Compression textuelle sur la base de règles issues d'un corpus de sms

Arnaud Kirsch
UCL - CenTAL - Place Blaise Pascal 1, Louvain-la-Neuve, 1348
arnaud.kirsch@student.uclouvain.be

RÉSUMÉ

La présente recherche cherche à réduire la taille de messages textuels sur la base de techniques de compression observées, pour la plupart, dans un corpus de sms. Ce papier explique la méthodologie suivie pour établir des règles de contraction. Il présente ensuite les 33 règles retenues, et illustre les quatre niveaux de compression proposés par deux exemples concrets, produits automatiquement par un premier prototype. Le but de cette recherche n'est donc pas de produire de "l'écrit-sms", mais d'élaborer un procédé de compression capable de produire des textes courts et compréhensibles à partir de n'importe quelle source textuelle en français. Le terme "d'essentialisation" est proposé pour désigner cette approche de réduction textuelle.

ABSTRACT

Textual Compression Based on Rules Arising from a Corpus of Text Messages

The present research seeks to reduce the size of text messages on the basis of compression techniques observed mostly in a corpus of sms. This paper explains the methodology followed to establish compression rules. It then presents the 33 considered rules, and illustrates the four suggested levels of compression with two practical examples, automatically generated by a first prototype. This research's main purpose is not to produce "sms-language", but consists in designing a textual compression process able to generate short and understandable texts from any textual source in French. The term of "essentialization" is proposed to describe this approach of textual reduction.

MOTS-CLEFS : résumé automatique, compression de texte, sms, lisibilité, essentialisation.

KEYWORDS : summarization, text compression, text messaging, readability, essentialization.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 309–322,
Grenoble, 4 au 8 juin 2012.

1. Introduction

Les locuteurs, en pleine "convergence numérique", sont amenés simultanément à échanger divers types de messages écrits dans de mêmes environnements (c'est le cas de Facebook où la même interface permet l'échange de mails, de messages instantanés et de sms) et de mêmes types de messages écrits dans des environnements divers (des emails peuvent être rédigés depuis un ordinateur, une tablette ou un GSM). Un mode de communication bref et informel traverse ainsi les genres et les *media*, mode mis en évidence par la présente recherche, qui veut proposer une aide à la rédaction de messages courts.

En cela, nous nous sommes naturellement intéressé au résumé automatique et à la compression de données, mais ces deux approches ne rencontrent pas totalement notre objectif : nous avons donc choisi de dériver de nouvelles règles de contraction émanant des pratiques des locuteurs. En effet, le résumé automatique, apparu en linguistique informatique avec Luhn à la fin des années 1950 (Luhn, 1958), à pour but la simplification d'un texte en ne retenant que les idées principales de celui-ci et en les rassemblant dans un nouveau texte grammaticalement acceptable ((Minel, 2004), (Yousfi-Monod, 2007)). La compression de données vise elle à modifier la représentation binaire des données pour en réduire le poids numérique. Elle nécessite décompression du côté du récepteur.

À l'inverse, la proposition qui fait l'objet de cet article possède trois caractéristiques essentielles. (1) Le processus de simplification se situe au **caractère** près : nous cherchons à isoler les caractères les plus essentiels à la transmission de la totalité de l'information contenue dans le message initial, ce qui nous éloigne du résumé automatique *stricto sensu* (Minel, 2004). Nous nous distinguons également de la compression textuelle (Yousfi-Monod, 2007) qui cherche à supprimer les segments non porteurs de sens dans un texte. (2) Le processus de simplification ne doit pas entraver la **compréhension** : quel que soit le taux de compression recherché, nous devons produire le texte le plus décodable possible, soit celui à partir duquel, malgré la disparition d'un certain nombre de caractères, un lecteur pourra le plus aisément retrouver le texte initial. Cela signifie qu'il y aura une borne minimale en deçà de laquelle il ne sera pas possible de descendre, au risque de rendre le texte de base totalement inintelligible (puisque nous entendons transmettre le texte tel que comprimé, sans intention de décompression à l'arrivée). (3) Enfin, il s'agit d'un processus de **compression** : seuls les mécanismes permettant de réduire la taille du message doivent être pris en compte. Pour respecter ces caractéristiques, nous nous sommes basé sur le corpus de sms de (Fairon *et al.*, 2006a), et avons extrait une série de règles de compression. Malgré les caractéristiques du corpus d'entraînement, nous ne projetons pas de reproduire de l'*écrit-sms* ((Cougnon et François, 2010) ; nous le noterons e-sms). Notre intention n'est donc pas d'inverser les systèmes de normalisation tels que ceux développés par (Yvon, 2008) et (Beaufort et al., 2010).

