Travaux neuchâtelois de linguistique, 2015, 63, 71-86

Classification des items inconnus de 88milSMS: aide à l'identification automatique de la créativité scripturale

Cédric LOPEZ[1], Mathieu ROCHE[2 3] & Rachel PANCKHURST[4]
[1]R&D, Viseo, Grenoble
[2]UMR TETIS, Cirad, Irstea, AgroParisTech, Montpellier
[3]LIRMM, UMR 5506 CNRS & Université Montpellier
[4]Praxiling, UMR 5267 CNRS & Université Paul-Valéry Montpellier

The *sud4science LR* project[1] aimed at studying a fairly recent form of written communication: SMS (Short Message Service). The first step of the project was to collect a large number of text messages from the general public. We initially gathered 93'085 SMS and our final corpus, entitled *88milSMS*, contains over 88'000 SMS.[2] In this article, we propose a novel approach (which is also applicable to other textual data) for classifying unknown items in *88milSMS*, based on two steps: 1) Classification of SMS in relation to 5 European languages (French, Spanish, English, German, Italian), 2) Classification of unknown items according to predefined classes (schedules, items containing special character(s), number(s), words without accents, or with repeated characters, etc.). We are then able to make a distinction between the truly "original" items which are widely used compared to those that are rarely used in the corpus. Based on examples mined in the different classes, we present a preliminary analysis of the obtained resource.

1. Introduction

Les nouvelles formes de communication électronique médiée développées ces dernières années soulèvent de nombreux problèmes linguistiques complexes. Avec l'adoption des téléphones mobiles, le service de messagerie SMS (Service de Messages Succincts) s'est largement développé à travers le monde, permettant la transmission de courts messages textuels en temps réel. Au départ contraint par le nombre de caractères maximum utilisables pour la rédaction d'un SMS et par la difficulté de maniement des claviers, l'écriture SMS apparaît et se développe rapidement sur les supports de communication du Web (réseaux sociaux, fora, blogs, etc.).

L'écriture SMS (désormais eSMS[3]) se caractérise par la présence de formes scripturales très riches: squelettes consonantiques ("slt" (*salut*)), abréviations

1 Projet sud4science Languedoc-Roussillon (LR): http://www.sud4science.org/

2 http://88milsms.huma-num.fr/

3 Cf. Panckhurst (2009) pour une typologie détaillée. Comme d'autres chercheurs, nous refusons l'appellation *langage SMS,* car il ne s'agit pas d'un langage, ou d'une langue, mais bien d'une *pratique scripturale*. Nous préférons l'appellation *écriture SMS* à *écrit SMS* (Cougnon 2015), car nous nous focalisons sur l'aspect dynamique de la langue, en mouvance constante. Pour Cougnon, il s'agit de "mettre l'emphase sur le résultat d'une pratique".

sémantisées (f (*fais/feras/faisais: tu f koi?*)), apocopes ("ordi" (*ordinateur*)), aphérèses ("zou" (*bisou*)), substitutions phonétisées plus ou moins complexes ("koi" (*quoi*), "boC" (*bosser*), "2m1" (*demain*)), agglutinations ("jattends" (*j'attends*)), suppressions de fins de mots muettes ("vou" (*vous*)), répétitions de caractères ("suuuuppppeeerrrr"), ajouts de caractères ("les zamours", "oki"), binettes/emoji ("^^", ":)", ☺) — la liste est longue. Ces phénomènes simulent parfois l'oralité, l'écrit, ou possèdent leurs propres marques et caractéristiques, le tout reflétant une langue vivante dynamique en mouvance constante.

Dans le but de permettre l'étude de l'ensemble de ces phénomènes, (Panckhurst & al. 2013) ont récemment recueilli plus de 90'000 SMS dans le cadre du projet *sud4science LR*[4]. Après plusieurs étapes de prétraitement des données (vérification, épuration, etc., Panckhurst & al. 2013) et d'anonymisation (Accorsi & al. 2014; Patel & al. 2013), le corpus final contient plus de 88'000 SMS authentiques, et est publié sous le nom de *88milSMS* (Panckhurst & al. 2014)[5].

Dans cet article, nous présentons une première approche permettant de classer automatiquement des items inconnus, qui apparaissent dans *88milSMS*, en vue d'une aide à l'analyse de l'eSMS (Panckhurst & al. 2013) et plus précisément à l'identification de la *créativité scripturale*[6]. Nous définissons un item lexical comme l'unité autonome constituante du lexique de l'eSMS (au moins dans le cadre de notre corpus), compris entre deux espaces. Ainsi, "jtrouve" est considéré comme un seul item lexical, alors que "je trouve" est considéré comme une suite de deux items lexicaux.

Une telle ressource contenant tous les items lexicaux SMS "inconnus" du français standard trouve son intérêt à la fois en linguistique et en informatique. D'un point de vue linguistique, cette ressource pourra faciliter l'étude à propos de la créativité scripturale, les mots du discours, l'agglutination, etc. D'un point de vue informatique, l'utilisation de la ressource sera utilisée dans la chaîne de traitement automatique des messages de blogs, fora, SMS, et réseaux

---

4 Le projet *sud4science LR* (www.sud4science.org) s'inscrit dans un projet plus global, international, lancé en Belgique en 2004: sms4science (www.sms4science.org, Fairon & al. 2006; Cougnon 2014; Cougnon & Fairon 2014).

