Towards Electronic SMS Dictionary Construction: An Alignment-based Approach

Cédric Lopez*, Reda Bestandji, Mathieu Roche**/***, Rachel Panckhurst******

*Viseo Research Center, France

**LIRMM, CNRS, Université Montpellier 2, France

*** TETIS, Cirad, Irstea, AgroParisTech, France

****Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3, France

clopez@objetdirect.com, reda.bestandji@hotmail.fr, mathieu.roche@cirad.fr, rachel.panckhurst@univ-montp3.fr

Abstract

In this paper, we propose a method for aligning text messages (entitled AlignSMS) in order to automatically build an SMS dictionary. An extract of 100 text messages from the *88milSMS* corpus (Panckhurst *el al.*, 2013, 2014) was used as an initial test. More than 90,000 authentic text messages in French were collected from the general public by a group of academics in the south of France in the context of the *sud4science* project (http://www.sud4science.org). This project is itself part of a vast international SMS data collection project, entitled *sms4science* (http://www.sms4science.org, Fairon *et al*. 2006, Cougnon, 2014). After corpus collation, pre-processing and anonymisation (Accorsi *et al*., 2012, Patel *et al*., 2013), we discuss how "raw" anonymised text messages can be transcoded into normalised text messages, using a statistical alignment method. The future objective is to set up a hybrid (symbolic/statistic) approach based on both grammar rules and our statistical AlignSMS method.

Keywords: SMS, alignment, electronic dictionaries

1. Introduction

Over the past few decades, new forms of communication have appeared with the development of various types of technology. Among these, SMS (Short Message Service) communication is constantly changing and evolving. Text messages include specific features, and can be defined as belonging to a particular form of *mediated electronic discourse* (Panckhurst, 2006).

As (Kobus *et al*., 2008) note, the characters in SMSs can be diverted from conventional use in order to encode linguistic form, emotions, feelings, and attitudes. Natural language processing tools can be used for semi-automatic processing of text messages, and may help to improve electronic dictionary construction.

In the context of the *sud4science* project (http://www.sud4science.org), more than 90,000 authentic text messages in French were collected (Panckhurst *et al*., 2013). A pre-processing phase (elimination of spurious information: text messages received twice or more from the same person; those received from abroad (since the study was limited to mainland France); advertisements; automatic SMSs received from telephone operators, etc.) and an anonymisation phase (Accorsi *et al*., 2012; Patel *et al.*, 2013) were then conducted. In this paper, we present the next step: an alignment method. Alignment consists of matching textual atomic elements from an authentic anonymised "raw" SMS (including abbreviations, repetition of letters, spelling mistakes, and so forth) with a transcoded "normalised" SMS (in our case, in the French language) (see Fig. 1). Alignment techniques can be useful for real-life applications, such as SMS vocalisation for landline phones (Guimier De Neef and Fessard, 2007b).

"Raw" SMS	Jspr	ktu		va	bien
Normalised SMS	J'espère	que	tu	vas	bien

Figure 1: Alignment between a "raw" and a normalised SMS ("*I hope you are well"*).

First, we present the *sud4science* project in order to situate the alignment phase within our processing chain (§ 2). Next, we present related work and introduce our alignment method applied to both authentic anonymised and normalised SMSs (§ 3). Then we evaluate our approach (§ 4) and finally present future work (§ 5).

2. The *sud4science* project

2.1 Context

An international project, entitled *sms4science* (coordinated by CENTAL, Centre for Natural Language Processing, Université Catholique de Louvain, Belgium), aims at gathering, building a world-wide data base and analysing authentic text messages (http://www.sms4science.org, Fairon *et al.,* 2006; Cougnon, 2014). Several related SMS data collections have taken place since the initial Belgian one (Reunion Island, 2008, http://www.lareunion4science.org/; Switzerland, 2009-2010, http://www.sms4science.ch/; Quebec, 2010, http://www.texto4science.ca/; French Rhône-Alps, 2010,

http://www.alpes4science.org/). The most recent initiative for the French language was the subsequent *sud4science* project. Over 90,000 authentic text messages in French were collected during a three-month period (spanning September 15th to December 15th, 2011), by academics in the Languedoc-Roussillon region of France, in the context of the *sud4science* project (*cf.* Panckhurst & Moïse, 2012; Panckhurst *et al*., 2013; Accorsi *et al*. 2012; Patel *et al*., 2013). All of the SMS data collections are linked to the original Belgian project, and the aim is to organise all of the anonymised data into a global database, containing authentic SMS writing in French, over a period of a decade, from various countries (*cf.* Cougnon, 2014). This will provide important sources and allow comparative pluridisciplinary research to take place in years to come.