Notre recherche, même si elle participe du résumé automatique, s'en distancie donc. En effet, selon la première caractéristique énoncée ci-dessus, nous ne cherchons pas à extraire les idées principales d'un texte, ni à produire un message grammaticalement acceptable : nous travaillons presque exclusivement au niveau des caractères et tentons de conserver toutes les informations du texte initial. Aussi nous avons préféré parler "d'essentialisation" de texte, dont voici la définition : "sous-procédé du résumé automatique focalisé sur la réduction du nombre de caractères". Le prototype de programme d'essentialisation automatique, ou essentialiseur, a notamment pour objectif

la création d'une application Twitter®[1]. Nous envisageons, à l'instar de ce qui se fait déjà en anglais (140it, 2012), qu'un utilisateur encode un texte plus grand et le réduise automatiquement à l'aide dudit programme.

Dans cet article nous aborderons tout d'abord la méthodologie que nous avons suivie pour établir nos règles. Nous expliquerons ensuite le fonctionnement global d'un premier prototype d'essentialiseur, en détaillant les règles retenues et les classements de celles-ci. Nous terminerons par l'évocation des difficultés et résistances rencontrées et des perspectives de recherche que nous envisageons.

2. Méthodologie

Notre proposition d'essentialisation se base sur le corpus de (Fairon *et al.*, 2006a). Ce corpus a été constitué dans le cadre du projet *sms4science* : 30.000 messages ont été transcrits et annotés en 2004. Ces messages sont des sms véritablement échangés entre usagers et permettent donc d'observer les pratiques réelles de simplification de la population. Nous avons travaillé sur trois sous-corpus : l'un contenant les messages d'exactement 160 caractères [2], un autre ceux de plus de 160 caractères, et le dernier les messages transcrits de plus de 160 caractères[3]. Les messages de 160 caractères ont constitué notre corpus d'entraînement, et les deux autres ont servi à valider nos observations.

Sur la base de ces sous-corpus, nous avons établi une liste de règles de contraction. Celles-ci consistent en la formalisation de certains phénomènes en vue de respecter deux enjeux majeurs : l'**intercompréhension** et la **compression**. Les phénomènes les plus récurrents sont retenus parce qu'ils sont supposés être les plus intelligibles. Parmi ceux-ci, nous ne nous intéressons qu'à ceux permettant de diminuer la taille du texte. À ces observations nous avons ajouté certaines règles qui ne sont pas issues de notre observation du corpus, par exemple la formalisation des dates et des heures (norme ISO 8601, *cf.* (ISO, 2012)). Enfin, nous avons ordonné ses règles et avons établi différents niveaux de contraction.

3. Le système

La deuxième étape de ce travail a consisté à implémenter un premier prototype d'essentialiseur. À ce stade de nos travaux, certains tests statistiques et une évaluation préliminaire sont rendus possibles, comme nous le montrons ensuite.

---

[1] Site de *microblogging*, soit d'échange de courts messages ne pouvant dépasser 140 caractères.

[2] Ancienne limite technique, la limite de 160 caractères est aujourd'hui une limite de facturation : un sms de 161 caractères sera facturé comme deux messages. Des études statistiques menées, notamment, sur notre corpus, ont démontré qu'il y avait un pic de sms de 160 caractères (Cougnon *et* François, 2010).

[3] Le premier corpus contenait 2223 sms, le deuxième 2298 et le troisième 8310.

3.1. Les règles

Nous avons établi 33 règles. Le TABLEAU 1 reprend chacune de ces règles, en donne une explication succincte et l'illustre par un exemple concret. Les codes utilisés renvoient aux classements présentés par la suite (*cf.* 3.3.).

Règle	Description	Exemple
[0101]	Réduction des smileys	• :-) ➔ :) • -___-" ➔ -_-"
[0302]	Remplacement des URL par une forme plus courte	• http://www.uclouvain.be/cental-cahiers.html#langsms ➔ http://tinyurl.com/7lumcf7
[0203]	Réduction des répétitions inutiles de caractères	• Mdrrrrrrr ➔ mdr • Looooool ➔ lol
[0504]	Réalisation des unités lexicales en logogrammes	• Vingt-quatre ➔ 24 • Et ➔ &
[0705]	Remplacement par un synonyme	• Travailler ➔ bosser
[0406]	Normalisation des dates et des heures au format ISO 8601	• Le 14 juillet 1989 vers 12h30 ➔ 1989-07-14 vers 12:30
[1107]	Suppression du pronom sujet quand les formes verbales sont non ambiguës (notamment les il impersonnels)	• Nous allons arriver ➔ allons arriver • Il faut partir ➔ Faut partir
[1208]	Si deux verbes consécutifs utilisent le même pronom sujet, et que la première occurrence du pronom sujet est maintenue, la seconde peut-être supprimée	• Je pense à toi. Je veux te revoir. ➔ Je pense à toi. Veux te revoir.
[0809]	Ellipse des "et" qui retournent une interrogation et des "ne" de la négation	• Je vais bien, et toi ? ➔ Je vais bien, toi ? • Je ne pense pas ➔ Je pense pas

