A translated corpus of 30,000 French SMS

Cédrick Fairon[1], Sébastien Paumier[1,2]

[1] Centre de traitement automatique du langage, UCLouvain, Place Blaise Pascal 1, B-1348 Louvain-la-Neuve
[2] Institut Gaspard Monge, Université de Marne-la-Vallée, 5bd Descartes, F-77454 Champs-sur-Marne
cedrick.fairon@uclouvain.be, paumier@univ-mlv.fr

Abstract

The development of communication technologies has contributed to the appearance of new forms in the written language that scientists have to study according to their peculiarities (typing or viewing constraints, synchronicity, etc). In the particular case of SMS (Short Message Service), studies are complicated by a lack of data, mainly due to technical constraints and privacy considerations. In this paper, we present a corpus of 30,000 French SMS, collected through a project in Belgium named "Faites don de vos SMS à la science" (Give your SMS to Science). This corpus is unique in its quality, its size and the fact that the SMS have been manually translated into "standard" French. We will first describe the collection process and discuss the writers' profiles. Then we will explain in detail how the translation was carried out.

1. Introduction

The development of communication technologies has contributed to the appearance of new forms in the written language: email, chat, forums and SMS (Short Message Service) are a daily source of new codes and abbreviations. Each of these means of communication has its own specificities (synchronous communication vs. asynchronous, extended keyboards vs. small keyboards, one-to-one vs. one-to-many, etc.), but they have in common the fact that most of their users adopt new habits in the way they write (they invent new abbreviations, use non-standard orthographic forms, etc.). This phenomenon has come under the scrutiny of sociologists and linguists, who have started to describe how the language is adapted and how users play with it in order to "make sense" faster (with fewer words; even fewer characters). Scientists involved in Natural Language Processing must also pay attention to this phenomenon because text processing software must be adapted for parsing non-standard texts.

The problem often raised by researchers concerning the study of these new forms of text is the shortage of reference corpora. This is especially true for SMS, for which the text collection is technically more complex than it is for emails, chats and forums. In fact, messages sent from phone to phone are difficult to collect, because it requires the collaboration of either the phone owners - but they are scattered - or the phone companies, but the latter of course have very strict legal regulations. As a consequence, there is no corpus big enough to permit large scale studies. In fact, this was one of the conclusions of an ATALA workshop dedicated to *new written forms* [1].

However, there are recent and ongoing projects aiming at building SMS corpora. For instance, the University of Singapore has built a corpus of 10,000 English SMS (How, 2005). For French, the University of Aix-en-Provence (J. Véronis) has a corpus of more than 1,000 manually transcribed SMS. In Italy, a student project from the University of Torino[2], entitled "SMS Monitor Studies", is intended to collect SMS from web donations (their database currently contains a few hundred messages). In all these cases, the collection was carried out by students and messages were manually copied from phone screens. These are two important limitations: first, these corpora are restricted to a certain category of SMS users (they are all somehow connected to the students: family, friends, etc.) and second, the copying stage may have introduced errors, including typing mistakes or (in-)voluntary corrections.

2. "Faites don de vos SMS à la science"[3]

In order to fill this gap, we organised a SMS collection in the French-speaking part of Belgium, which has lead to the constitution of a large corpus. In order to facilitate the data collection, a toll free short code was made available to the public and a call for participation was broadcast by the main national media (press, radio, television). Participants were invited to send copies of their SMS to the free number and also to fill in an online sociolinguistic form accessible on the Internet, that required the following information: gender, age, native language, other language(s) spoken at home or at work, level of education, employment, home location, work location, frequency of use of SMS, use of a dictionary, recipients of SMS, ability to "decrypt" SMS, SMS writing habits, etc.

From October 2004 to December 2004, we received more than 75,000 SMS donated by more than 3,200 people with different linguistic backgrounds and levels of education, employment, etc. Among them, 2,500 answered our form: they are aged from 12 to 65, divided into 1200 men and 1500 women. They live in 480 different towns, with a regular geographic repartition (there was, however, a high participation in the region around the University of Louvain where the project was organised). Our goal was to build a reference corpus that could serve as a solid base for linguistic studies. It was therefore important to collect samples representing the whole diversity of SMS, written by as large a number of different people as possible.

[1] http://up.univ-mrs.fr/~veronis/je-nfce/resumes.html, visited on August 28, 2005.
[2] http://www.e-allora.net/SMS/ms_index.php, visited on August 28, 2005.
[3] "Give your SMS to Science". This project was lead by the Centre for Natural Language Processing (CENTAL) and the Centre for studies on Roman lexicons (CELEXROM) that are both part of the Université catholique de Louvain (see http://www.smspourlascience.be, http://cental.fltr.ucl.ac.be, and http://celexrom.fltr.ucl.ac.be).

