
16	2955	bisous	41	1393	oui	66	885	se	91	570	gros
17	2854	ai	42	1350	h	67	875	soir	92	559	fais
18	2810	en	43	1345	bon	68	859	alors	93	541	midi
19	2748	on	44	1320	qu	69	857	quand	94	538	quoi
20	2722	ne	45	1305	aime	70	855	ma	95	532	encore
21	2717	il	46	1272	une	71	832	trop	96	520	comme
22	2604	t	47	1259	fait	72	820	là	97	498	salut
23	2445	un	48	1224	tout	73	794	heure	98	497	voir
24	2242	l	49	1218	du	74	783	chez	99	485	peut
25	2190	n	50	1175	dans	75	775	peux	100	480	dit

Table C.1 – 100 tokens plus fréquents du corpus

Annexe D

Typologie de SMS

Typologie de SMS selon Panckhurst (2009, 2015,2017)

Substitution

Phonétique	Entière : o (eau), 7 (cet) Partielle : ossi (aussi), allé (aller), bizes (bises) Avec variation : k (que), kikou (coucou)
Graphique	Élision, typographie, majuscules / minuscules : m en, est ce que Icônes, symboles, rébus : à + (à plus), de grandes @ (oreilles) Avec variations : bisoux (bisous), mwa (moi)

Réduction

Phonetique	Raccourcissements morpho-lexicaux : initialismes - alphabétismes & acronymes : ASV, mdr, tvb, tlm, lol troncations - apocope ; aphérèses : ordi (ordinateur), 'lut, Net Entière : c (c'est), d (des/dé/dès), v (vais) Variation : ui (oui)
Graphique	Suppression des suffixes muets ; préfixes : vous (vous), peu (peut), ôte (hôtel), douch (douche) Contractions consonantiques / coupures & abréviations : dc (donc, pr (pour), ds (dans) ; double consonne : ele(elle), poura (pourra) ; abréviations sémantiques : t (te/tu), p (peut/-peux) Agglutinations : jattends (j'attends)

Suppression

Graphique	Typographie & ponctuation : [...] se genre de truc pr le site je pense ke ca devré allé vite je vou envéré [...] Signes Diacritiques : ca (ça), voila (voilà)

Ajout

Graphique	Répétition (caractères, ponctuation) : suuupppeer ! ! ! ! Ajout d'un caractère muet : peux (peu), as (a) Représentations sémiologiques (smileys/emoji) Onomatopée : snif, bof
Phonétique	Partielle substitution phonétique : a) ajout d'un caractère, pas de modification phonétique : reparler (reparlé) b) liaisons : zaime (aime) substitution phonétique avec variation : oki (ok), ouaip (oui)

Table D.1 – Typologie modifiée de l'écriture SMS figurant dans Panckhurst 2009, 2015

