Texto4Science: a Quebec French Database of Annotated Short Text Messages

Philippe Langlais[(1)], **Patrick Drouin**[(2)]
Amélie Paulus[(2)], **Eugénie Rompré Brodeur**[(2)], **Florent Cottin**[(1)]

[(1)] DIRO, [(2)] OLST
Université de Montreal
CP 6128 Succursale Centre-Ville
H3C3J7 Montreal, Québec, Canada
felipe@iro.umontreal.ca, patrick.drouin@gmail.com
http://www.texto4science.ca

Abstract

In October 2009, was launched the Quebec French part of the international sms4science project, called texto4science. Over a period of 10 months, we collected slightly more than 7000 SMSs that we carefully annotated. This database is now ready to be used by the community. The purpose of this article is to relate the efforts put into designing this database and provide some data analysis of the main linguistic phenomenon that we have annotated. We also report on a socio-linguistic survey we conducted within the project.

Keywords: sms4science and texto4science projects, database of Quebec French SMSs, socio-linguistic survey

1. Introduction

Short Text Services are used by a huge community of users for an increasing number of purposes, including advertising (Bamba and Barnes, 2007; Wei et al., 2010), voting (for instance for TV shows), or even political campaigning (Ahvazi, 2004). According to Wikipedia, more than 4 trillion text messages (or SMSs) have been exchanged in 2008. Communicating by short text messages is not anymore reserved to mobile phone users, and is nowadays pervasive on discussion forums, as well as on Twitter;[1] even if the type of device used for typing messages likely influence to some extent the quality of the texts produced.

A large number of works are devoted to study this medium of communication, focussing on various of its aspects, including social ones (Reid and Reid, 2004; Leung, 2006; Wajcman et al., 2007; Baron, 2008), linguistic ones, *e.g.* (Anis, 2001; Cougnon, 2010) as well as educational ones (Scornavacca et al., 2007; So, 2009). We refer the interested reader to the website of the sms4science's project[2] for an extensive list of articles related to SMS.

Because text messaging in particular, and cyberlanguages more generally are becoming ubiquitous, it is natural to see a growing interest in technological aspects related to these medium of communication. Text completion for traditional touchtone phone keypads has been among the first applications studied (MacKenzie et al., 2001). Since then, the development of smart keyboards designed to ease the entering of text on smart phones and other portable devises (e.g. tablets, game consoles, etc.) has evolved quite drastically. The Swipe application[3] as well as the Swiftkey keyboard[4] are two striking illustrations of how fast the mobile phone technology is evolving.

One of the most recently studied application is text message normalization, that is, the transformation of SMS-like texts into their "standardized" version. This has been studied for instance for the English language by (Aw et al., 2006; Choudhury et al., 2007), as well as for the French language, *e.g.* (Yvon, 2010; Beaufort et al., 2010).

Perhaps (or hopefully) more marginally, technologies are deployed in car environment, either for reading SMSs by a text-to-speech synthesis system, as in the *Ford Sync* system, or to assist a driver to answer a message while driving (Ju and Paek, 2010). Also, Munro (2010) describes the service deployed by a consortium of volunteer organizations named "Mission 4636" during the earthquake which stroke Haiti in January 2010. This service routed SMSs alerts reporting trapped people and other emergencies to a set of volunteers who translated Haitian Creole SMSs into English, so that primary emergency responders could understand them. Lewis (2010), in the same context, describes how the Microsoft Translation team developed a statistical translation engine (Haitian Creole into English) in as less as 5 days.