3. Preprocessing the corpus

By the end of the collection campaign, we had received 73,127[4] raw SMS. The first operation was to reassemble messages of more than 160 characters that were split into several SMS[5]. This operation was only partially automated, because each phone operator uses a different splitting method, and we had no access to these protocols. Then, we had to remove SMS that did not conform to the project recommendations (commercial messages, forwarded SMS obviously written by another person, SMS written for the attention of our team, non-French SMS, graphical SMS, etc). At this stage, we also removed duplicates and made some encoding corrections, such as restoring the € symbol or the *c* with cedilla.

The second major preliminary work was to remove personal information form the messages (names, addresses, street numbers, phone numbers, bank account numbers, urls of blogs and personal web sites, etc.). This was not a trivial issue as personal information may appear in many different contexts and can sometimes be very unexpected as in the following (forged) example: *I live in Martin Street, the red house with a green pig above*. In all cases where there was any doubt, data were removed.

In order to preserve the global meaning of messages, personal data were replaced by tags such as `{???,.EMAIL}`[6].

4. Translating the corpus

4.1. Why translate ?

An in-depth study of the first 5,000 SMS we have received has revealed the great variability of word forms (Fairon, Klein, Paumier 2006a). The only limit in variation seems to be the writers' imagination (and the need, sometimes limited, to "make" sense): word spelling variations are totally unpredictable. As a consequence, the exploitation of the corpus is difficult. In fact, when you look into the corpus to find attestations you have to guess at the spelling variants of the word you are looking for. Also, this extreme variation makes reading more difficult, and sometimes even impossible, for the untrained eye. For these two reasons, we have decided to "translate" or "transliterate" the corpus into "standardised" French (this work is discussed in §4.2) and we have built what could be called a bilingual corpora. In this database, each SMS is stored in parallel with its translation in standardised French.

4.1.1. Readability

The difficulty in reading SMS is the first reason for translating the database. In fact, many of the spelling phenomena make reading difficult. For instance, there are SMS that do not contain spaces (or very few) and that mix upper case and lower case letters in a non-conventional manner:

```
BijourMonAmourDiMoiJPeVnirChéToi?GRi1DOtrPr
FerPaséLTpsJusk19hPuiGCorPasé1SuperMatinéeJ
MeSuiFéTrétéDTtLèNomPuiMmSiCPrDodoCLeMèmMsG
TroBzoinDTVoirEnf1SiTuVeBi1?JTM
```

In others SMS, it is the number of non-standard abbreviations and text transformations that make the message difficult to decrypt:

```
Hep.cfè plésir dav 1msg dtoi.g u math l1di
é ca abof éT.ier scienc é Go,alèz.toi oci
bon merd.la jaten lbus é ca gèl.mè d
couch,bonè,gan,écharp...lol.bis a+
```

Also, it is often necessary to understand the codes, usages and habits of SMS writers, which are not necessarily comprehensible for someone who is not familiar with SMS practices. Moreover, some sequences that could have been taken as errors were in fact regular constructions, used throughout the whole corpus. In such cases, having the translation can avoid misunderstandings.

4.1.2. Usability

The second motivation for translating the corpus is that it facilitates the exploration of messages. In fact, the random aspect of word variants makes it impossible to list them *a priori*. Without such a list, you cannot find all the utterances of a given sequence, except if you review the whole corpus, which is tedious. However, it becomes very easy to perform such searches if you have the translation, because you just have to look for the "standard" form of your sequence, and then examine the corresponding SMS. For instance, you can easily find variants of the word *soirée*:

```
Merci. Bisous, bonne soirée...

Bonne swarée et a+??? Bisouxx

Bizzøux bone soiré

G pase bon soire.Now g mal tet.
```

The consultation interface distributed with the corpus offers powerful facilities for expressing complex requests, such as finding all SMS that contain a noun followed by an adjective ending with the *able* suffix.

4.2. Translation protocol

4.2.1. "Standardised" French

We have chosen not to use the technical term "standard French" which has a more normative overtone. The notion of "standard French" is also not very clear and even controversial as it tends to obscure the complexity of language variation phenomena (geographical variation, oral vs. written, language level, etc.). In the context of SMS translation, it is even harder to determine a norm, because some phenomena are typical of this mode of

[4] Participants kept sending messages after the end of the official collection period and at the last count we had reached over 75,000 SMS.

[5] This is what happens when you send a SMS longer than 160 characters; it is split and sent as several single messages of a maximum of 160 characters.