5 Corpus *88milSMS*: http://88milsms.huma-num.fr

6 Dans le cadre de cet article, notre terme *créativité scripturale* se veut générique et renvoie à différents phénomènes, qui questionnent encore et toujours en sciences du langage: la *néologie* (la créativité lexicale par suffixation (SMS, n° 52041: “ça se passe bien la *voituration*?”), mots-valises ("mdr j'avais une réponse bien cinglante, mais rien que de répondre, ça annule la *cinglicité* (?) de la chose...")), la *néographie* (des variantes de graphies qui constituent des "écarts ludiques" (Anis 1998: 132), qui s'éloignent de la langue standardisée et qui sont très présentes et très instables dans l'écriture SMS: abréviations, troncations, notations sémio-phonologiques ou graphies phonétisées, etc.), l'écriture *non-intentionnée* ("fautes" de saisie, etc.). Nous ne prétendons pas répondre à ces questions, notamment concernant la frontière parfois ténue entre néologie et néographie, mais nous n'avons pas besoin d'une distinction fine ici.

sociaux. Elle constitue en effet un premier pas vers le transcodage automatique de l'écriture non standard vers l'écriture standardisée (Beaufort & al. 2010) qui permettra d'améliorer la qualité des applications fondées sur un traitement automatique de l'eSMS, par exemple dans un contexte médical (Stenner & al. 2011; Vetulani & Marciniak 2011), ou de reconnaissance vocale (Bove 2005).

Dans la suite de l'article, nous commençons par identifier, de manière automatique, les items non standard (§ 2) en classant les SMS par langue (§ 2.1) puis, en se concentrant sur le classement des items issus de SMS français (§ 2.2). Enfin, nous présentons et discutons les résultats et la ressource obtenue (§ 3) avant de conclure (§ 4).

2. Identification automatique d'items originaux

L'objectif du travail est d'aider l'utilisateur à identifier les items non standard (INS) dans le corpus *88milSMS*, dans le sens où ces items n'existent dans aucun dictionnaire de langue française. Parmi les INS, nous proposons d'identifier les items non standard originaux (INSO). Un traitement manuel serait complexe principalement à cause de deux points:

- la définition d'un "item non standard original";
- la taille du corpus, supérieure à un million d'items lexicaux.

La question de l'originalité d'un item est largement discutable selon que l'on s'intéresse aux variations lexicales/scripturales, aux créations de termes, ou encore à l'alternance codique, par exemple. Afin de ne pas biaiser l'interprétation de la ressource produite et de ne pas contraindre son utilisation à une application donnée, nous avons considéré que les INSO sont des items lexicaux que nous ne sommes pas en mesure de classer de façon triviale (par exemple par horaires, pseudonymes, termes du français, etc.). L'hypothèse sous-jacente est de considérer que les items n'ayant pu être classés dans les catégories prédéfinies sont potentiellement des INSO.

La taille du corpus nous incite à proposer un traitement automatique. L'identification des INS revient donc à un problème de classification selon des catégories prédéfinies. Même si ce traitement ne permet pas de dresser une liste exhaustive des classes d'INS du corpus, celui-ci a tout de même le mérite d'aider l'utilisateur dans la tâche d'identification. Dans la suite, nous décrivons les deux étapes principales de notre approche: le classement des SMS par langue (§ 2.1), puis la définition des filtres pour le classement des items en français (§ 2.2).

2.1 Classement des SMS par langue

Le classement des SMS par langue est indispensable pour assurer que la ressource produite en sortie de l'étape de classification des items ne contient pas de biais dû aux différentes langues utilisées dans *88milSMS*. Par exemple, le SMS "see you soon" doit être classé dans la catégorie "anglais", sans quoi chacun des items le composant serait confronté à des critères définis *a priori* pour le français. Nous avons restreint la classification au français et à 4 langues officielles de pays limitrophes à la France. Au total, nous considérons donc 5 langues: français, anglais, espagnol, allemand, italien. À ces 5 classes, nous ajoutons 4 classes mixtes qui permettent de tenir compte de la présence de deux langues[7] dans un même SMS: français-anglais, français-espagnol, français-allemand, français-italien. Ainsi, notre système doit être en mesure de classer les SMS selon 9 classes. Notons que d'autres langues apparaissent dans le corpus (notamment le shimaoré/mahorais et l'arabe)[8].

L'approche largement adoptée pour la classification de textes par langue repose sur l'identification des *n*-grammes de caractères, suites de *n* caractères (Cavnar & Trenkle 1994) caractéristiques des langues (par exemple, les trigrammes "the" et "les" sont respectivement associés à l'anglais et au français). À partir de listes établies pour chaque langue, il est alors possible de prédire la langue d'un texte. Dans notre contexte, nous nous intéressons à l'étude des spécificités scripturales des SMS. Ainsi, les *n*-grammes "classiques" risquent de se révéler peu pertinents étant donné la présence d'inventions langagières, d'abréviations, d'erreurs d'accentuation, de variantes d'orthographe propres à l'eSMS (par exemple, les *n*-grammes extraits à partir du SMS "C como kon fai pr ls k do" seront tout à fait inappropriés). Par ailleurs, la découverte de ces *n*-grammes caractéristiques sera étudiée, d'une certaine mesure, dans la suite du processus (voir § 2.2).