Règle	Description	Exemple
[1010]	Suppression des mots répétés (ne concerne pas les pronoms personnels)	• Hello hello ➔ Hello
[2411]	Réduction aux initiales des noms propres composés et des noms communs composés les plus courants	• Pierre-Yves ➔ P.Y. • Week-end ➔ w.e.
[3012]	Simplification des répétitions de pronoms personnels	• Nous nous reverrons ➔ nous reverrons
[1513]	"tu" et "vous", suivis de voyelles, deviennent, respectivement, "t'" et "z'"	• Tu arrives ➔ t'arrives • Vous arrivez ➔ z'arrivez
[1614]	Réduction des "tu" et "je" à l'initiale	• Tu vas ➔ tvas • Je pense ➔ jpense
[2115]	Fusion des mots composés, locutions, etc. Reconnus par ailleurs	• Porte-monnaie ➔ portemonnaie
[2016]	"bisou(s)" en fin de texte > "x"	• À plus, bisous. ➔ à plus, x
[2317]	Les monosyllabiques courants sont réduits à leur squelette	• Temps ➔ tps
[2218]	Certaines fins de mots courantes sont réduites à leur squelette	• Internement ➔ internemt
[1319]	Certains mots très courants du texte sont réduits à leur squelette	• Beaucoup ➔ bcp
[1420]	Apocopes des mots les plus fréquents	• Anniversaire ➔ anniv
[2721]	Suppression des consonnes finales muettes, sauf les marques du pluriel	• À travers ➔ à traver
[2522]	Si pas d'ambiguïté, suppression des marques muettes du pluriel	• Journaux ➔ journau
[1723]	Suppression des schwas peu prononcés et fusion avec le mot suivant	• Dernièrement ➔ dernièrment • Je ne sais pas -> je nsais pas

Règle	Description	Exemple
[3124]	Suppression de tous les schwas ; les monosyllabiques fusionnent avec les mots qu'ils précèdent	• Je me grouillerai ➔ jmgrouillerai
[1825]	Suppression des "h" muets	• Hérisson ➔ érisson
[2626]	Simplification des doubles consonnes	• Notamment ➔ notament
[2827]	Les phonèmes transcrits par plusieurs caractères sont remplacés par des caractères uniques	• Je voulais que tu viennes ➔ je voulè ke tu vienes
[2928]	Phonétisation des syllabes par la lettre, le signe ou le chiffre	• J'ai envie de toi ➔ G envie 2 toi
[0629]	Simplification des abréviations	• P.-S. ➔ PS
[1930]	Suppression des points finaux, des apostrophes, des traits d'union et simplification des points de suspension (réduits à deux points successifs)	• J'imagine ➔ jimagine • Penses-tu ➔ pensestu
[3231]	Suppression des doubles points et points-virgules	• Je disais : comment vas-tu ? ➔ je disais comment vas-tu ?
[0932]	Les espaces séparant deux systèmes graphiques différents sont supprimés	• J'ai passé 1 bonne journée ➔ j'ai passé1bonne journée • Je suis dégouté ! Et il est là ! ➔ je suis dégouté!Et il est là!
[3333]	Suppression de tous les espaces : *scriptura continua* (sauf entre deux caractères numériques)	• Je ne crois pas ➔ jenecroispas

TABLEAU 1 - *Présentation des règles*

3.2. Remarques sur le tableau

Nous regroupons ici une série de remarques concernant le Tableau 1 :

- L'ordonnancement des règles n'est pas figé : en fonction des résultats d'une enquête qualitative, ou en cas d'ajout ou de suppression d'une règle, l'ordre global peut être repensé : il y a une grande variété de types de compression d'un même texte, nous en

choisissons une, sans prétendre qu'elle soit la meilleure, afin qu'elle serve de ligne de conduite ;