The phases of the *sud4science* project are indicated in Fig. 2) acquisition of our SMS corpus (entitled *88milSMS*), 2) pre-processing and anonymisation (*cf.* Accorsi *et al*., 2012; Patel *et al*., 2013), 3) transcoding and normalisation, 4) alignment.

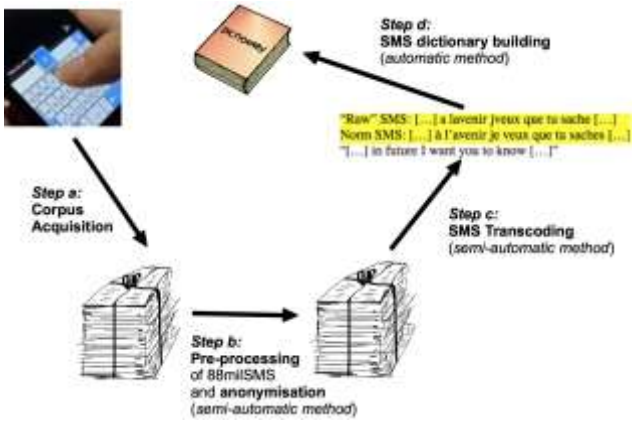

Figure 2: Overall process of the *sud4science* project

2.2 Anonymisation

The anonymisation of our *88milSMS* corpus was conducted in two phases (Accorsi *et al*., 2012, Patel *et al.*, 2013): an automatic phase and a manual one, in order to minimise computer errors. Researchers are legally required to eradicate traces of identification from text messages, given that they contain personal, therefore potentially sensitive, information[1]. Not only first names, nicknames, surnames need to be rendered anonymous, but also any information which can help recognise an SMS donor, the recipient, or a third party (postal addresses, emails, URLs, codes, places, telephone numbers, brand names, firms, etc.).

The main purpose of the automatic approach was to process our corpus using a wide set of dictionaries (country names, cities, first names, etc.) and anti-dictionaries (containing words that do not require anonymisation, such as common nouns, adjectives, verbs, adverbs, prepositions, articles, etc.) in order to identify and hide words that must be anonymised.
When a word is anonymised, it is replaced by a code conforming with the following format:

<[Tag]_[#characters]> where "Tag" indicates the type of the word (e.g. First name, Last name) and "#characters" indicates the number of characters in the word. For instance, the following SMS:

"Coucou Patrice, ça va?" (*Hi Patrice, how are you?*)

becomes:

"Coucou <PRE_7>, ça va?"

Thus, "Patrice" is replaced by < PRE_7> where:
- PRE: First name (*prénom*)
- 7: number of characters in "Patrice"

Once the words of the corpus are identified (thanks to dictionaries and anti-dictionaries), the corpus is processed according to the labels and the words are either (a) anonymised, (b) ignored, or (c) highlighted so that the human linguist expert annotators can then manually process the data, owing to non-identification of a word in a dictionary, or ambiguous identification in both types of dictionaries. Words that only appear in the dictionary are automatically anonymised. Words that only appear in the anti-dictionary are not anonymised. Other words become candidates for the semi-automatic anonymisation (since they are considered as ambiguous or unknown).

SMS writing is often very creative, rendering the anonymisation process even more difficult: first names may (or may not) be capitalised (*Cédric*/*cédric*); characters may be repeated (*Cééééédric*); diminutive/abbreviated forms appear (*Nico* for *Nicolas,* Nicholas), words may contain spelling 'mistakes' (*surment*, instead of *sûrement*, most likely), no accents (*desole*, instead of *désolé*, sorry), incorrect accents (*dèsolè*, instead of *désolé*, sorry), onomatopoeia (*mouhahaha,* ha ha ha), elision without apostrophes (*jexplique,* I explain), agglutination (*jtaime*, I love you), etc. So, in order to take into account the specific nature of SMS data, we added different heuristics to solve some of the above problems.