D'autres approches s'appuient sur l'apprentissage automatique pour construire un modèle permettant de classer les textes selon chaque langue apprise (Vo-Trung 2004; Okanohara & Tsujii 2009). Cette approche nécessite au préalable d'avoir un jeu de données annoté. De tels jeux de données sont disponibles pour les textes homogènes, créés la plupart du temps en utilisant les métadonnées décrivant la langue utilisée (par exemple, dans les articles journalistiques, pages web, ou dans les tweets), évitant ainsi une lourde tâche d'annotation manuelle. Mais la tâche est plus complexe lorsque l'on souhaite

---

[7] D'autres classes sont représentées dans *88milSMS*. Nous avons par exemple rencontré des SMS contenant plus de 3 langues (sms n° 41015: *Hello ! Que tal? Yes, a bientôt*); voir également n° 32957 et n° 78099). Dans le cadre de cet article, nous nous focalisons sur les classes les mieux représentées.

[8] Notons que les SMS écrits en shimaoré/mahorais ont fait l'objet d'une annotation manuelle lors d'une étude réalisée par des étudiants stagiaires. Au total, 335 SMS ont été classés ainsi et ne constituent pas ici un objectif de classement automatique.

annoter un jeu de données constitué de SMS. D'une part, à notre connaissance, il n'existe aucun corpus de SMS annotés par langue[9], et d'autre part, nous considérons les SMS comme des textes hétérogènes, i.e. contenant parfois des mots de plusieurs langues (par exemple, "peut dormir at home?", SMS n° 928, doit être classé dans une même classe "français et anglais").

Démunis d'un jeu de données annoté pour les SMS[10] et souhaitant éviter d'utiliser des *n*-grammes de caractères, nous avons choisi une approche de classement fondée sur lexique. Pour constituer notre lexique, nous avons utilisé la liste des mots les plus fréquents utilisés pour l'anglais, l'espagnol, l'allemand, et l'italien (500 à 1'000 mots par langue), selon Wiktionary[11].

Nous avons ensuite comparé les termes des 4 listes entre eux, ainsi qu'au plus grand lexique des formes fléchies du français disponible au format électronique, le *Lexique Électronique des Formes Fléchies du Français* (LEFFF, cf. (Sagot 2010)). Les termes présents dans plusieurs listes ont été éliminés. De même, les termes présents à la fois dans une liste et dans le LEFFF ont été éliminés, puisqu'ils ne peuvent être considérés comme étant spécifiques à la langue en question. L'ensemble de ces listes constituent ainsi notre modèle. Dans la suite de l'article, nous nommerons "descripteurs" les mots spécifiques à une langue.

Afin de considérer la spécificité de notre corpus, notre expertise nous a permis d'ajouter manuellement des descripteurs à notre modèle. Par exemple, nous savons que "bjr" (*bonjour*), "slt" (*salut*) ou "jtm" (*je t'aime*) sont des descripteurs pertinents pour les SMS français.

Nous obtenons ainsi un lexique contrôlé, spécifique à chaque langue, adapté aux SMS: 88 descripteurs pour l'anglais, 55 pour l'espagnol, 62 pour l'allemand, 83 pour le français, et 45 pour l'italien.

À chaque SMS est attribué un score pour chaque langue. Ce score correspond à la somme des occurrences de chaque descripteur dans un lexique donné (par exemple, FR=4 signifie que 4 descripteurs du français ont été identifiés automatiquement dans un SMS). Les résultats sont structurés au format XML, annotés avec les balises FR, EN, SP, DE, et IT indiquant chaque langue considérée.

Par exemple, dans un SMS contenant "at home", "at" est un mot fréquent en anglais; en revanche, "home" ne figure pas dans la liste des 500 à 1'000 mots fréquents en anglais. Comme "at" n'apparaît ni dans le LEFFF ni en tant que

---

9 Nous avons très récemment eu connaissance d'un corpus suisse annoté par langue (voir Stark & al. 2009-2014; Cathomas & al. ce volume).

10 Nous projetons de constituer un corpus d'apprentissage hétérogène à partir de différents corpus de SMS de langues différentes, de la même façon que (Lui & al. 2014).

11 http://www.wiktionary.org/, voir "listes de fréquences".

descripteur des autres langues considérées, il sera classé parmi les 88 descripteurs ayant été retenus pour l'anglais, alors que "home" sera rejeté.

Formellement, notre algorithme de classement est le suivant:

Soit S un SMS.

Soit $L \in$ *{français, anglais, espagnol, italien, allemand}.*

Soit *Ns,l* le nombre de descripteurs de S pour la langue $l \in L$.

Soit *m(S)* le nombre maximum de descripteurs de *l* trouvés pour S.

1. Si aucun descripteur n'a été trouvé, on classe le SMS par défaut en français.
2. Pour chaque *Ns,l = 1 = m(S)* on classe S dans la classe *l*
3. Pour chaque *Ns,l > 1*, on classe le SMS dans la/les langue(s) *l* correspondante(s).

Les trois étapes de l'algorithme correspondent aux trois cas de la fig. 1, ci-dessous.