Annexe E

Graphe de reconnaissance

Graphe

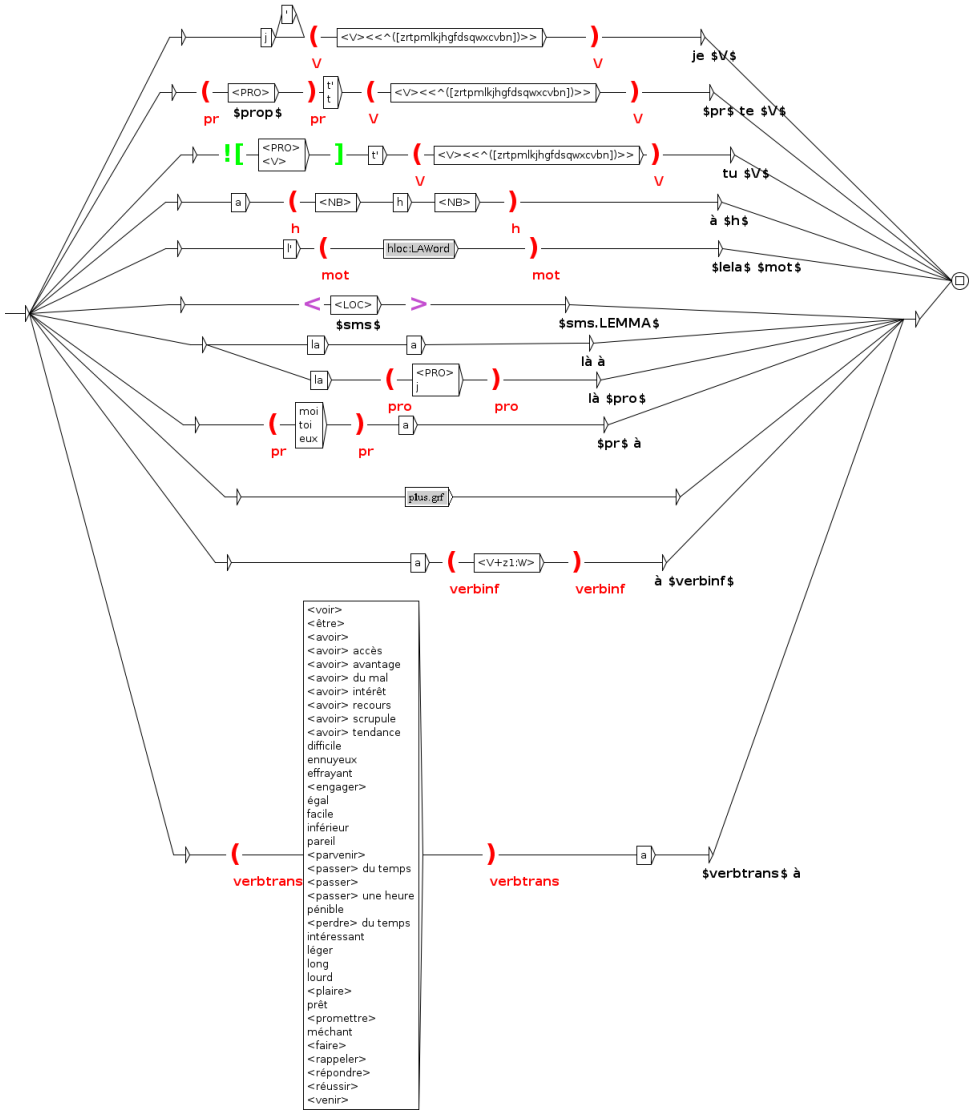

FIGURE E.1 – Graphe d'expressions

Annexe F

Corpus de normalisation

Extrait du corpus parallèle smsfra-fra

Sentence pair (14) source length 60 target length 54 alignment score : 1.55442e-27

je<prn><tn><p1><mf><sg> ne<adv> savoir<vblex><pri><p1><sg> pas<adv> ou<cnjcoo> être<vbser><pri><p3><pl> le<det><def><mf><pl> loutrons!<sent> le<prn><pro><p3><f><sg> bbsitter le<prn><pro><p3><mf><pl> avoir<vbhaver><pri><p3><sg> chercher<vblex><pp><m><sg> partout<adv> .<sent> le<det><def><m><sg> moz ne<adv> repond pas<adv> .<sent> j espere kils être<vblex><pri><p3><pl> devant<pr> le<det><def><f><sg> maison<n><f><sg> avec<pr> le<det><def><mf><pl> bops .<sent> je<prn><tn><p1><mf><sg> crise<n><f><sg>!<sent> j ose pas<adv> appeler<vblex><inf> le<det><def><mf><pl> bops j avoir<vbhaver><pri><p1><sg> trop<adv> peur<n><f><sg> s<n><m><sp> il<prn><tn><p3><m><pl> ne<adv> être<vbser><pri><p3><pl> pas<adv> avec<pr> eux<prn><tn><p3><m><pl> .<sent>