The software successfully anonymised over 70% of the corpus. The remaining 30% was then retained for the semi-automatic phase, requiring human expert linguist annotators. The following tags were used for the anonymisation process: PRE (First Name), NOM (Last Name), SUR (Nickname), ADR (Address), LIE (Place), TEL (Telephone Number), COD (Code), URL (URL), MAR (Brand Name), MEL (Email), Other.

In order to improve the system and to reduce the workload in the manual validation phase, we decided to implement a second technique, based on supervised machine learning. This method learns from annotated training data, and is able to make predictions on new test data. Our objective is to label at message level (not at word level).

Here is the list of features extracted from each SMS:
- The number of words from the SMS that are in the

[1] Our University's *National Commission for Data Protection and Liberties* (CNIL, http://www.cnil.fr) correspondent, Nicolas Hvoinsky, made a declaration rendering the anonymisation operation compulsory by 30 September 2013.

dictionary of abbreviated forms specific to text messages (anti-dictionary).
- The number of words that are in the LEFFF dictionary of French (anti-dictionary).
- The number of words that are in the dictionary of first names. This is particularly useful for the types of messages to be anonymised.
- The number of words that are in the dictionary of country names.
- The length of the SMS.
- The number of words in upper case in the text.
- The average word length of the SMS.
- The number of pronouns in the text message.
- The number of numbers.
- The number of punctuation marks.
- The number of words with repeated characters.

We tested several algorithms from Weka (Hall *et al*., 2009). The most appropriate algorithm (which we use) for this task seems to be the Decision Trees algorithm (DT). Other classifiers, such as SVM (Support Vector Machine) and Naive Bayes, use separation planes or probabilities, and these numbers are not understandable for a human examiner.

Finally, we combined the learnt model with the symbolic system, according to the following idea: the manual analysis by an expert is useful only if there is a disagreement between the two automatic methods.

After anonymisation of the *88milSMS* corpus (for further details, *cf.* Patel *et al.*, 2013), and normalisation, in this paper, we propose an initial approach for aligning SMSs (entitled AlignSMS), in order to automatically build an SMS dictionary.

3. The AlignSMS approach

3.1 Related work

In the literature, there are two types of SMS alignment approaches: symbolic or statistical. (Guimier De Neef and Fessard, 2007a) deal with the normalisation issue in the same way that automatic spelling correction is applied, using a symbolic approach based on phonetic, morphosyntactic cutting, and taking into account repetition of letters. In all, 2,000 grammatical rules are used. (Kobus *et al.*, 2008) use phonetic non-determinist contextual rules describing grapheme/phoneme correspondences (140 rules are used, supplemented by a dictionary containing 23,000 words). Nevertheless, numerous spellings are distinct from the phonetic form (e.g., consonant contractions/clippings and abbreviations: *bcp* (beaucoup, *many*), dc (donc, *therefore*), pr (pour, *for*), ds (dans, *in*); double consonants: ele (elle, *she*), poura (pourra, *will be able to*); semantic abbreviations//initialisms (abbreviations reduced to initials): t (te/tu, *you*) p (peux/pas, *can*, *not*) but they are modelled with basic dictionaries. In these approaches, transcoding techniques from "raw" to normalised text messages are not resolved. Aw *et al.* (2007) view the SMS normalisation task as a 'translation' problem from the SMS form to the standard English language form and suggest adapting a phrase-based statistical MT model. Choudhury *et al.* (2007) prefer implementing the noisy channel metaphor (Shannon, 1948). The algorithm for aligning is based on the concept of *pivots*. The unaligned words appearing between two consecutive pivots are then recursively aligned by searching for more pivots between them. Beaufort *et al.* (2008) highlight that traditional alignment methods based on edition distance are unusable within the context of SMS alignment. Indeed, the phenomenon of agglutination and the sequential character of traditional implementation (for example with Viterbi's algorithm (Viterbi, 1967)) infer that such an approach is unsatisfactory for our problem solving. His approach consists of identifying the distance of minimal edition between the "raw" SMS and the normalised SMS, using finite-state machines. The objective is to select the best alignment candidate among all of the possible alignments between the "raw" SMS and the normalised SMS.