Exemples	Format XML
Cas 1 • français	<sms id="64158" text="bon courage" EN="0" SP="0" DE="0" IT="0" FR="0" />
Cas 2 • français et anglais	<sms id="928" text="<PRE_5> **peut** dormir **at** home?" FR="1" EN="1" />
Cas 3.1 • français	<sms id="424" sms="Hey **yo** ! Viens quand tu veux péqufier mon appart , **ce** sera ac grand plaisir ! Enfin **le** week prochain je rentre sr Tlse , ms sinon ceux **qui** suivent ss problème ! **du** coup **on se** voit la semaine prochaine ! Bisous vieille grognasse" SP="1" FR="6" />
Cas 3.2 • français et anglais	<sms id="1648" sms="Seriously?? ! ! Wow such a coincidence ! ! **She** ll probably give **you** a 100 % **yes** j en **suis** sure ! ! I ' m watching csi :) j **ai** cours a 8h demain :/ FML tu commence **quand le** boulot? <3" EN="3" FR="4" />

Fig. 1: Exemples de SMS classés par langue(s).

De façon générale, de nombreux termes spécifiques à une langue peuvent également être des créations scripturales dans le cadre des SMS. Par exemple, "el" est à la fois le déterminant espagnol et une contraction du pronom français "elle". Dans notre système, ces cas peuvent contribuer à une classification erronée, et montrent donc une limite à notre approche (cf. § 3 pour une évaluation de notre approche).

Notre classification des SMS par langue a montré que le corpus *88milSMS* était composé d'une très grande majorité de SMS en français (de l'ordre de 97% du vocabulaire utilisé est spécifique au français). Ces SMS en français sont alors retenus et exploités pour la prochaine étape décrite en section 2.2.

2.2 Classement des items

Notre approche d'identification d'INS consiste à fournir le corpus *88milSMS* en entrée du système et à obtenir en fin de traitement un ensemble de classes permettant d'aider à l'identification automatique de la créativité scripturale. Le système est développé en Java.

Le corpus *88milSMS* est d'abord segmenté. Les segments habituellement considérés dans les approches de classification sont les mots ou les phrases. La segmentation par phrases n'est pas pertinente ici car notre objectif est d'identifier un ensemble d'items. Aussi, dans le contexte de la segmentation de SMS, il ne semble pas pertinent de prendre le mot comme segment puisque le lexique utilisé pour la rédaction de ces messages n'est pas formellement défini. De plus, il est complexe d'identifier automatiquement les frontières des mots au sein d'une chaîne de caractères issue de données textuelles de type SMS (par exemple, "a2min lami" = "à demain l'ami"). Notre objectif étant d'identifier des items lexicaux non standard, nous considérerons donc qu'un segment, ou item lexical, est une suite de caractères compris entre deux espaces (dans l'exemple précédent nous obtenons ainsi deux segments: "a2min" et "lami"). Notons qu'un prétraitement a consisté à ajouter une espace avant et après chaque élément de ponctuation lorsque ce dernier était absent. Au total, nous obtenons ainsi plus d'un million d'items lexicaux.

Notre approche consiste à déterminer, dans un premier temps, trois ensembles distincts, que nous nommerons "classe": C1, C2 et C3.

La classe C1 recevra les items standard, c'est-à-dire les items présents dans le LEFFF, avec et sans accents. C2 recevra les INS reconnus grâce à des filtres que nous définissons ci-après, et C3 recevra tous les items qui n'ont été classés ni dans C1 ni C2 et qui peuvent donc correspondre à une forme de créativité scripturale non retenue par nos filtres. Les classes C1, C2, et C3 sont disjointes, i.e. un même item ne peut apparaître dans deux classes différentes.

L'objectif du travail étant d'identifier automatiquement les INS pour le français, nous cherchons, en premier lieu, à élaguer l'ensemble des items de *88milSMS* qui seraient également présents dans le LEFFF. Ainsi, la classe C1 contenant les items standard est constituée de deux sous-classes:

- ***C1.1: items standard présents dans le LEFFF.***
Le filtre consiste ici à comparer un à un les items français de *88milSMS* avec les items du LEFFF. Les items présents dans le LEFFF sont attribués à la classe C1.1.

- ***C1.2: items standard présents dans le LEFFF sans accents.***
Nous mettons en place un filtre permettant de comparer les items avec les mots du LEFFF auxquels nous avons supprimé les accents. La classe C1.2 accueille donc les items correctement orthographiés selon les normes du français standard mais dont l'accentuation est absente (par exemple: *qualites*, *degat*, *precisions*, *europeen*).

Cette première étape a permis de construire une approche automatique qui catégorise les items standard. Dans la suite, nous proposons une sous-catégorisation des items non standard (INS) de C2 (items non contenus dans C1):

- C2.1: items composés d'un caractère unique. Cette sous-classe contient les items constitués d'un seul caractère, incluant les caractères spéciaux, les chiffres, lettres, etc. Une telle classe est par exemple utile pour l'étude des abréviations sémantisées telles que *c* pour *c'est/ces/ce…* ou *t* pour *t'es/tu….*

- C2.2: items assimilables à des horaires. Cette sous-classe contient les items représentant une heure, ou plus généralement un rapport avec le temps. Le filtre correspondant est une succession de tests recherchant la présence d'une suite de caractères spécifiques telle qu'un chiffre suivi de la lettre "h" ou des lettres "min". Par exemple, nous identifions *12h30*, *23:56*, *8heures*, *10minaperdre*, *6-7h,* etc.

- C2.3: item avec allongement. Les termes de cette sous-classe ont subi une répétition de caractères, qui simule un allongement vocalique, et ce sur au moins un caractère (par exemple: *Jarriiiiiiiiive*, *Huuuummm*, *Meeerciiiiiii*, *tkkkkt*). Le filtre mis en œuvre compare chaque caractère avec le caractère suivant. Si plus de deux caractères sont répétés, l'item est classé dans C2.3. Rappelons que les mots possédant deux mêmes caractères consécutifs issus de la langue standard (par exemple: *passe*, *embrasse*, *apprendre*) ont précédemment été classés dans C1. Si nous considérons qu'un allongement est un critère répondant à l'originalité des items, alors il faut considérer que C2.3 contient bon nombre d'INSO.