NULL ( ) je<prn><tn><p1><mf><sg> (1) ne<adv> (2) savoir<vblex><pri><p1><sg> (3) pas<adv> (4) ou<cnjcoo> (5) être<vbser><pri><p3><pl> (6) le<det><def><mf><pl> (7) surnom<n><m><sg> (8) 8<num> ( )!<sent> (9) le<prn><pro><p3><f><sg> (10) baby ( ) sitter (11) le<prn><pro><p3><mf><pl> (12) avoir<vbhaver><pri><p3><sg> (13) chercher<vblex><pp><m><pl> (14) partout<adv> (15) .<sent> (16) le<det><def><m><sg> (17) surnom<n><m><sg> (18) 3<num> ( ) ne<adv> (19) repond (20) pas<adv> (21) .<sent> (22) j (23) espere (24) qu ( ) il<prn><tn><p3><m><pl> (25) être<vblex><pri><p3><pl> (26) devant<pr> (27) le<det><def><f><sg> (28) maison<n><f><sg> (29) avec<pr> (30) le<det><def><mf><pl> (31) surnom<n><m><sg> (32) 4<num> ( ) .<sent> (33) je<prn><tn><p1><mf><sg> (34) crise<n><f><sg> (35)!<sent> (36) j (37) ose (38) pas<adv> (39) appeler<vblex><inf> (40) le<det><def><mf><pl> (41) surnom<n><m><sg> (42) 4<num> ( ) j (43) avoir<vbhaver><pri><p1><sg> (44) trop<adv> (45) peur<n><f><sg> (46) s<n><m><sp> (47) il<prn><tn><p3><m><pl> (48) ne<adv> (49) être<vbser><pri><p3><pl> (50) pas<adv> (51) avec<pr> (52) eux<prn><tn><p3><m><pl> (53) .<sent> (54)

Sentence pair (15) source length 23 target length 23 alignment score : 7.49593e-07

ah<ij> ,<cm> t<n><m><sp> être<vbser><pri><p2><sg> avec<pr> ton<det><pos><m><sg> binome!<sent> tant<adv> pis<adj><mf><sp> ,<cm> on<prn><pro><ref><p3><mf><sp> prendre<vblex><fti><p3><sg> le<det><def><m><sg> thé<n><m><sg> tout<predet><m><sg> seul<adj><m><pl> (<lpar> avoir<vblex><pri><p3><sg> 200<num> ,<cm> quoi<prn><itg><mf><sp>)<rpar>

NULL ( ) ah<ij> (1) ,<cm> (2) tu<prn><tn><p2><mf><sg> (3) être<vblex><pri><p2><sg> (4) avec<pr> (5) ton<det><pos><m><sg> (6) binôme (7)!<sent> (8) tant<adv> (9) pis<adj><mf><sp> (10) ,<cm> (11) on<prn><pro><ref><p3><mf><sp> (12) prendre<vblex><fti><p3><sg> (13) le<det><def><m><sg> (14) thé<n><m><sg> (15) tout<predet><m><pl> (16) seul<adj><m><pl> (17) (<lpar> (18) à<pr> (19) 200<num> (20) ,<cm> (21) quoi<prn><itg><mf><sp> (22))<rpar> (23)

Sentence pair (16) source length 18 target length 17 alignment score : 1.83629e-18

*chala aller<vblex><pri><p1><sg> à<pr>+le<det><def><m><sg> dodo mon<det><pos><m><sg> coeur<n><m><sg> .<sent> .<sent> .<sent> avoir<vblex><pri><p3><sg> tout<prn><tn><m><sg> avoir<vblex><pri><p3><sg> l<n><m><sp> heure<n><f><sg> mi *amore!<sent>

NULL ( ) *chala (1 15 16) je<prn><tn><p1><mf><sg> ( ) aller<vblex><pri><p1><sg> (2) à<pr>+le<det><def><m><sg> (3) lit<n><m><sg> (4) mon<det><pos><m><sg> (5) coeur<n><m><sg> (6) .<sent> (7) .<sent> (8) .<sent> (9) à<pr> (10) tout<predet><m><sg> (11) à<pr> (12) l<n><m><sp> (13) heure<n><f><sg> (14) mon<det><pos><m><sg> ( ) amour<n><m><sg> ( )!<sent> (17)

Sentence pair (17) source length 15 target length 13 alignment score : 1.29875e-06

ben être<vblex><pri><p1><sg> dans<pr> mon<det><pos><m><sg> labo .<sent> sans<pr> message<n><m><sg> je<prn><tn><p1><mf><sg> pouvoir<vbmod><pri><p1><sg> pas<adv> chavoir emoticon

NULL ( ) ben (1) je<prn><tn><p1><mf><sg> ( ) être<vblex><pri><p1><sg> (2) dans<pr> (3) mon<det><pos><m><sg> (4) laboratoire<n><m><sg> (5) .<sent> (6) sans<pr> (7) message<n><m><sg> (8) je<prn><tn><p1><mf><sg> (9) ne<adv> ( ) pouvoir<vbmod><imp><p2><sg> (10) pas<adv> (11) savoir<vblex><inf> (12) emoticon (13)