3.2 Linguistic transcoding

Once the anonymisation processing has taken place, the "raw" SMS data are ready to be *transcoded* into "standardised" [2] French to allow for prospective computational linguistic applications (including syntactical parsers). The idea is to rectify spelling and grammar, yet not "inject" more than is necessary for clear understanding. Many problems arise with the transcoding phase, from a linguistic point of view, and are not necessarily easy to resolve, since linguists working from different perspectives may have differing opinions on just how to conduct the shift from "raw" text messages to standardised French. In our research, we have applied several general rules: a) modified words are transcoded to standardised French (this includes any type of *substitution*: 'o' (eau, *water*); *reduction*: 'ordi' (ordinateur, *computer*), 'zou' (bisou, *kiss*); *suppression*: ca (ça, *that*), voila (voilà, *there*), *addition* (suuuupppeeerrr!!!!!! (*super!*)), etc. appearing within a text message); b) if an item appears in the Petit Robert 2014 (PR14) dictionary, it is retained as is when transcoded: e.g. 'frérot' is a popular form of 'brother' in French, but is not modified to 'frère' since the entry exists in the PR14; this is also the same for foreign words: 'week-end' is maintained in the standardised version, since it appears in the PR14; a word such as 'relou' (lourd, *that's a pain*) from the French inverted 'verlan' form exists in the PR14 so is transcoded as is; an English acronym, 'lol' is in PR14, but the French equivalent, 'mdr' (mort de rire, laughing out loud) is not; the latter is transcoded but the former is not; c) certain syntactical forms are transcoded in order for automatic parsers to be able to process them correctly: a missing 'ne' (in the French, *ne…pas*) will not be indicated, however an ellipsis will be: 'suis arrive' with a 'missing' pronoun (je suis arrivé, *I have arrived*), or an

[2] From a linguistic perspective, the term 'standardisation' is preferable to 'normalisation', the latter being often used in natural language processing references. In the computational paragraphs of this paper we use 'normalisation'.

abbreviated 'oral' form with elision, 't'as' (*you have*) will be transcoded into 'tu as', to allow for automatic processing.

Both the original text message and the transcoded one are available after the process has been completed. This phase is useful for the general public, or those working on SMS analysis and who want to be able to read and compare "raw" and/or processed SMS quickly.

3.2 Our alignment method

The algorithm we propose is based on the *pivot principle* (Choudhury *et al.*, 2007) and uses the results of a normalisation method (Beaufort *et al.*, 2010). We align "raw" anonymised SMSs (called *RS*) with normalised SMSs (called *NS*) according to the following steps:

1. Identification of textual blocks to be aligned,
2. Identification and alignment of invariant blocks (i.e. *pivot* blocks),
3. Deducting alignments based on step 2,
4. Identification of non-aligned blocks and subsequent processing with manual alignments.

In our study, three cases of alignment are highlighted with associated examples:

- *Case 1:* A word in the RS aligned with a single word in the NS ("often as possible").

RS	plus	svt	possible
NS	plus	souvent	possible

- *Case 2:* Two or more words in the RS are aligned with a single word in the NS ("come on leave me alone").

RS	Vasi	lâche	moi
NS	Vas-y	lâche-moi	

- *Case 3:* One word within the RS is aligned with two or more words in the NS ('you had').

RS	T'as		eu
NS	Tu	as	eu

Case 1 is the easiest to handle, since each original word corresponds to a single normalised word. In case 2, we assume that a couple (RS, NS) might be normalised when an aggregation of RS appears. Finally, case 3 needs original processing because one or more couples (RS, NS) are possible (e.g. (T' => Tu), (T' => T, as in 'Tu t'es trompé', *you made a mistake*), etc.). In order to tackle this issue, we propose an approach that generates different *candidate couples*. These are built by taking into account characters of an RS word one by one in order to form blocks of characters. Similarity between blocks and NS is computed. This similarity is based on the Jaccard measure (Geng and Hamilton, 2006). The aim is to select the candidate couple with the highest score.

Let us consider an example.