- C2.4: item avec caractère spécial. Nous testons ici la présence d'un caractère spécial dans chaque item. Les caractères spéciaux considérés sont tous les caractères d'un clavier alphanumérique AZERTY classique, hors chiffres et lettres (30 caractères spéciaux au total). Les items de la classe C2.4 contiennent au moins un caractère spécial (par exemple: *Conn*rd*, *resto+cine*, *appeler/texto*, *dés~annule*, *thèse/antithès/synthèse*, *fish&chips*). Cette sous-classe contient les binettes contenant un ou plusieurs caractères spéciaux (par exemple: ^^ ou ;)).

- C2.5: présence d'un chiffre. La sous-classe C2.5 contient tous les items incluant un chiffre (qui n'ont pas été précédemment repérés, par exemple, dans C1). Le filtre correspondant teste simplement la présence d'au moins un chiffre au sein de l'item. Nous obtenons par exemple, *numb3rs*, *mc2*, *106ounette*, *3615ma-vie*, *Ar5ggggggggh*. Ces items peuvent dès lors être considérés comme des INSO dont l'originalité réside dans la présence de chiffres.

- C2.6: binettes[12]. Cette sous-classe contient les binettes d'après une liste construite en deux temps: 1) binettes acquises sur le Web[13], 2) binettes ajoutées manuellement d'après notre expertise sur le corpus. Les binettes inconnues de notre liste pourront être découvertes dans la sous-classe C2.4. Au total, nous disposons d'une liste de 54 variantes de binettes (par exemple: ":-)" et "+.+"). Cette liste ne contient aucun 'emoji' (émoticône graphique).

Enfin, la classe C3 contient les items qui ne sont ni dans C1 ni dans C2.

- C3: items non standard originaux (INSO). Cette catégorie contient les items qui n'ont pas été classés dans les catégories précédentes et ne nécessite donc pas la mise en place de filtre spécifique. Nous obtenons ainsi des items néologiques tels que *cinglicité*, *voituration* ou encore des items néographiques agglutinés tels que *tatende*, *tetrangle*. Les items présents dans C3 sont donc potentiellement des items non standard originaux (INSO).

Il est important de noter que les sous-classes de C2.1 à C2.6 ne sont pas disjointes: plusieurs sous-classes peuvent contenir un même item. Par exemple, *Ar5ggggggggh* doit être classé dans C2.3 et C2.4.

Nous avons défini 3 classes et 8 sous-classes qui représentent l'ensemble des items présents dans les SMS "français" identifiés à l'étape précédente (§ 2.1). D'autres classes et sous-classes peuvent être ajoutées dans le but de classer plus finement les items en fonction des objectifs visés.

Dans la suite, nous évaluons l'approche de classement des SMS par langue, et nous analysons les classes générées automatiquement d'un point de vue quantitatif et qualitatif afin de mettre en relief la pertinence de nos propositions et le type d'items non standard originaux identifiables.

---

12 'Binette' est le terme (québécois) que nous utilisons pour évoquer 'smiley', 'émoticône', 'frimousse', par exemple: ";)", "^^", ";)", ":D", etc. Dans un travail ultérieur, nous effectuerons le classement des 'emoji' (les binettes graphiques) qui nécessitent un repérage Unicode.

13 Notamment https://support.skype.com/fr/faq/FA12330/qu-est-ce-que-la-liste-complete-d-emoticones

3. Evaluations et analyse

Dans un premier temps, nous évaluons notre algorithme d'identification de la langue (§ 3.1) puis nous discutons les INSO extraits au regard des différentes sous-classes que nous proposons (§ 3.2).

3.1 Classification par langue

Cette section a pour objectif de présenter la qualité des classes générées par notre système. Nous avons appliqué notre algorithme de classification automatique sur l'ensemble du corpus *88milSMS*. Le protocole consiste à évaluer manuellement tous les SMS classés par notre système. Cependant, devant l'ampleur de la tâche qui exige la lecture attentionnée de chaque SMS, nous avons limité la taille de la classe "français" à 500 SMS. Au total, 1'329 SMS ont été évalués manuellement.

La fig. 2 présente les résultats de classification par rapport aux classifications réelles issues des experts. Un exemple de lecture est le suivant: Le système a automatiquement attribué 551 SMS à la classe "anglais" (EN), parmi lesquels 476 sont effectivement écrits en anglais, 63 sont écrits à la fois en anglais et en français et 12 sont en français. La fig. 2 permet de mettre en avant le taux d'exactitude, c'est-à-dire le taux de SMS bien classés qui s'élève à 89%. Ainsi, les résultats montrent que le système permet de générer des classes pertinentes, bien que l'on doive noter une faiblesse pour le classement des SMS italiens et allemands, mal représentés dans notre corpus.