- Certaines règles, comme [0101] et [0203] proviennent typiquement de caractéristiques propres au corpus sur lequel nous nous sommes basé : le but de notre système étant de pouvoir travailler sur n'importe quel type de texte, nous devons donc y intégrer des règles plus spécifiques à certains textes qu'à d'autres ;
- [0203] seule : il est clair que ces répétitions sont porteuses de sens également, et que les supprimer revient à enlever une partie du sens encodé initialement. C'est néanmoins un choix que nous posons de les réduire, afin de gagner quelques caractères ;
- Un logogramme (mentionné en [0504]) est un "[d]essin représentatif d'une notion (logogramme sémantique ou idéogramme) ou d'une suite phonique constituée par un mot (logogramme phonétique ou phonogramme)." (TLFi, 2012) ;
- Les synonymes, tels qu'évoqués par la règle [0705], soulèvent un problème évident : il n'existe que très peu de synonymes parfaits, et une substitution peut donc fréquemment modifier une partie du sens apporté. Pour cette approche préliminaire, nous avons établi manuellement une courte liste de synonymes, dont les variations sémantiques portent plutôt sur le registre de langue ("bosser" pour "travailler", par exemple). Ce choix est certes discutable, mais est en lien avec notre remarque préliminaire : nous cherchons à cadrer avec un mode de communication bref et informel ;
- De nombreuses règles, comme [1107], [1208], [0809], etc. se basent sur le constat que le français est parfois redondant (répétitions, reformulations, etc.). Nous cherchons à définir ces redondances (et d'autres) pour les réduire au maximum et donc gagner de l'espace ;
- La règle [2115] prévoit la fusion des mots composés et autres locutions (puisque plus facilement identifiables/compréhensibles par le lecteur). Si nous avions simplement décidé de fusionner les collocations[4] se serait posée la question de l'évaluation du degré de figement des syntagmes. Nous avons donc tout d'abord repris une liste finie de mots composés que nous avons augmentée de locutions relevées automatiquement dans notre corpus[5] ;
- Les règles relatives aux squelettes consonantiques ([2317], [2218] et [1319]) sont établies, elles aussi, sur la base de notre corpus : nous avons relevé de très nombreux squelettes et en avons dégagé des constantes. La difficulté de ces trois règles est de déterminer quelles unités sont plus facilement compréhensibles lorsqu'elles sont autant réduites. En ce qui concerne leurs fréquences d'apparition, nous nous sommes basé sur une évaluation statistique de notre corpus ;
- Lors de la phonétisation des syllabes [2928], les lettres à lire pour leur valeur phonétique seront encodées en majuscule. Aussi, avant l'application de cette règle, toutes les majuscules orthographiques sont réduites, afin d'éviter tout risque de

---

[4] Au sens de « cooccurrences statistiquement privilégiées ».

[5]Pour ce faire, nous avons mesuré la fréquence d'apparition de chaque mot dans les transcriptions des sms du corpus, ainsi que celle de leurs contextes gauche et droit. Nous avons donc pu établir les fréquences d'apparition de tous les syntagmes du corpus et avons retenu les plus présents. Nous avons utilisé les transcriptions afin d'éviter les problèmes d'instabilité orthographique.

confusion. La question se pose cependant de savoir quel système suivre lors de la fusion des caractères [3333] : maintenir la phonétisation ou concaténer les mots non phonétisés en utilisant des majuscules pour marquer l'emplacement des espaces supprimés ? Cette question devra être tranchée après enquête qualitative ;

- Nous distinguons quatre systèmes graphiques pour établir la règle [0932] : les lettres, les chiffres, les signes de ponctuation et les symboles. Cette distinction est établie par nous-même. La *scriptura continua*[6] peut sembler excessive. Elle est cependant présente dans quelques sms, et apparaît à quelques reprises dans l'histoire de l'écriture (par exemple durant l'Antiquité). Ainsi nous décidons de la maintenir.

3.3. Leur classement

Nous avons ordonné ces règles selon deux classements distincts. L'un dépendant de l'ordonnancement informatique, l'autre prenant en compte l'application du système.

Le premier classement répond à deux exigences : d'une part classer les règles selon leur influence sur la lisibilité (ce classement est représenté par les deux premiers chiffres du code d'une règle) ; d'autre part selon l'ordre dans lequel elles doivent être appliquées pour ne pas se gêner mutuellement : par exemple, il faut formaliser les dates avant de phonétiser les noms des mois (les deux derniers chiffres du code illustrent ce second ordre).