As noted by Fairon et al. (2006), most of the aforementioned studies can only be conducted thanks to the availability of corpora of SMSs in different languages. Such corpora are available in some languages. For instance, the Nus SMS corpus gathers 10 117 English SMSs collected from students of Singapore University that were asked to type (over a webform), messages they received or sent. Some messages acquired from chats complement the collection. A live version of this corpus is also available online (Chen and Kan, 2011); in October 2011, it was gathering 28 724 English SMSs provided by 116 contributors and 29 100 Chinese ones by 515 contributors. Also, the British English SMS Corpora[5] gathers slightly more than 800 SMSs. A number of corpora are also available for the French language (some being proprietary); the largest collection coming from the sms4science project, an

[1] http://twitter.com/

[2] http://www.sms4science.org/?q=fr/node/4

[3] http://www.swype.com/

[4] https://market.android.com/details?id=com.touchtype.swiftkey&hl=fr

[5] http://mtaufiqnzz.wordpress.com/british-english-sms-corpora/

international project coordinated by the Catholic University of Louvain in Belgium. This international project gathers corpora in various languages by asking participants to forward SMSs they already sent. This way of acquiring text messages avoids typos introduced by copying text messages over a webform, and somehow guarantees that the messages sent are real ones. French corpora collected in different regions or countries are already available: Belgium (sms4science), Reunion island (LaRéunion4science), Switzerland (sms4science) and France (smsAlpins as sud4science).

This paper describes the design, acquisition and specificities of the Quebec French corpus of the sms4science project, namely the texto4science corpus.[6] The reminder of this article is as follows. We describe in section 2. the diary of the project as well as our annotation guidelines. In section 3. we give a description of the databases we designed: one consists in annotated SMSs, the other gathers answers of made by contributors to a socio-linguistic survey about SMSs. Section 4. lists a number of projects we are working on that are exploiting this database.

2. Texto4Science

Officially launched in October 2009 (we received our first SMS in November, 23rd 2009), we went through a number of different stages that we detail in section 2.1. We also made a number of annotation choices that depart from other branches of the sms4science project, and that we describe in section 2.2.

2.1. Diary of the project

Money & Approval We received a grant from the Multidisciplinary Center in Emergent Technologies (CITÉ)[7] from University of Montreal in February 2009. This was the starting point of our project. In parallel to this, and since we were dealing with human subject, we had to get our project accepted by the ethical committee of University of Montreal. This turned out to be more complicated than we first thought. We submitted our project to this committee during spring 2009 and obtained a first certificate in June 2009. This certificate had to be modified in order to take into account our policy for recruiting participants. We received a second certificate in September 2009. Finally, the committee disapproved the stickers we printed out for advertising the project and we eventually received a third and last certificate in November 2009.

Technical aspects In order to collect text messages, we rented a phone line with a short number (202202), which impacted our budget significantly (1 500$CA for opening the line, and 370$CA a month). Depending of the SMS plan subscribed by a person, sending a message to this number might cost money. Therefore we tried (from June to September 2009) to obtain from the different telephone operators, as well as from the Canadian Wireless Telecommunications Association[8] the removal of those fees. Although one of this association's main purpose is to promote mobil phone technology, and SMSs in particular, we were not able to come to a satisfactory end. Certainly, this impacted the number of messages we received during the collection phase.[9] One key point in our project was to get hands on the SMSs sent to us. This was done thanks to the help of *Adenyo Télécom Mobile Inc*,[10] an industrial partner who kindly provided us a technical platform for easing the management of the messages we received.

Advertisement Attracting participants to the project was much more complicated than we initially thought, and different strategies have been tried. We first contacted the communication services of University of Montreal which helped us to promote locally the project. In particular, 60 letter-format posters of the project have been put at several strategical points within the University. We also animated on a daily basis a discussion group on Facebook[11] and Twitter, two major social networks. One student was recruited for assisting in this time consuming task, otherwise conducted by the second author of this article. We also printed stickers that we distributed into strategic places in Montreal, such as cafes, pubs, universities, high schools and the like. This costed us approximatively 600$CA as well as quite a lot of energy for distributing them. To further encourage the forwarding of SMSs, we randomly selected each week a winner to whom we offered a gift, such as a prepaid phone card. This turned out to be difficult to organize since this is assimilated to a lottery, which poses legal issues we had to go through.