		Classes réelles								
		FR	EN	SP	IT	DE	FR-EN	FR-SP	FR-IT	FR-DE
Classes prédites	FR	494	0	1	0	0	5	0	0	0
	EN	12	476	0	0	0	63	0	0	0
	SP	1	0	25	0	0	0	6	0	0
	IT	5	0	0	9	0	0	0	2	0
	DE	2	0	0	0	2	0	0	0	0
	FR-EN	21	4	0	0	0	169	0	0	0
	FR-SP	0	0	3	0	0	0	7	0	0
	FR-IT	8	0	0	11	0	0	0	1	0
	FR-DE	2	0	0	0	0	0	0	0	0

Fig. 2: Matrice de contingence des résultats du classement des langues.

Finalement, l'évaluation montre que les classes générées sont satisfaisantes en vue d'une exploitation dans la suite du processus (classification des items).

3.2 Analyse des résultats et ressource obtenue

Le traitement décrit dans § 2.2 permet d'obtenir une ressource facilitant l'exploration des items lexicaux originaux dans le corpus *88milSMS*. La ressource obtenue est divisée en 11 fichiers au format XML. Chaque fichier correspond à une classe ou sous-classe, et associe à chaque item son nombre d'occurrences dans le corpus. La fig. 3 présente une synthèse de la ressource générée.

Nous observons un nombre d'items différents compris entre 83 (pour C2.1) et 22'211 (pour C1.1).

L'item mono caractère (C2.1) le plus utilisé apparaît 74'922 fois. Il s'agit du point qui semble indiquer une volonté de segmenter les syntagmes. Notons que le point n'est pas utilisé dans les binettes.

L'item le plus fréquent dans C2.2 est "19h", utilisé 172 fois dans notre corpus. Le nombre important d'items différents attribués à la classe C2.2 (500 items) tend à montrer l'utilisation des SMS comme un outil de prise de rendez-vous.

Dans la classe C2.3 contenant des items présentant une répétition de caractères, les plus fréquents sont "Mdrrr", "mdrrr", "Mdrrrr", "Biisoux", et "Loool", utilisés entre 66 et 741 fois. Le maximum de répétitions observées sur un item s'élève à 171 "a" successifs (dans "Aaaa[…]aaah", SMS n° 348).

Les caractères spéciaux sont le plus fréquemment utilisés dans la ponctuation et construction de binettes, tel que le montrent les items classés dans C2.4. Parmi les mieux représentés, nous trouvons le point et l'apostrophe ainsi que les binettes " :)", "^^", ":D", " :p", utilisés entre 1'382 et 74'922 fois. D'autres formes d'utilisation apparaissent, comme la censure dans "pu+in", "Conn*rd", la valorisation d'un item dans "*ironie*", ou le remplacement d'un item par un symbole équivalent phonétique, par exemple "La+Belle".

1'971 items forment la classe C2.5. Les plus fréquents sont les chiffres et nombres, avec un maximum atteint pour le chiffre *2,* ce qui s'explique par son transcodage en la préposition "de". Viennent ensuite les expressions de positionnement telles que "2eme", "3eme", les expressions monétaires telles que l'item "5€" qui apparait 9 fois, ainsi que d'autres items tels que "mp3" pour le support numérique ou encore "w9" pour la chaîne télévisée.

La classe C3 contient 17'891 items. Comme nous l'avons vu, celle-ci contient les items qui n'ont pas été classés dans les catégories précédentes. Afin de donner un aperçu de la qualité globale du contenu de C3, nous mentionnons ci-après 20 INSO parmi ceux qui ont le plus faible nombre d'occurrences (1 occurrence) et les 20 INSO qui ont le plus grand nombre d'occurrences (entre 219 et 3'341 occurrences):

- 1 occurrence: *zeit*, *Fbk*, *evenemens*, *mavei*, *GDF*, *haceler*, *parcequel*, *souvenai*, *svpm*, *lattendais*, *Hmhm*, *Jmendors*, *kmbien*, *vaho*, *Bàh*, *tavécour*, *Alellujah*, *estt*, *estr*, *alongei*
- Entre 214 et 3'341 occurrences: *Jai (214)*, *Lool (219)*, *avc (229)*, *Tkt (237)*, *Jte (246)*, *Pk (248)*, *weekend (253)*, *Beh (263)*, *Dsl (264)*, *cest (283)*, *LOL (300)*, *méme (306)*, *Parce (347)*, *jsuis (425)*, *çà (459)*, *week (552)*, *jte (608)*, *tkt (688)*, *parce (1'009)*, *lol (3'341)*

Les items les plus fréquents sont les abrègements morpho-lexicaux (acronymes/sigles: "lol", "mdr", etc.), puis des squelettes consonantiques avec substitution tels que "tkt" (688 occurrences, pour "t'inquiète"), des réductions en agglutinations "jte" (608 occurrences, pour "je te"), des apocopes ("ordi" 313 occurrences), puis des onomatopées comme "Beh" (263 occurrences) ou encore des substitutions graphiques comme "mwa" (108 occurrences, pour "moi"), etc. (Panckhurst 2009). On note également la présence d'items appartenant à d'autres langues (provenant d'un classement erroné à l'étape de l'identification de la langue), ou plus généralement d'items absents du LEFFF qui ne peuvent en aucun cas être considérés comme des INSO (*GDF*, *Alellujah*…).