Le second classement, orienté vers l'application finale, envisage deux types d'utilisation : le premier, qualitatif, où l'utilisateur doit choisir entre quatre niveaux d'essentialisation prédéfinis, et le second, quantitatif, où le programme connaît le seuil de caractères sous lequel ramener le texte initial et applique les règles jusqu'à y parvenir (ou non).

3.4. Les quatre niveaux

Revenons plus en détail sur les quatre niveaux d'essentialisation que nous proposons. Chaque niveau permet de définir une séquence de règles à appliquer pour atteindre un certain degré d'essentialisation. Nous envisageons d'abord deux exemples produits par notre prototype, puis nous les commentons et détaillons les choix théoriques sous-jacents.

3.4.1.Définitions

Superficiel : ce niveau touche uniquement à ce que nous jugeons accessoire, aux caractères qui ne sont là que pour fluidifier la lecture, comme les répétitions émotives de caractères ("looooool"), les espaces précédant certains signes de ponctuation ("Non !"), l'utilisation de logogrammes ("vingt-quatre" > "24" ; "et" "&"), la suppression de certains mots redondants, par ailleurs souvent absents du langage oral ("ne" de la négation), etc. Ce niveau ne supprime donc que des caractères que nous pourrions qualifier de sémantiquement moins pertinents.

Conventionnel : Il s'agit d'appliquer une série de règles qui sont des pratiques fréquentes de l'écrit (bien au-delà des seuls sms), ou des transcriptions de l'oral. Nous pensons par exemple aux apocopes de mots fréquents ("anniv"), à l'effacement des schwas toujours silencieux ("effacement"), aux abréviations courantes ("tps", "ds",

[6] Le terme *scriptio continua* est également employé.

"qq"...), à l'élision de certains pronoms sujets, ou à leurs simplifications ("zavez vu", "faut y aller"...). Nous commençons ici à atteindre plus conséquemment l'orthographe et la syntaxe, mais selon des pratiques qui, par leur fréquence dans notre corpus et notre propre perception, semblent rapidement déchiffrables.

Morpho-syntaxe du sms : Ce niveau se concentre sur des phénomènes couramment trouvés dans des sms, notamment la phonétisation de certains phonèmes transcrits par plusieurs caractères, la contraction en squelettes consonantiques de certaines fins de mots, ou de certaines unités lexicales fréquentes. Nous ne recensons ici que les phénomènes les plus fréquents afin que nos messages restent les plus compréhensibles possible : en effet nous partons du principe que les phénomènes les plus courants seront les plus intelligibles.

Cryptage : On parle ici de l'application de toutes les règles observées dans notre corpus, afin de gagner un maximum de caractères. Il ne s'agit plus de tenter de conserver un minimum de décodabilité, le but est la compression : phonétisation de toutes les syllabes, suppression d'un partie de la ponctuation, suppression des espaces, etc.

3.4.2.Application

Les quatre niveaux d'essentialisation que nous avons définis servent à dégager un formalisme. Ils ont été posés arbitrairement, puisque seule la définition des niveaux est importante. Le but est d'obtenir une description stricte des attentes et conditions de ces niveaux, afin de choisir les règles qui y correspondront le mieux. Quel que soit le nombre

Niveau	Forme	Taille
Corpus	Hi tite puce,g pensé a tfèr 1pti sign 2vi,tu m'mank grav.pq pa svoir 2m1?alé pass 1bonnuit	*90*
Forme standard	Hi petite puce, j'ai pensé à te faire un petit signe de vie, tu me manques grave. Pourquoi pas se voir demain ? Allez passe une bonne nuit	*138*
Superficiel	Hi petite puce,j'ai pensé à te faire1petit signe de vie,tu me manquegrave.Pourquoi pas se voir demain?Allez passe1bonne nuit	*124*
Conventionnel	Hi ptit puc,jai pensé à tfair1ptit signde vie,tu mmanqugrav.Pourquoi pas svoir demain?Allez pass1bonnnuit	*105*
Morpho-syntaxe du sms	hi ptit puc G penC à t9èr1pttsign2vi tmmankgrav pourkoi pasvoir2m1alé pas1bOn n8	*80*
Cryptage	hiptitpucGpenCàt9èr1pttsign2vitmemankgravpourkoip asevoir2m1alépas1bOnn8	*71*

TABLEAU 2 - *Essentialisation d'un message illustrant la proximité*

de niveaux définis, l'important réside dans la gradation continue entre ceux-ci, tant du point de vue de l'intercompréhension que de la compression.

Deux exemples permettent d'illustrer les résultats de chaque niveau. Le premier exemple[7] est un texte illustrant l'immédiateté (*cf.* TABLEAU 2), le second[8] exposant la distance[9] (*cf.* TABLEAU 3).