After a number of initiatives, the project eventually became relayed by medias. Patrick Drouin presented the project in 4 radio and 2 TV shows. 6 articles in Montreal newspapers also related the project, so did a tenth of blogs maintained by journalists or institutions (e.g. Radio Canada). Retrospectively, 2 events had a strong impact on the number of messages we received. The first one was when Patrick Drouin was invited in November, 26th 2009 to Radio Canada for a popular radio-show called "Christiane Charette".[12] He was accompanied by *Biz*, the singer of the popular hip hop group from Quebec: *Locolocass*.[13] Among other things, this group defends the role of the French language in Canada; which explains why *Biz* spontaneously accepted to assist us in promoting the project. This event corresponds to the first peak in the number of messages we received over time, as clearly shown in Figure 1. The second peak we observe was when Patrick Drouin was interviewed at a popular TV show called *Salut, Bonjour* on TVA channel (March, 4th 2010).

Annotation We recruited two students (who co-authored this paper) at the Linguistic department of University of Montreal. They were in charge of annotating the mes-

[6] SMSs are called *textos* in the French speaking part of Canada.

[7] http://www.cite.umontreal.ca/

[8] http://www.cwta.ca/

[9] People to whom we asked to send SMSs often told us they won't because of the fees associated to a call to our short number.

[10] http://www.adenyo.com/

[11] http://www.facebook.com/

[12] http://www.radio-canada.ca/emissions/christiane_charette/2009-2010/chronique.asp?idChronique=97193

[13] http://fr.wikipedia.org/wiki/Loco_Locass

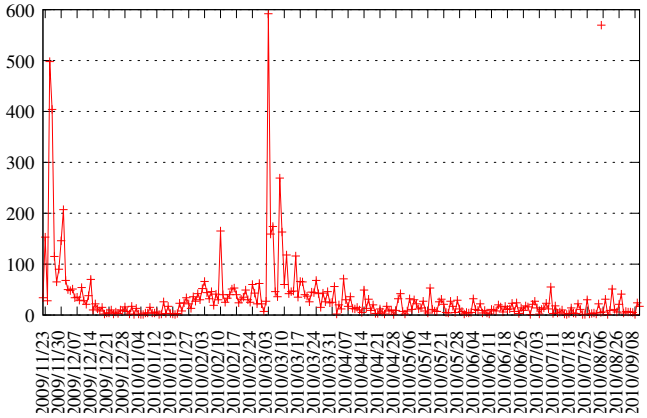

Figure 1: Number of SMSs received over time.

sages we received. The annotation took place in roughly 3 stages. In a first one, they annotated a number of messages, using plain text annotation, typed with `Excel` as an interface, and without much instructions regarding the way messages should be annotated. This stage only served to apprehend the various phenomenon encountered in SMSs. In a second stage, we analyzed those phenomenon[14] and decided to adopt XML for annotating SMSs. We developed an XML schema that was used within the friendly XML editor <oXygen/>.[15] In a third stage, we refined the XML schema used for annotating, and the annotators revised the annotation they produced.

Corpus Production A bachelor student in computer science (who co-authored this paper) joined the team for 3 months and developed XML tools for preparing the 7 274 SMSs that were available at that time. The anonymisation of the messages and the spotting of several inconsistencies in the database were the most consuming part of the transformation process. He also developed a prototype which transforms a French message into a likely SMS form.

2.2. Annotation policy

Figure 2 illustrates the annotation provided in the Belgium `sms4science` database, which basically takes the form of an `Excel` spreadsheet, where one column stands for the original message and another one stands for the normalization or transcription. While most words are normalized, it happens that some are not. In this example, the form `bizo` is not normalized because it is marked as ambiguous. Also, the form `jtm`, which very likely stands for `je t'aime` (I love you) is not normalized either.

Our annotation scheme departs from the one just described in several aspects. First, we decided to produce an XML database. This allows a clean separation of the text data from the annotation, which is not the case of the format shown in Figure 2. It also facilitates the alignment of the normalized form with the original SMS. Last but not least, it allows the validation of the database against an XML schema (or a DTD), as well as manipulating the database with XSLT transformations. Those last two points proved

[14] This is documented in (Drouin et al., 2010).