RS	Jmexquz	
NS	Je	m'excuse

The following comparisons are conducted, one by one, on each character, moving from right to left, according to Fig. 3 below:

a. "z" (last character of RS) is compared to "m'excuse" (NS),
b. "uz" (second-to-last character of RS) is compared to "m'excuse" (NS),
c. "quz" (third-to-last character of RS) is compared to "m'excuse" (NS),
d. "xquz" (fourth-to-last character of RS) is compared to "m'excuse" (NS),
e. "exquz" (fifth-to-last character of RS) is compared to "m'excuse" (NS),
f. "mexquz" (sixth-to-last character of RS) is compared to "m'excuse" (NS),
g. "Jmexquz" (seventh-to-last character of RS) is compared to "m'excuse" (NS).

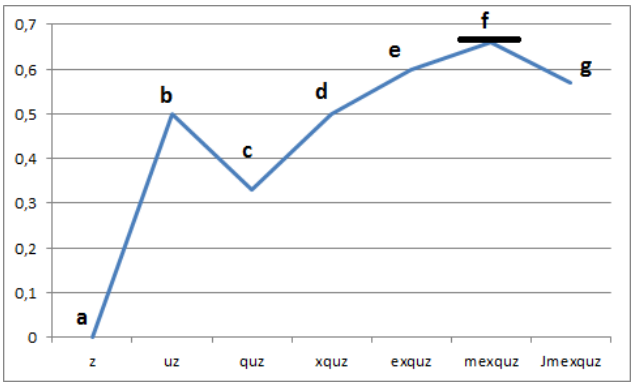

Fig. 3: Comparison function between "Jmexquz" and "Je m'excuse" (in English *I apologise*).

The maximum value is obtained with mexquz (f). This enables to select the candidate ("mexquz", "m'excuse"). This alignment of words enables to construct a specific SMS dictionary presented in the following section.

4. Evaluation

In our study, we randomly extracted 100 SMS (totalising 2,148 words) from the *88milSMS* corpus. For this task, we manually evaluated each normalised word. In our experiment, we obtained:

- 1,814 words that did not require normalisation,

- 186 relevant normalisations (i.e., normalisations which were correctly conducted by the automatic system, e.g. *pr* => *pour* "for"),

- 127 irrelevant normalisations (i.e., normalisations which were incorrectly conducted by the automatic system, e.g. *chambre u* ("University residence room") => *chambre tu* (past form of verb "to be quiet", as initialisms can be ambiguous, and are not necessarily known by the automatic system or the dictionaries),

- 21 words which were "ignored" by the automatic system (because a contextual clue is necessary, e.g., "2" exists as a number but in the following text message is used phonetically in French as a substitution for "of" ("bcp 2 cour", "beaucoup de cours" (lots of lectures).

We use Precision and Recall, which are classic methods of evaluation in text mining.
Precision corresponds to the ratio of the number of words correctly normalised by the system (in this case, 186) and the total number of words annotated by the system (in this case, 313). Recall corresponds to the ratio of the number of words correctly normalised by the system and the total number of words the system should have normalised (in this case, 334).

In these experiments, with the basic algorithm, Precision and Recall are respectively 0.59 (i.e. 186/313) and 0.55 (i.e. 186/334). These weak results justify the necessity to improve the method with several alignment heuristics summarised in this paper. After their application only 6 alignment errors were identified.

These initial experiments with the extract of 100 text messages from the *88milSMS* corpus allow to build a dictionary with 728 words. For each word, different variations are found with the AlignSMS method (e.g., travail → taff, taf ('work'); Montpellier → Mtp, Montpel, montpel, montpellier).

5. Conclusion

In this paper, we have presented an automatic approach for SMS alignment. Based on a comparison function, our algorithms allow splitting "raw" SMSs into sequences in order to align them with normalised SMSs. Using this approach, we generate a freely available specific resource for SMS analysis tasks.

We plan to outline special treatment for punctuation and specific SMS features such as repetition of characters ("Il caaaaaille" (*il caille*, colloquial, "it's cold") and cross-language code-switchings which are not easy to normalise ("aujourd'hui" can appear as "today" in text messages). The final objective is to set up a hybrid (symbolic/statistical approach), based on both grammar rules and our statistical AlignSMS method.