Niveau	Forme	Taille
Forme standard	Je rappelle que les banques ont payé plus à l'État belge que ce qu'on leur a donné. Les garanties sur Dexia ont rapporté bien plus que le milliard qui a été mis dans Dexia. Tout a été payant.	*191*
Superficiel	Je rapelle que lebanques ont payé plus à l'État belge que ce qu'on leur a doné.Les garanties sur Dexia ont rapporté bien plus que le milliard qui a été mis dans Dexia.Tout a été payant.	*185*
Conventionnel	Jrapell qulebanqus ont payé plus à lÉtat belgque cquon leur a doné Les garanties sur Dexia ont rapporté bien plus que lmilliard qui a été mis ds Dexia.Tout a été payant.	*169*
Morpho-syntaxe du sms	jrapèl klébanks on pèyé + à léta bèlj ke ce kon leur a dOné lé garantisur dèxia on rapOrT bi1 + ke lmiliar ki a éT mi ds dèxia tout a éT pèyan	*144*
Cryptage	j r a p è l k e l é b a n k o n p è y é +àlétabèljkecekoleuradOné.légarantisurdèxiaonrapOrT bi1+klmiliarkiaéTmidandèxiatoutaéTpèyan	*110*

TABLEAU 3 - *Essentialisation d'un message illustrant la distance*

3.4.3.Analyse des résultats

Le niveau **superficiel** enregistre un taux de réduction de 14% pour le premier exemple, et de seulement 3% pour le second. Le taux moyen de compression de ce niveau se situe à hauteur de 9%. Les textes produits ne sont pas très réduits, mais restent compréhensibles : la forme des mots n'est pas atteinte. Il reste quelques erreurs produites par notre prototype, notamment "lebanques", qui aurait dû rester "les banques", et la même chose pour "maquegrave". Une assez grande différence que l'on

[7] Il s'agit d'un sms issus du corpus.

[8] (Libre Belgique, 2011 : 5)

[9] L'opposition immédiat *versus* distance est proposée par (Koch et Österreicher, 2001). La distance communicative dénote le degré d'implication (immédiat) ou de détachement (distance) des locuteurs dans le discours.

trouve entre les taux de compression des deux messages est probablement due au fait que le premier message est du même type que notre corpus d'entraînement.

Au niveau **conventionnel**, le sms obtient un taux de réduction de 24% et l'extrait de journal 10,5%. À nouveau, l'écart est assez large entre les deux. Le taux moyen se situe à 18%. Par rapport au premier niveau, la progression est assez marquée du point de vue de la compression. En ce qui concerne la compréhension, le texte devient déjà plus hermétique, principalement lorsque des schwas ont été supprimés, avec l'espace qui les suivait. Cette agglutination de mots, si elle permet de gagner des caractères, gêne le décodage du texte. Cependant, si nous regardons le troisième niveau, l'écart perceptif entre celui-ci et le deuxième qui nous occupe semble plus important qu'entre les deux premiers niveaux. Notre impression est donc que la gradation n'est pas continue entre les trois premiers niveaux.

Les taux de réduction des deux exemples du niveau **morpho-syntaxe du sms** sont respectivement de 42% et 25%. Le taux moyen de ce troisième niveau est de 34%. Les deux textes reprennent certaines caractéristiques des sms et leur aspect est similaire à l'écrit sms. Nous pouvons d'ailleurs apprécier cette ressemblance en comparant le niveau 3 du premier exemple à la version originale de ce sms dans notre corpus. Il y a des différences, (rappelons que nous ne cherchons pas à produire de l'e-sms) mais nous devons aussi garder à l'esprit que les possibilités de combinaison des règles sont très nombreuses, et que notre système est déterministe : il produira toujours la même sortie pour un même texte. À l'inverse, un locuteur n'emploiera pas forcément les mêmes techniques d'un message à l'autre.

Les deux taux de réduction du niveau **cryptage** convergent assez fort, puisque le premier texte atteint un taux de réduction de 49% et le second de 43%. Le taux moyen est de 44%. La compression est excellente, puisque près de la moitié du texte a été supprimée. La compréhension est par contre bien plus délicate. Il semblerait qu'il manque trop de caractères pour que la lecture reste fluide. En effet, l'absence d'espace empêche le lecteur de délimiter les unités lexicales. Les deux textes produits ressemblent à une suite continue de phonèmes, à l'instar de ce qu'est le signal sonore.