La fig. 3 montre la répartition du nombre d'items par classe et met en évidence le nombre élevé d'items dans C3 (17'891 items). Nous remarquons que les items de C3 peuvent se distinguer en 3 sous-classes:

- **items non standard largement utilisés par les utilisateurs.** Il s'agit des items qui apparaissent les plus fréquemment, tels que les acronymes susmentionnés et certains squelettes consonantiques. Ces items peuvent être repérés par le nombre d'occurrences élevé des items de C3. Nous pouvons considérer que ces INS sont adoptés par les utilisateurs. On peut dès lors se poser la question de leur originalité.
- **items non standard rarement utilisés (qui apparaissent rarement dans notre corpus).** Nous pouvons proposer deux sous-classes pour ces items qui correspondent:
    - soit à une écriture non intentionnée (erreur de touche sur le clavier, par exemple) comme cela semble être le cas dans "dimache", "qiue", "szrviette" ou encore "confortablemenr". Ces items apparaissent une seule fois dans le corpus. En effet, les items originaux ayant une graphie non intentionnée se traduisent par un nombre d'occurrences faible dans le corpus, s'expliquant par le fait qu'une même erreur sur un même item est peu probable.
    - soit à une réelle volonté de la part de l'utilisateur de créer un item original. Nous les distinguons des items de la sous-classe précédemment décrite en ce sens qu'ils sont réellement originaux, c'est-à-dire que la graphie ne semble pas due à une erreur de

> frappe. C'est le cas par exemple de "Keskitariv", "coa" "juskau", "fmlle", qui apparaissent une seule fois dans le corpus. Lorsque ces items apparaissent plus d'une fois dans le corpus, mais avec un nombre d'occurrences toutefois faible (de l'ordre d'une dizaine dans notre corpus), on peut considérer qu'il ne s'agit ni d'inventions de la part de l'utilisateur, ni de graphies largement adoptées par la communauté. Cette proposition est renforcée par l'observation de nombreuses graphies pour un même mot (par exemple pour "coa": "koi", "qoi", "qua", etc.). Nous pourrions qualifier certains items de "naissants" afin d'émettre l'idée qu'il s'agit d'un item qui pourrait être adopté prochainement par une plus large partie des "textoteurs". Cela pourrait être le cas par exemple de "depech" (2 fois), "sapel" (2 fois), "mkini" (4 fois), "kun" (4 fois), "staprem" (5 fois), "komen" (5 fois), "moua" (5 fois), ou "kome" (6 fois).

Les constatations précédentes nous amènent à évoquer la question de l'adoption d'un item dans le vocabulaire SMS (en considérant que ce vocabulaire correspond à l'ensemble des graphies utilisables pour la rédaction d'un SMS). Nous avons repris l'ensemble des items appartenant aux classes C2.1 à C2.6 ainsi que C3 (c'est-à-dire tous les items de notre corpus *88milSMS* absents du LEFFF). En représentant leur nombre d'occurrences (fréquence) normalisé sur un graphe (non représenté dans cet article), on constate qu'une majeure partie des INS employés dans *88milSMS* constituent la longue traîne (entre 0 et 1%), et que de très rares items originaux sont plus fréquemment employés (seulement 8 INS ont une fréquence supérieure à 1%) avec une fréquence maximum de 4,05% pour la binette "^^".

Le vocabulaire spécifique aux SMS ne respecte donc pas la loi de Zipf (observation empirique concernant la fréquence des mots dans un texte). Or, la loi de Zipf est valide quelle que soit la langue sur laquelle elle est appliquée. Quels que soient les sujets ou les auteurs, elle présente toujours la même allure. Aujourd'hui, les INS ne sont donc pas suffisants pour composer un vocabulaire complet. La créativité lexicale ne semble donc pas *réduire le vocabulaire en rassemblant derrière un simple mot une multitude de significations* (Bully 1969), comme cela est interprété dans le cas de la loi de Zipf, mais au contraire d'augmenter le vocabulaire.

Parmi la multitude d'INS proposés par les utilisateurs, certains seront ainsi adoptés, voire figés, par la communauté (par exemple "tkt" utilisé 908 fois dans notre corpus).

Catégorie	Nombre d'items différents	Item ayant le nombre d'occurrences maximum pour chaque classe
C1.1	22'211	*de* (24'908)
C1.2	2'530	*a* (16'017)
C2.1	83	*.* (74'922)
C2.2	500	*19h* (172)
C2.3	2'537	*Mdrrr* (741)
C2.4	481	*.* (74'922)
C2.5	1'971	*<3* (1'540)
C2.6	54	*:)* (6'704)
C3	17'891	*lol* (3'341)

Fig. 3 : Description synthétique de la ressource générée.

4. Conclusion

Dans le but d'aider à l'identification de la créativité lexicale au sein du corpus *88milSMS*, nous avons proposé, dans cet article, une approche automatique de classification des items inconnus. L'application, développée en Java, a permis de générer une ressource constituée de 11 classes à laquelle l'utilisateur peut accéder selon ses propres objectifs de recherche (étude de la créativité lexicale, de l'agglutination, etc.). Notons que de nouvelles classes peuvent être ajoutées facilement.

La mise à disposition d'une telle ressource est fondamentale dès lors que l'on s'intéresse à des tâches telles que le transcodage automatique de SMS en français standardisé. Plus généralement, les outils linguistiques mis à disposition de la communauté du Traitement Automatique du Langage Naturel sont peu robustes lorsqu'il s'agit d'analyser du texte brut issu des réseaux sociaux, tels que Twitter, Facebook, ou encore des blogs ou fora. Dans ce contexte, un prétraitement visant à transcoder l'écrit non standard vers une écriture standard nécessite de nombreuses ressources telles que celle que nous proposons dans nos travaux.