[6] This particular tag format is used because it is compatible with Unitex (Paumier, 2005), an Open Source corpora parser that we have re-used to develop the CD –ROM query interface (Fairon, Klein & Paumier, 2006).

communication and some are borrowed from others. For instance, we find in the SMS language, abbreviations (and many other phenomena) that are typical of spoken language, such as the abbreviation *à toute* in place of *à tout à l'heure*. If we want to "translate" into "standard" French, should we write : *à toute à l'heure* or *à toute* ? As we can see from this example, this is not a simple problem. It was necessary to design a protocol for handling problems one by one. For each problem, a decision was taken by a consensus between the three researchers involved in the project (C. Fairon, S. Paumier and J.-R. Klein).

Several people collaborated [7] in the translation work, but only one person (S. Paumier) was in charge of reviewing the work and ensuring the standardisation and strict respect of the translation protocol. As he is French, he also spotted words and expressions that are typical of the Belgian use of French. These phenomena have also been tagged in the corpus.

4.2.2. Translation rules

The translation has lead to the organisation of the corpus in a six-column data grid:

Name	Content
IdSMS	Index of the SMS in the database
User	Number standing for a GSM number. This information was used for reassembling parts of SMS larger than 160 characters and to link messages with the sociolinguistic profile of the author (if available)
Sex	When this information was available, it was used to check gender agreements, in particular for past participles
Flag	Message annotations (e.g. is case of hesitation about the way to translate)
Message	Original SMS (already anonymised)
Trans.	Translation in "standard" French

For each message, the first step was to check if the preprocessing tasks had been applied correctly: i.e to verify if the content was appropriate (preprocessing should have removed all commercial SMS or SMS addressed to us, etc.) and that there was no more personal information in the message. Then, the translation work was carried out, observing two general rules:

- original SMS are not modified for any reason. If needed, annotation is added in the Flag column;
- protocol must be strictly observed.

The protocol was designed to restore "standard" French and at the same time preserve as much as possible of the original messages. Most of the protocol rules were created before the translation of the corpus, but some were added over time to solve unforeseen problems. Here is the subset of the rule list.

Foreign words: keep them, but correct the spelling if necessary.

```
Sory = Sorry
```

Punctuation marks: restore minus sign and apostrophes.

```
Pa d adresse = Pas d'adresse
j-espère = j'espère
```

Mathematical symbols: keep them if used for their mathematical sense, replace them otherwise.

```
à+ = à plus
Les yankees vienn d'= 5-5 = Les Yankees viennent d'égaliser 5-5
```

Abbreviations: keep them if the abbreviated form is common, otherwise replace. Surround with square brackets in case of doubt.

```
Alors, pour vend. c'est ok? = Alors, pour vendredi c'est ok?
```

Smileys: leave them except when they stand for a word

```
Et oui :-) l'amouuur. = Et oui :-) l'amour.
jte fais un gr :-* = je te fais un gros bisou
```

Spaces and new lines: respect the original SMS and do not restore spaces, except when a word is attached to a number

```
v1 vers19h = viens vers 19h
```

Acronyms and sigla: keep the most common ones (DVD, SMS, etc) or those typical of SMS (lol, mdr, etc), but type them in uppercase.

Letter repetitions: remove them, except when they are essential for onomatopoeia. In that case, limit to 3 letters.

```
aaaaaammouuurr = amour
mmmmmmh = mmmh
```

Phonetic transformations: replace if there is no ambiguity; otherwise, annotate the message.

```
mon ti chéri = mon petit chéri
```

Onomatopoeia and interjections: do not modify the spelling, except in the case of letter repetitions (see above).

Spelling: restore correct forms and when necessary add feminine marks if the writer is a woman. In case of doubt, the word that may be marked will be placed in the annotation column, with the gender mark in parentheses. For instance, if we have:

```
jsui oqp
```

we will translate it into:

```
je suis occupé
```

and put the annotation `occupé(e)` into the Flag column, in order to highlight the doubt about the gender of the writer. The same strategy will be used in any similar case of ambiguity. For instance, if the name `Lauren` is ambiguous between the feminine name `Lauren` and the masculine name `Laurent`, we will write `Lauren` in the translation and put `Lauren(t)` in the Flag column.

Proper names: replace if there is no ambiguity; otherwise, leave them unmodified.

```
caouanne = caouanne
2nise = Denise
```

[7] We have to pay a special tribute to Bernadette Dehottay who dealt on her own with over 25,000 messages.

Numbers: keep them when they actually represent a numeric value; otherwise, convert them to letters.

```
j'ai 1 peu froid=j'ai un peu froid
la paj 84 = la page 84
```

Neologisms: leave them unmodified and explain them with an annotation in the Flag column.