3.5. Évaluation

Étant donné que nous sommes à une étape préliminaire de notre recherche, nous n'avons pas encore pu réaliser d'évaluation qualitative[10] de ce travail. Nous proposons donc une évaluation en deux parties : une évaluation quantitative d'une part, et le relevé manuel des limites de notre système, d'autre part.

3.5.1.Évaluation quantitative

Pour établir les taux moyens de compression, nous avons utilisé un corpus de test composé de cent textes courts ou extraits répartis comme suit : cinquante transcriptions de sms tirées de notre corpus, quarante extraits de journaux et dix extraits littéraires[11]. Nous avons mesuré le taux moyen de compression produit par les quatre niveaux mentionnés ci-dessus sur l'ensemble de notre corpus de test. Il nous est impossible de mesurer la divergence de nos résultats par rapport à une référence, celle-ci n'existant

[10] Validation des choix théoriques par un échantillon de testeurs humains.

[11] De dix auteurs différents.

pas. À l'inverse du résumé automatique, nous ne pouvons essentialiser manuellement, ni comparer les résultats d'un système équivalent.

Au mieux, dans le cas des 50 sms, aurions-nous pu comparer nos taux de compression à ceux des locuteurs. Cependant, la variabilité de l'écrit sms nous empêcherait de savoir à quel niveau d'essentialisation comparer le sms réel. Et un taux moyen ne serait pas plus éclairant. Nous devons donc nous limiter à une observation des taux moyens de chaque niveau. Le but est d'évaluer la gradation de la compression entre les quatre niveaux proposés.

Les chiffres ainsi obtenus viennent, d'une certaine façon appuyer notre première impression : la progression n'est pas continue entre les quatre niveaux d'essentialisation tels qu'ils sont définis actuellement. Nous passons en effet de 9% à 14% puis à 34% pour atteindre 44% au dernier niveau. Le deuxième niveau semble donc ne pas être un bon intermédiaire entre le premier et le troisième ; ou alors ceux-ci sont-ils respectivement trop conservateur et trop destructeur. Nous devrons attendre d'avoir obtenu les résultats d'une évaluation humaine pour le déterminer et corriger ces premières propositions.

Cependant, malgré certains soucis de gradation et quelques problèmes d'implémentation, nous obtenons assez rapidement des taux de réduction intéressant. Le dernier niveau peut aller jusqu'à réduire de moitié la taille du texte initial, et le deuxième s'approche de 15% de réduction. Considérant que notre but est de transmettre un texte maintenant toutes les nuances du texte d'origine et générant le moins d'ambiguïté possible, ces premiers résultats sont assez encourageants, même s'ils ne nous permettent pas encore d'évaluer objectivement l'évolution de la lisibilité des textes produits.

3.5.2.Limites du système

Il convient d'être assez critique au regard de nos premiers résultats. Nous rencontrons en effet deux types de problèmes avec notre premier prototype d'essentialiseur : le premier se situe au niveau de l'algorithme général d'application des règles, l'autre au niveau de certaines règles elles-mêmes.

Tout d'abord, bien que nous en ayons conscience et ayons tenté d'éviter ce type d'inconvénients, certaines règles gênent l'application des suivantes, voire annulent leurs résultats. Par exemple, la règle [2031] supprime une série de signes de ponctuation qui pourraient se retrouver dans des adresses web préalablement réduites par la règle [0302]. Ou encore, si la règle [0504], qui remplace, notamment, les nombres écrits en toutes lettres par leur équivalent en chiffres arabes ne s'applique pas correctement, toutes les autres règles qui s'appuient sur des chiffres seront induites en erreur. Ensuite, certaines règles qui semblent évidentes pour l'esprit humain ne sont pas toujours les plus simples à formaliser pour la machine. Il en va ainsi de la règle [0504], que nous avons déjà citée : des trente-quatre, c'est elle qui fut la plus difficile à implémenter, et à optimiser. Il existe en effet une multitude de configurations possibles pour l'énonciation de nombres, et nous n'avons considéré que les plus fréquentes, partant du double constat que nous pourrons améliorer cette règle lors des prochaines étapes de notre réflexion et qu'il y aura probablement fort peu de cas où les utilisateurs entreront des nombres en toutes lettres.

Mais le principal problème que nous rencontrons reste la variabilité de la langue : comme nous l'avons déjà précisé ci-dessus, le français, comme d'autres langues naturelles, offre d'innombrables possibilités de variations que nous ne pourrons pas

toutes envisager, or certaines d'entre elles amèneront notre système à commettre des erreurs et à manquer son objectif (par exemple les dates : nous tentons de les formaliser, mais un utilisateur pour aussi bien entrer "le 2 juillet 1989" que "2 du 7 89" ou encore "7-2-89"). Ce premier prototype doit donc encore être amélioré.