À moyen terme, il sera intéressant de mener une étude comparative des résultats obtenus à partir du corpus *88milSMS* avec d'autres corpus de SMS, mais aussi avec d'autres types de corpus comme des données en français issues de Twitter (Bouillot & al. 2012).

Par ailleurs, notre prochain travail consistera à transcoder semi automatiquement les INS découverts dans le corpus *88milSMS*: une partie automatique consistera à aligner les items de notre ressource (nous

évaluerons en particulier l'apport de cette ressource à notre méthode d'alignement (Lopez & al. 2014) avec des items déjà présents (et transcodés) dans d'autres ressources); une partie manuelle sera nécessaire afin de transcoder les INS n'ayant pas été référencés par ailleurs. Ensuite, il sera envisageable d'utiliser notre ressource transcodée manuellement en tant que corpus d'apprentissage afin de la fournir en entrée d'un système d'apprentissage automatique. Ceci pourrait représenter une aide précieuse pour l'identification de certaines règles et permettre ainsi un transcodage automatique des INS à venir.

Bibliographie

Accorsi, P., Patel, N., Lopez, C., Panckhurst, R. & Roche, M. (2014): Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques. In: Cougnon, L.-A. & Fairon, C. (éds.): SMS Communication. A linguistic approach. Amsterdam (John Benjamins), 11-28.

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

Beaufort, R., Roekhaut, S., Cougnon, L.-A. & Fairon, C. (2010): A hybrid rule/model-based finite-state framework for normalizing SMS messages. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 770-779.

Bouillot, F., Poncelet, P., Roche, M., Ienco, D., Bigdeli, E. & Matwin, S. (2012): French Presidential Elections: What are the Most Efficient Measures for Tweets? In: Proceedings of Politics, Elections and Data Workshop (PLEAD'2012). CIKM Workshop, ACM, 23-30, Maui, USA, 23-30.

Bove, R. (2005): Étude de quelques problèmes de phonétisation dans un système de synthèse de la parole à partir de SMS. Actes de RECITAL 2005, 625-634.

Bully, P. (1969): Zipf, créateur de la linguistique statistique. In: Communication et langages, 2 (1), 23-28.

Cavnar, W. B. & Trenkle, J. M. (1994): N-gram-based text categorization. In: Ann Arbor MI, 48113 (2), 161-175.

Cougnon, L.-A. (2015): Langage et sms. Une étude internationale des pratiques actuelles. Cahiers du CENTAL, 8. Louvain-la-Neuve (Presses universitaires de Louvain).

Cougnon, L.-A. & Fairon, C. (éds.) (2014): SMS Communication. A linguistic approach. Amsterdam (John Benjamins).

Fairon, C., Klein, J.-R. & Paumier, S. (2006): SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation. Louvain-la-Neuve (Presses universitaires de Louvain). Manuel+CD-Rom. Disponible: http://www.smspourlascience.be/ (28.6.2015)

Lopez, C., Bestandji, R., Roche, M. & Panckhurst, R. (2014): Towards Electronic SMS Dictionary Construction: An Alignment-based Approach. In: Proceedings LREC, Reykjavik, Iceland, 26-31 May, 2833-2838.

Lui, M., Lau, J. H. & Baldwin, T. (2014): Automatic detection and language identification of multilingual documents. In: Transactions of the Association for Computational Linguistics, 2, 27-40.

Okanohara, D. & Tsujii, J. (2009): Text Categorization with All Substring Features. In: Proceedings of the 2009 SIAM International Conference on Data Mining, 838-846.

Panckhurst, R. (2009): Short Message Service (SMS): typologie et problématiques futures. In Arnavielle, T. (coord.): Polyphonies, pour Michelle Lanvin, Université Paul-Valéry Montpellier 3, 33-52.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M. & Verine, B. (2013): Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS. In : Épistémè – revue internationale de sciences sociales appliquées, 9: Des usages numériques aux pratiques scripturales électroniques, 107-138. Disponible: https://hal.archives-ouvertes.fr/file/index /docid/923618/filename/panckhurst_detrie_lopez_moise_roche_verine_v16.pdf (28.6.2015)

— (2014): 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.

Patel, N., Accorsi, P., Inkpen, D., Lopez, C. & Roche, M. (2013): Approaches of anonymisation of an SMS corpus. In: Proceedings of CICLING (Conference on Intelligent Text Processing and Computational Linguistics), LNCS, Springer Verlag, March 24–30, 2013, University of the Aegean, Samos, Greece, 77-88.

Sagot, B. (2010): The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In: Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta. Disponible: http://hal.inria.fr/inria-00521242/ (28.6.2015), 2744-2751.

Stark, E., Ueberwasser, S. & Ruef, B. (2009-2014): Swiss SMS Corpus. University of Zurich. Disponible: https://sms.linguistik.uzh.ch (1.7.2015)

Stenner, S. P., Johnson, K. B. & Denny, J. C. (2011): PASTE: patient-centered SMS text tagging in a medication management system. In: Journal of the American Medical Informatics Association, 19 (3), 368-374.

Vetulani, Z. & Marciniak, J. (2011): Natural language based communication between human users and the emergency center: POLINT-112-SMS. In: Human Language Technology. Challenges for Computer Science and Linguistics. Berlin/Heidelberg (Springer), 303-314.

Vo-Trung, H. (2004): SANDOH - un système d'analyse de documents hétérogènes. In: Proceedings JADT, 1177-1184.