Sentence pair (18) source length 12 target length 11 alignment score : 3.83462e-06

hello!<sent> non<adv> pas<adv> ce<det><dem><m><sg> soir<n><m><sg> .<sent> .<sent> .<sent> hs bizz

NULL ( ) hello (1)!<sent> (2) non<adv> (3) pas<adv> (4) ce<det><dem><m><sg> (5) soir<n><m><sg> (6) .<sent> (7) .<sent> (8) .<sent> (9) hors<pr> ( ) service<n><m><sg> (10) bise (11)

Sentence pair (19) source length 16 target length 14 alignment score : 7.8515e-07

c pas<adv> fo .<sent> .<sent> .<sent> mais<cnjcoo> mon<det><pos><f><sg> semaine<n><f><sg> avoir<vbhaver><pri><p3><sg> être<vbser><pp><mf><sp> charger<vblex><pp><f><sg> .<sent> bizz

NULL ( ) ce<det><dem><m><sg> (1) n ( ) être<vbser><pri><p3><sg> ( ) pas<adv> (2) faux<adj><m><sp> (3) .<sent> (4) .<sent> (5) .<sent> (6) mais<cnjcoo> (7) mon<det><pos><f><sg> (8) semaine<n><f><sg> (9) avoir<vbhaver><pri><p3><sg> (10) être<vbser><pp><mf><sp> (11) charger<vblex><pp><f><sg> (12) .<sent> (13) bise (14)

Extrait du dictionnaire bilingue

<e r="LR"> <l>cocot</l> <r>cocotte</r> </e>

<e> cocote</e>

<e r="LR"> <l>cocott</l> <r>cocotte</r> </e>

<e> cocotte</e>

<e r="LR"> <l>cocou</l> <r>coucou</r> </e>

<e> <l>cocsice</l> <r>coccis</r> </e>
<e r="LR"> <l>cod</l> <r>code<s n="n"/><s n="m"/><s n="pl"/></r> </e>
<e> code<s n="n"/></e>
<e> coder</e>
<e> codis</e>
<e> coehlo</e>
<e> coeur<s n="n"/></e>
<e> coffee</e>
<e> coffre<s n="n"/></e>
<e r="LR"> <l>cofirmer</l> <r>confirmer<s n="vblex"/><s n="inf"/></r> </e>
<e> cogitations</e>
<e> cogite</e>
<e> cogiter</e>
<e> cognin</e>
<e> coiffer<s n="vblex"/></e>
<e> coiffeur<s n="n"/></e>
<e> coin<s n="n"/></e>
<e> coince</e>
<e> <l>coincee</l> <r>coincée</r> </e>
<e> coinche</e>
<e> coincidence</e>
<e r="LR"> <l>coincée</l> <r>coincé</r> </e>
<e> coincés</e>
<e> col</e>

<e r="LR"> <l>coli</l> <r>colis<s n="n"/><s n="m"/><s n="sp"/></r> </e>
<e> colin<s n="n"/></e>
<e> coline</e>
<e> colis<s n="n"/></e>
<e> collage</e>
<e> collants</e>
<e> collecter<s n="vblex"/></e>
<e> collectif<s n="adj"/></e>
<e> collection<s n="n"/></e>
<e> collector</e>
<e> <l>college</l> <r>collège<s n="n"/><s n="m"/><s n="sg"/></r> </e>
<e r="LR"> <l>collegue</l> <r>collègue<s n="n"/><s n="mf"/><s n="sg"/></r> </e>
<e r="LR"> <l>collegues</l> <r>collègue<s n="n"/><s n="mf"/><s n="pl"/></r> </e>
<e> coller<s n="vblex"/></e>
<e> collier<s n="n"/></e>
<e> <l>colloc</l> <r>collocation</r> </e>
<e> colloque<s n="n"/></e>
<e> collu</e>
<e r="LR">collège<s n="n"/><s n="m"/><s n="sg"/></e>
<e> collège<s n="n"/><s n="m"/><s n="pl"/></e>
<e> collègue<s n="n"/></e>
<e> colmar</e>