4. Conclusions

La première étape de cette recherche ouvre de très nombreuses perspectives d'optimisation. D'un point de vue technique, pour corriger les problèmes mentionnés ci-dessus, nous envisageons notamment de marquer certains caractères, de sorte qu'ils ne puissent plus être modifiés. Nous continuons par ailleurs à réfléchir à un meilleur algorithme de conversion de nombres. Au niveau du système lui-même, d'autres règles pourraient être ajoutées, s'appuyant éventuellement sur d'autres types de corpus qu'un corpus de sms, pour y observer d'autres mécanismes. Enfin les quatre niveaux que nous avons présentés devront également être affinés, et de nouveaux pourraient être ajoutés au système.

Une évaluation qualitative effectuée par des locuteurs nous permettra d'améliorer notre estimation de la lisibilité des textes essentialisés, éventuellement de repenser l'organisation des règles ou des niveaux du système actuel, mais également de valider nos propositions quant à l'essentialisation en général. Il reste donc de nombreuses pistes de réflexion, et nous n'apportons ici que nos premières propositions.

Si le champ de recherche ici présenté peut sembler restreint ou limité à quelques applications ludiques, nous affirmons qu'il n'en est rien. Il ouvre d'une part la voie à une approche originale de l'étude des phénomènes de contraction présents dans les sms en ce qu'il postule leur application formelle. D'autre part, il propose une nouvelle méthode de compression textuelle, pertinente notamment dans le cadre de sites de microblogging comme twitter® ou dans d'autres situations de communication textuelle soumises à une forte contrainte d'espace (affichage de messages d'alerte, de notifications). Mais il reconnaît surtout l'existence d'un nouveau mode de communication bref et informel, et propose une aide à son utilisation.

Remerciements

Nous tenons à remercier Cédrick Fairon qui nous a soufflé l'idée de cette recherche. Nous souhaitons que Louise-Amélie Cougnon soit également reconnue pour son indéfectible soutien, sa disponibilité et son aide précieuse. Enfin merci à MM Beaufort, Bouraoui et Watrin pour leurs conseils avisés.

Références

140it, 2012-03-20 http://140it.com.

2011-11-05, La Libre Belgique, Bruxelles.

BEAUFORT Richard, COUGNON Louise-Amélie, FAIRON Cédrick *et* ROEKHAUT Sophie, 2010, "Une approche hybride traduction/correction pour la normalisation des sms", *in TALN*, Montréal.

COTTIN Florent, 2011, *Le "pourrisseur de texte" du RALI*, Université de Montréal.

COUGNON Louise-Amélie *et* FRANÇOIS Thomas, 2010, *Étudier l'écrit sms. Un objectif du projet sms4science*, Linguistik Online.

COUGNON Louise-Amélie *et* LEDEGEN Gudrun, 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*, Oxford.

FAIRON C., KLEIN J. *et* PAUMIER S., 2006a, *sms pour la science. Corpus de 30.000 sms et logiciel de consultation*, Presses universitaires de Louvain, Louvain-la-Neuve.

FAIRON C., KLEIN J. *et* PAUMIER S., 2006b, *Le langage sms, étude d'un corpus informatisé à partir de l'enquête "faîtes don de vos sms à la science"*, Cental (Cahier du Cental), Louvain-la-Neuve.

GLOR HOWARD Paul, 1993, *The Design and Analysis of Efficient Lossless Data Compression Systems*, Brown University, Providence.

International Organisation for Standardization, 2012-03-20 www.iso.org.

KOCH Peter *et* ÖSTERREICHER Wulf, 2001, « Gesprochene Sprache und geschriebene Sprache », *in Lexikon der romanistischen Linguistik*, Günter Holtus, Tübingen, pp. 584-627.

LUHN H.P., 1958, "The Automatic Creation of Literature Abstracts", *in IBM Journal of Research and Development*, pp. 159-165.

MINEL Jean-Luc HDR, 2004, *Le résumé automatique de textes : solution et perspectives*, Sorbonne, Paris.

TLFi - Trésor de la Langue Française informatisé, 2012-03-24, www.cnrtl.fr/definition/.

YOUSFI-MONOD Mehdi, 2007, *Compression automatique ou semi-automatique de textes par élagage des constituants effaçables : une approche interactive et indépendante des corpus*, Thèse de doctorat à l'Université de Montpellier II, Montpellier.

YVON François, 2008, *Réorthographier des sms*, LIMSI, Paris.