6. Acknowledgements

The SMS data described in this paper was collected within the framework of the *sud4science* project (http://www.sud4science.org). It is part of a vast international SMS data collection project, entitled *sms4science* (http://www.sms4science.org), and was initiated at the CENTAL (Centre for Natural Language Processing, Université Catholique de Louvain, Belgium) in 2004. Our work is supported by the MSH-M (Maison des Sciences de l'Homme de Montpellier, France, http://www.msh-m.fr/), the DGLFLF (Délégation générale à la langue française et aux langues de France, http://www.dglflf.culture.gouv.fr/) and PEPS CNRS ECOMESS (HuMaIn). Many thanks to our colleagues, Catherine Détrie, Claudine Moïse, Bertrand Verine, and to our students (Pierre Accorsi, Frédéric André, Yosra Ghliss, Camille Lagarde-Belleville, Ahmed Loudah, Aghiles Lounes, Michel Otell, Zakaria Mokrani, Namrata Patel, Takfarinas Sider, Anthony Stifani, Tarik Zaknoun), who conducted internships or specific research tasks at various stages of the project.

7. References

Accorsi, P., Patel, N., Lopez, C., Panckhurst, R., & Roche, M. (2012). Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques, *Lingvisticæ Investigationes*, Special Issue: SMS Communication: A Linguistic Approach, John Benjamins, 35:2, 163-180.

Aw, A., Zhang, M., Xiao, J., & Su, J. (2006). A phrase-based statistical model for SMS text normalization. In *Proceedings of the COLING/ACL on Main conference poster sessions* (pp. 33-40). Association for Computational Linguistics.

Beaufort, R., Roekhaut, S., & Fairon, C. (2008). Définition d'un système d'alignement SMS/français standard à l'aide d'un filtre de composition. *Proceedings of JADT 2008*, 155-166.

Beaufort, R., Roekhaut, S., Cougnon, L. A., & Fairon, C. (2010). A hybrid rule/model-based finite-state framework for normalizing SMS messages. In: Hajič, Jan *et al*. (Eds.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, July 11-16, 2010. © 2010 Association for Computational Linguistics, 770–779.

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., & 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.

Cougnon L.-A. (forthcoming, 2014) *Langage et sms. Une étude internationale des pratiques actuelles*. Presses universitaires de Louvain.

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

Geng L., Hamilton, H.-J., (2006). *Interestingness Measures for Data Mining: A Survey*, ACM Comput. Surv., Vol 38(3), Article No. 9

Guimier de Neef, E., Debeurme, A., & Park, J. (2007a). TILT correcteur de SMS : évaluation et bilan qualitatif. *Proceedings of TALN 2007*, 123-132.

Guimier de Neef, É., & Fessard, S. (2007b). Évaluation d'un système de transcription de SMS. In Proceedings, *26th International Conference on Lexis and Grammar, Bonifacio, France, October 2-6, 2007.*

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009). The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, Volume 11, Issue 1.

Kobus, C., Yvon, F., & Damnati, G. (2008). Transcrire les SMS comme on reconnaît la parole. *Actes de TALN 2008*, 128-138.

Panckhurst, R. (2006). Le discours électronique médié : bilan et perspectives, in A. Piolat (Ed.). *Lire, écrire, communiquer et apprendre avec Internet*. Marseille : Éditions Solal, 345-366.

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., et Verine B. (2013). Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS. Épistémè — revue internationale de sciences sociales appliquées, 9 : Des usages numériques aux pratiques scripturales électroniques, 107-138

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (forthcoming, 2014), 88milSMS. A corpus of authentic text messages in French, produit par l'Université Paul-Valéry Montpellier III et le CNRS, avec l'autorisation de 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.

Panckhurst, R., & Moïse, C., (2012b). French text messages. From SMS data collection to preliminary analysis, *Lingvisticæ Investigationes*, Special Issue: SMS Communication: A Linguistic Approach, John Benjamins, 35:2, 290-317.

Patel, N., Accorsi, P., Inkpen, D., Lopez, C., & Roche, M. (2013). Approaches of anonymisation of an SMS corpus. In *Computational Linguistics and Intelligent Text Processing* (pp. 77-88). Springer Berlin Heidelberg.

Shannon C.E. (1948). *A mathematical theory of communication*. Bell System Technical Journal, 27: 379-423, 623-656.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, *13*(2), 260-269.