<e> <l>colo</l> <r>colonie<s n="n"/><s n="f"/><s n="sg"/></r> </e>
<e> <l>coloc</l> <r>colocataire</r> </e>
<e> colocation<s n="n"/></e>
<e> <l>colocs</l> <r>colocataires</r> </e>
<e> colombes</e>
<e> colombia</e>
<e> colombien<s n="adj"/></e>
<e> colon</e>
<e> colorectal</e>
<e> <l>colos</l> <r>colonie<s n="n"/><s n="f"/><s n="pl"/></r> </e>
<e> coltar</e>
<e r="LR"> <l>colègu</l> <r>collègue<s n="n"/><s n="mf"/><s n="pl"/></r> </e>
<e r="LR"> <l>colègue</l> <r>collègue<s n="n"/><s n="mf"/><s n="sg"/></r> </e>
<e> colère<s n="n"/></e>
<e r="LR"> <l>colégu</l> <r>collègue<s n="n"/><s n="mf"/><s n="sg"/></r> </e>
<e r="LR"> <l>com</l> <r>comme<s n="rel"/><s n="adv"/></r> </e>
<e r="LR"> <l>coman</l> <r>comment<s n="adv"/><s n="itg"/></r> </e>
<e> combat<s n="n"/></e>
<e> combe</e>

<e> <l>combi</l> <r>combinaison<s n="n"/><s n="f"/><s n="sg"/></r> </e>
<e> combien<s n="prn"/></e>
<e> combien<s n="rel"/></e>
<e> combinaison<s n="n"/></e>
<e> comble<s n="n"/></e>
<e> combler</e>
<e> comboir</e>
<e> comboire</e>
<e r="LR"> <l>come</l> <r>comme<s n="rel"/><s n="adv"/></r> </e>
<e> <l>comedi</l> <r>comedie</r> </e>
<e> <l>comedies</l> <r>comédie<s n="n"/><s n="f"/><s n="pl"/></r> </e>
<e> comeg</e>
<e r="LR"> <l>comen</l> <r>comment<s n="adv"/><s n="itg"/></r> </e>
<e r="LR"> <l>comence</l> <r>commencer<s n="vblex"/><s n="pri"/><s n="p3"/><s n="sg"/></r> </e>
<e> comencer</e>
<e r="LR"> <l>coment</l> <r>comment<s n="adv"/><s n="itg"/></r> </e>
<e r="LR"> <l>comentaire</l> <r>commentaire<s n="n"/><s n="m"/><s n="sg"/></r> </e>
<e> <l>comere</l> <r>comère</r> </e>
<e> comestible</e>
<e> coming</e>

<e> comique<s n="n"/></e>
<e> <l>comite</l> <r>comité<s n="n"/><s n="m"/><s n="sg"/></r> </e>
<e> <l>comm</l> <r>commentaire<s n="n"/><s n="m"/><s n="sg"/></r> </e>
<e r="LR"> <l>command</l> <r>commander<s n="vblex"/><s n="pri"/><s n="p3"/><s n="sg"/></r> </e>
<e> commande<s n="n"/></e>
<e> commander<s n="vblex"/></e>
<e r="LR"> <l>commantaire</l> <r>commentaire<s n="n"/><s n="m"/><s n="sg"/></r> </e>
<e> comme</e>
<e> comme<s n="pr"/></e>
<e> comme<s n="rel"/></e>
<e r="LR"> <l>commemt</l> <r>comment<s n="adv"/><s n="itg"/></r> </e>
<e r="LR"> <l>commen</l> <r>comment<s n="adv"/><s n="itg"/></r> </e>
<e> commencants</e>
<e> commencer<s n="vblex"/><s n="inf"/></e>
<e> commencer<s n="vblex"/><s n="pii"/><s n="p1"/><s n="sg"/></e>
<e> <l>commencer<s n="vblex"/><s n="imp"/><s n="p2"/><s n="sg"/></l><r>commencer<s n="vblex"/><s n="pp"/><s n="m"/><s n="sg"/></r> </e>
<e> commencer<s n="vblex"/><s n="pri"/><s n="p2"/><s n="sg"/></e>