[15] `http://www.oxygenxml.com/`

original: `Ccou princes jpens trop a ti et tme mank tro.. Jesper kon svoi bi1tot.. Gro gro bizo jtm`

annotation: `Coucou princesse je pense trop à toi et tu me manques trop.. J'espère qu'on se voit bientôt.. Gros gros {bizo,.AMBIG} jtm`

Figure 2: Example of an annotated SMS in the Belgium `sms4science` corpus. `AMBIG` indicates that the form `bizo` is ambiguous. This very likely stands for the French word `bisous` (kisses).

to be very useful when came the time for checking and anonymizing the database. Second, we extended the number of linguistic phenomenon annotated. We decided to provided as much annotation as possible for easing the normalization of SMSs into proper French. We feel this leaves more opportunities for fine-grained linguistic analyses. On the downside, however, we must admit that the annotation scheme we designed is much more complex to apprehend, and that dealing with XML data for a non XML literate person can be tedious.

Figure 3 provides an example of an SMS in the `texto4science` database. Each message (element `texto`) has a specific identifier (attribute name `ID`), and is stamped by the date at which it has been received (attribute `date`). The donator identifier (element `user_id`) indicates which person sent this message; it can be used to consult the sociolinguistic information provided by the person (if any). The text message received is encoded in the `orig` element, after anonymisation took place (replacement of names, phone numbers and the like by generic names). The `transcrip` element contains the annotation of the original message. There are basically two families of elements that are used to annotate an SMS: those that denote a fact (such as a missing punctuation), and those that denote errors (such as typos), in which case the `forme` attribute contains the correct text. Last, the `norm` element contains the normalization of the SMS. It has been automatically produced by applying XSLT transformations to the `transcrip` element.

2.3. Annotation schema

In accordance with the annotation conducted within the `sms4science` database, we annotated missing negations (element `negat`). An example of annotation is provided in Figure 3 where the French negative particle `ne` is absent, as is often the case in spoken language. We also decided to annotate missing punctuations (element `ponc`) because text messages are known to miss punctuation signs, which might pose problems for instance to SMS-to-speech converters.

We also annotated the presence of a number of specific units encountered in SMSs. For instance we marked each abbreviation we found along with their plain form (element `abrev`). Abbreviations are one of the main characteristics of SMSs, and we analyze them in some de-

```
<texto ID="2009120927" date="2009-12-09">
   <user_id>user_111250</user_id>
   <orig>
        Salut Florent, je sais pas ou tu es, mais si tu peux connecte toi sur
        Skype, j y serai une partie de la soiree! Ciao
   </orig>
   <transcrip>
        Salut <prenom sexe="masc">Florent</prenom>, je<negat forme="ne"/>
        sais pas <ortho forme="où">ou</ortho> tu es, mais si tu peux
        <ponc forme=","/><typog forme="connecte-toi">connecte toi</typog>
         sur Skype, <typog forme="j'y">j y</typog> serai une partie de la
        <ortho forme="soirée">soiree</ortho>! Ciao<ponc forme="."/>
   </transcrip>
   <norm>
        Salut Florent, je ne sais pas où tu es, mais si tu peux, connecte-toi
        sur Skype, j'y serai une partie de la soirée! Ciao.
   </norm>
</texto>
```

Figure 3: Example of an annotated and anonymized SMS. We slightly edited this XML instance for the sake of readability.

tails in section 3.1. Similarly, we annotated symbols (element `symb`) that often serve as abbreviations (e.g.. `@` for the word `at`), but that can be used as punctuation marks as well. Smileys are typical of SMSs and are being marked (element `binet`), as well as various marks used for indicating laughing (element `rire`), such as `ah ah` or `Mouahahah!!`. First names (element `prenom`), family names (element `nom`), numbers (element `numero`) as well as email addresses (element `mail`), url (element `web`) and normal addresses (element `addresse`) are marked as well, in order to facilitate the anonymisation of the messages. Also, parts of text messages that are in another language are marked by the `bloc_lang` element along with the language being recognized (attribute `langue`).