Obvious errors: restore the correct form and annotate the message.

```
que l on ce fiancie = que l'on se fiance
```

Unexpected or incomprehensible symbols: leave them unmodified and annotate the message

```
pgrm>bouf hor concour = pgrm>bouffe hors
concours
```

Character Case: uppercase SMS will be put in lower case, with upper case letters at the beginning of the message and after punctuation marks (! . ?). We will not modify upper case letters used by the writer to separate sentences. All sigla will be put in upper case (`SMS`, `DVD`, etc.) and proper names will be capitalized. Words that are in upper case for emphatic reasons will not be modified.

```
chez pierrette = chez Pierrette
Te voilà oncle ET parrain = Te voilà
oncle ET parrain
SALUT COMAN SA VA = Salut comment ça va
```

Typing errors: if an error is obviously due to a typing error on the phone keyboard, correct it and annotate the message. For instance, if there is:

```
vx tu ke je vienoe
```

translate it into:

```
veux-tu que je vienne
```

and annotate the SMS with `vienoe:vienne` in order to highlight the error.

Missing words: if it is obvious that a word or a group of words is missing, insert a tag in the message in order to highlight this absence. The tag will be as descriptive as possible. In the following example, we can assume that the pronoun `il` is missing:

```
Ici fé cho = Ici {il,.PRO+MISS} fait
chaud
```

When a missing negation cannot be restored because of a contraction (`t'a pas vu mon cd?`), we will not modify the message, but annotate it. If we cannot specify what is missing, we will use the tag `{???,.MISS}`.

5. The corpus

Initially, we had planned to translate the whole corpus, but the task turned out to be much more complicated and time-consuming than expected. So, we decided to limit the translation to 30,000 SMS. This subset is composed of randomly selected messages usually associated with a sociolinguistic profile. But we deliberately added 11% of SMS with no associated profile (in order to avoid the bias of selecting only people who have access to the Internet). This bank of 30,000 SMS contains messages from 1,736 authors with a profile and 700 authors with no profile. It is published in the form of a CD-Rom (Fairon, Klein & Paumier, 2006b) that contains various computer-readable formats of the corpus:

- A raw text file (3Mb in Latin 1).
- A spreadsheet document that includes the columns described in §4.2.2.

The corpus is also distributed as a database linked to a graphical interface that provides tools for searching[8] and sorting original and translated messages as well as author profiles. This user-friendly interface is particularly convenient for linguists and computer non-specialists.

6. Conclusion

This SMS corpus is unique in its size (30,000 SMS), its accuracy (it contains only authentic data collected by electronic means without hand transcription that could alter messages), the number of contributors and the amount of meta-data that are provided (sociolinguistic data and tags highlighting missing words, neologisms, Belgian expressions, etc.). Moreover, SMS have been manually "translated" to create a bilingual corpus in which each message is aligned with its translated version (allowing users to search in standard French and retrieve all the SMS variants). This provides a high added value to the corpus and opens new perspectives for studies of SMS language (Fairon, Klein & Paumier, 2006c).

Acknowledgments

We would like to thank Proximus, Ogilvy and NEWay, our private partners who supported us and helped in the technical preparation of the project. We also thank all the CENTAL members who have made this project possible and in particular Bernadette Dehottay (for her coordination work and the translation of SMS) and Claude Devis (for the creation of the CD-Rom and the development of its interface).

References

Anis, J. (2001). *Parlez-vous texto ?* Paris: Le Cherche-Midi.

Fairon, C., J. Klein, S. Paumier. (2006a). "Le language SMS, preuve d'1compétence". In J.-J. Didier *et alii* "*Le français m'a tuer". Actes du colloque* L'orthographe française à l'épreuve du supérieur. Cahiers du Cental, 1, Louvain-la-Neuve: Presses universitaires de Louvain.

Fairon, C., J. Klein, S. Paumier. (2006b). *Le Corpus* SMS pour la science. *Base de données de 30.000 SMS et logiciel de consultation*. CD-Rom. Cahiers du Cental, 3(2), Louvain-la-Neuve: PUL.

Fairon, C., J. Klein, S. Paumier. (2006c). *"Faites don de vos SMS à la science". Étude du langage SMS au départ d'un corpus informatisé*. Cahiers du Cental, 3(1), Louvain-la-Neuve: PUL. [forthcoming]

Paumier, S. (2002). *Unitex user manual.* [http://www-igm.univ-mlv.fr/~unitex/manuelunitex.pdf].

Pierozak, I. (2003). Le "français tchaté": un objet à géométrie variable. In *Ecrits électroniques, échanges, usages et valeurs*, *Langage et société,* 104, Paris: PUF.

YIjue, H. and K. Min-Yen (2005). Optimizing predictive text entry for short message service on phones. In *Proceedings of Human Computer Interfaces International* (HCII05). Las Vegas, July 2005.

[8] As mentioned above, the search engine is based on the open source software Unitex (Paumier, 2002).