<e r="LR"> <l>commenece</l> <r>commencer<s n="vblex"/><s n="pri"/><s n="p3"/><s n="sg"/></r> </e>
<e> comment<s n="adv"/></e>
<e r="LR">commentaire<s n="n"/><s n="m"/><s n="sg"/></e>
<e> commentaire<s n="n"/><s n="m"/><s n="pl"/></e>
<e> commerce<s n="n"/></e>
<e> commercial<s n="adj"/></e>
<e> commercial<s n="n"/></e>
<e> commode</e>
<e> commun<s n="adj"/></e>
<e> communiquer<s n="vblex"/></e>
<e r="LR"> <l>comnence</l> <r>commencer<s n="vblex"/><s n="pp"/><s n="m"/><s n="sg"/></r> </e>
<e> compa</e>
<e> compagnie<s n="n"/></e>
<e> compagnon<s n="n"/></e>
<e> comparer<s n="vblex"/></e>
<e> compassion</e>
<e> compatible<s n="adj"/></e>
<e> compatriote<s n="n"/></e>
<e> compensation<s n="n"/></e>
<e> compenser<s n="vblex"/></e>
<e> <l>compet</l> <r>compétition<s n="n"/><s n="f"/><s n="sg"/></r> </e>
<e> <l>compiler<s n="vblex"/><s n="prs"/></l> <r>compiler<s n="vblex"/><s n="pri"/></r> </e>
<e> complet<s n="adj"/></e>

<e> complete</e>
<e r="LR"> <l>completement</l> <r>complètement<s n="adv"/></r> </e>
<e> complexe<s n="n"/></e>
<e> complication<s n="n"/></e>
<e r="LR"> <l>compliker</l> <r>compliquer<s n="vblex"/><s n="pp"/><s n="m"/><s n="sg"/></r> </e>
<e> compliment</e>
<e> compliments</e>
<e> <l>compliquer<s n="vblex"/><s n="pri"/><s n="p3"/><s n="sg"/></l><r>compliquer<s n="vblex"/><s n="pp"/><s n="m"/><s n="sg"/></r> </e>
<e> compliquer<s n="vblex"/><s n="pp"/><s n="m"/><s n="sg"/></e>
<e> complètement<s n="adv"/></e>
<e> complètment</e>
<e r="LR"> <l>complétement</l> <r>complètement<s n="adv"/></r> </e>
<e> compléter<s n="vblex"/></e>
<e> composant<s n="n"/></e>
<e> <l>composer<s n="vblex"/><s n="imp"/><s n="p2"/><s n="sg"/></l><r>composer<s n="vblex"/><s n="pp"/><s n="m"/><s n="sg"/></r> </e>
<e> composer<s n="vblex"/><s n="pp"/><s n="m"/><s n="sg"/></e>
<e> composition<s n="n"/></e>
<e> compote</e>
<e r="LR"> <l>compren</l> <r>comprendre<s n="vblex"/><s n="pri"/><s n="p1"/><s n="sg"/></r> </e>