On top of describing typical units that are present in (or absent from) text messages, we also annotated some errors and their correction (via the attribute `form`). Different kind of errors are encoded with different element types. Typos are marked by the element `coquille` (e.g. `trsiste` instead of `triste` (sad)), while typographical errors are tagged by the `typog` element (e.g. `repose toi` instead of `repose-toi`). Spelling errors are marked by the element `ortho`, as in `nous meme` for the form `nous-même` (ourselves). Syntactical errors — e.g. `idée cadeau` (idea gift), instead of `idée de cadeau` (idea for a gift) — are tagged with the `synt` element. Also, the `accord` element indicates an agreement error, e.g. `Dit-lui` for `dis-lui` (tell him). We also annotated each error involving casing (element `majus`). Most often, they concern the absence of a capitalized letter at the beginning of a message or a proper name. Last, there were some forms we could not transcribe. We tagged `forme_inconnue` and `element_inconnu` each word-form and symbol we could not interpret respectively.

The frequency of each element in our database is provided in Table 1. Clearly, the SMSs we received are characterized by a high rate of missing punctuations (`ponc`), a large number of spelling errors (`ortho`) and a lot of abbreviations (`abrev`). Slightly more than 35 000 annotations are present in the current version of the database, that is, roughy 5 annotations per message.

element	count	element	count
ponc	9027	rire	622
ortho	7496	coquille	267
abrev	5082	symb	183
synt	2924	forme_inconnue	121
majus	1738	nom	112
binet	1605	element_inconnu	28
accord	1505	numero	21
bloc_lang	1480	adresse	17
prenom	1038	mail	1
typog	1006	web	1
negat	818		

Table 1: Counts of the 21 annotation types in our database.

Note that we decided to use non recursive XML elements in our annotation schema. This means that only one element can be associated to a given portion of a text message. For instance, an abbreviation in English such as `asap` (`as soon as possible`) is annotated as being an English bloc of text (element `bloc_lang`) in our database, but not as an abbreviation. Sometimes, the attribute `comment` is used to document such situations.

3. The database

The `texto4science` database takes the form of a tar file composed of three files: an XML file encoding the anonymized and manually annotated SMSs, an XML file encoding the answers of contributors to a socio-linguistic survey, as well as a bunch of tools that facilitate the treatment of both databases.

3.1. The Database of SMSs

At the time of writing, we treated a total of 7 274 text messages, sent by 360 different persons (or more exactly phones). We received 420 (5.8%) SMSs written in English, 6 written in Spanish, 1 written in Italian and 5 in other languages. Those SMSs are part of the database, but did not receive any annotation (apart that they are written in a language other than French). The main characteristics of the SMSs we annotated are reported in Table 2.

	SMS	normalized
Number of tokens	90 298	104 268
Number of types	11 750	9 279
Number of hapax	7 215	5 401
	char	word
Compression rate	10.5%	13.4%

Table 2: Main characteristics of the 6842 SMSs written in French. A simple set of regular expressions has been applied in order to tokenize the material.

It is often believed that SMS messaging is geared toward shorter texts, due in part to the length limitation applied by most operators. We observed in our database that the compression rate is only about 10%; the average SMS length being 58 characters, while their transcriptions are 65 character long on average.

One possible use of an annotated corpus consists in compiling a dictionary dedicated to SMSs. Some are already available, such as `www.dictionnaire-sms.com/`, `www.sos-sms.net` or `www.deblok.net/dicosms`, but are nevertheless never large enough to account for the great creativity in SMSs. Cougnon and Beaufort (2010) present a methodology to semi-automatically build up a dictionary out of an SMS corpus. Building such a dictionary from our database is simply a matter of querying `abrev` elements. As an illustration of this, we collected thanks to an xPath query, the 10-most frequent forms abbreviated in our database, as well as their 3 most common abbreviations. This is reported in Table 3.