Liste des tableaux

2.1 Trois classifications de définitions de la CMO interactive dans Mahmoud et Auter (2009) 28

2.2 Exemples d'apocope et d'aphérèse 33

3.1 Exemple de données anonymisées 47

3.2 Tableau récapitulatif de codes 48

3.3 Exemple de découpage de SMS 50

3.4 Exemple de transcription 51

3.5 Description du corpus 57

4.1 Répartition du pourcentage d'étiquettes attribuées à l'usage de SMS 68

4.2 Répartition du nombre d'envoi de SMS par semaine 69

4.3 Informations démographiques de participants 72

4.4 Exemple de formule TTR 74

4.5 Principales caractéristiques lexicométriques du corpus alpes4science 74

4.6 Principales caractéristiques lexicométriques de 4 projets SMS francophones dans Cougnon (2015) 75

4.7 La diversité lexicale en fonction de l'âge 76

4.8 La diversité lexicale en fonction du niveau d'études 79

4.9 Taux de réduction moyen en % selon le type de clavier 83

4.10 Table récapitulatif de 2 à 4-gramme) 84

4.11 Spécificités du discours numérique selon Marcoccia (2016) 86

4.12 Typologie SMS pour le corpus alpes4science 88

4.13 Graphies SMS plus fréquentes dans le corpus 89

5.1 Exemple de normalisation 101

5.2 Extrait du dictionnaire de locutions 106

5.3 Exemple de traduction sans des étiquettes morphosyntaxiques 118

5.4 Exemples erronés de normalisation 122

5.5 Tableau récapitulatif des données SMS 123

5.6 Résultats d'évaluation 124

5.7 BLEU scores de l'état de l'art 125

5.8 Catégories d'erreurs 129

5.9 Résultats d'évaluation - corpus 88milSMS 130

5.10 Exemples de traduction du corpus 88milSMS 131

6.1 Annotation d'entités nommées 136

6.2 Trois catégories d'entités nommées selon MUC (1995) 140

6.3 État de l'art des systèmes à base de règles 156

6.4 État de l'art des systèmes à apprentissage - suite 157

6.5 État de l'art sur les systèmes systèmes hybrides - suite 157

6.6 Typologie d'entités nommées SMS 163

6.7 Typologie d'entités nommées SMS - suite 164

7.1 Les types du corpus d'expérimentation 169

7.2 Caractéristiques du corpus de référence 170

7.3 Résultats globaux de trois systèmes 182

7.4 Écart des résultats de F-mesure obtenus sur la transcription manuelle 183

7.5 Exemples d'annotation par les trois systèmes 183

C.1 100 tokens plus fréquents du corpus 241

D.1 Typologie modifiée de l'écriture SMS figurant dans Panckhurst 2009, 2015 . . . 244

LISTE DES TABLEAUX

Table des figures

2.1 Modèle de la communication interactive bidirectionnelle 29

3.1 Affiche du projet alpes4science (http ://www.sms4science.org/ ?q=en) 41

3.2 Méthodologie de collecte 43

3.3 Interface d'anonymisation, figure tirée de Chabert *et al.* (2012) 46

3.4 Interface d'anonymisation, figure tirée de Chabert *et al.* (2012) 47

3.5 Interface de transcription, figure tirée de Chabert *et al.* (2012) 50

4.1 Répartition du sexe dans les projets 64

4.2 Répartition des sujets selon l'âge dans le corpus 64

4.3 Répartition des sujets selon le niveau d'études 65

4.4 Répartition géographique des sujets selon leur lieu d'origine 66

4.5 Rapport Type/Token et Type/Token standardisé en fonction de l'âge 77

4.6 Rapport Type/Token Corrigé et Moyenne Mobile du Rapport de Types/Tokens en fonction de l'âge 78

4.7 Rapport Type/Token Corrigé et Moyenne Mobile du Rapport de Types/Tokens en fonction du niveau d'études 80

4.8 Catégories grammaticales dans le corpus 80

4.9 Clavier azerty 82

4.10 Clavier alpha-numérique 82

4.11 Longueur moyenne d'unités lexicales par type de clavier pour les messages transcrits et bruts 82

5.1 Phase de traitement 102

5.2 Graphe principal de reconnaissance 104

5.3 Graphe de reconnaissance d'émoticônes 105

5.4 Graphe avec mode morphologique 106

5.5 Graphes de transitions de formes avec apostrophe 107

5.6 Graphe de transitions pour les articles *le* et *la* 108

5.7 Graphe de reconnaissance du caractère + ... 108

5.8 Graphe de transitions pour la forme *à* 109

5.9 Graphe de reconnaissances d'unités contenant des formes consécutives 110

5.10 Les huit modules du système Apertium dans Ramırez-Sánchez *et al.* (2006) . . 112

5.11 Exemple de flexion "abeille" dans le dictionnaire monolingue *fra.dix* 113

5.12 Entrées du dictionnaire monolingue *fra.dix* 114

5.13 Entrées du dictionnaire bilingue *fra-es.dix* 115

5.14 Illustration d'alignement mot-à-mot 117

5.15 Schéma d'induction du dictionnaire bilingue 117

5.16 Exemple de normalisation avec ambiguïté 119

5.17 Chaîne de traitement Apertium 120

6.1 Graphe hiérarchique de types définies pour le système QALC dans Poibeau (2005) 141

6.2 La notion d'entité nommée dans Daille *et al.* (2000) 143

6.3 Exemple de message tweet contenant un emoji sur Barbieri *et al.* (2016) 165

6.4 Typologie d'entités nommées pour les emoji 165

7.1 Extrait de concordance en mode "merge" de normalisation de casse sur Unitex 172

7.2 Graphique générale de performance de systèmes sur l'ensemble d'entités reconnues 179

7.3 Capture d'écran de message d'erreur du service *Nero* pour l'entité nommée *Agnès* 180

7.4 Capture d'écran de message d'erreur du service *Nero* 181

C.1 Taux en (%) des formes verbales 239

C.2 Taux en (%) des formes pronominales 240

TABLE DES FIGURES

E.1 Graphe d'expressions 246