We observe that some abbreviations are ambiguous, such as `dispo` which stands for both `disponibilités` (availability) and `disponible` (available) or the form `1` which stands both for `une` (a, feminine) and `un` (a, masculine). The most ambiguous abbreviation we found is `txt` which is used for various morphologically derived forms of the word `texte` (text): `texte`, `texteras`, `texté`, `textes`, `texter`, `texto`.

We also noticed a great variability in common expressions such as `à plus tard` (see you latter) which we found to be written as `a+` (9), `aplus` (2), `a pluche` (1), `a plus` (1), `aplush` (1), `a+` (1) and `à plus` (1), where the figures in parentheses indicate the frequency of the form in the database. Similarly, we found 6 different ways of writing the word `demain` (tomorrow): `2m` (3), `dmain` (2), `2mai` (1) which is likely a typo, `2main` (1), `2min` (1), `dmin` (1). Fairon et al. (2006) noticed as well a large number of variants for this word.

As we shown, 1605 smileys (element `binet`) have been annotated. We tagged a total of 98 different smileys in our corpus, this is much less than the 900 ones observed in (Beaufort et al., 2010). The 3 most frequent ones are `:)` (37.3%), `:P` (6.8%) and `:(` (6.2%). Some less common smileys are nevertheless creative such as `>;->` (which we observed 3 times), or `(>-_-<)` which we observed only once. In the same vein, we annotated 622 laughing marks in our database (element `rire`), for a total of 113 different forms; many being variants of the same form (e.g. `hahahaha`, `hahah`), others being more surprising (e.g. `L0olehw'zz`).

Perhaps one characteristic of our database is the high amount of annotations it gathers. We already mentioned that the annotated SMSs (those that are at least partially written in French) have an average of over 5 annotations. Table 4 provides the

of the number of annotations per message. It is interesting to note that only 258 of them (3.8%) do not have any annotation, which means that they are written in standard Quebec French, without any noticeable error or SMS-like idioms (see Section 3.2.). On the other hand, one message received no less than 86 annotations. It is reported in Figure 4, along with its transcription. As we can see, this is a rather long message: 563 character long, whilst the limit imposed by most SMS protocols is 160 characters.[16] We pinpoint that the transcription of this message contains several forms that are typical of Quebec French (the most obvious being marked in bold). In the current version of the

form	freq	Top-3 abbreviations					
okay	310	ok	(251)	k	(56)	okk	(3)
c'est	283	c	(270)	c'	(5)	cé	(5)
dans	174	ds	(168)	dan	(4)	dn	(2)
que	141	ke	(128)	q	(13)		
pour	120	pr	(120)				
avec	92	ak	(48)	aek	(22)	ac	(11)
tu es	88	t	(69)	te	(6)	té	(5)
parce que	82	pcq	(51)	pck	(20)	paske	(6)
il	71	y	(70)	ii	(1)		
vous	71	vs	(70)	vou	(1)		

Table 3: 10 most frequently abbreviated forms and their 3 most frequent abbreviations.

n_a	n_s	n_a	n_s	n_a	n_s	n_a	n_s	n_a	n_s
0	258	5	621	10	189	15	54	...	
1	978	6	531	11	143	16	49	44	1
2	896	7	376	12	113	17	37	46	2
3	901	8	317	13	102	18	21	51	1
4	825	9	241	14	74	19	19	86	1

Table 4: Number of SMSs (n_s) with a given number of annotations (n_a).

[16] Some operators offer the possibility to split a message into several ones, we regrouped those messages into a single one whenever possible before annotating them.

database, those forms are not annotated as such. This is planed as future work. Still, the use of words (most often verbs) borrowed from the English language are marked as such, as the verb `feeler` which is borrowed from the English verb `to feel`.
Our database deserves a more systematic analysis of its content than we can afford in this paper.

3.2. Database of contributors' profile

3.2.1. The survey

Volunteers who gave their SMSs to the `texto4science` project were invited to fill up a webform containing 23 questions designed for collecting their profile. The answers provided are organized into an XML file which is part of the data we distribute. For obvious reasons, some information has been withdrawn from the database, such as phone numbers. Still, it is possible to cross this database with the one of SMSs, since contributors have been serialized similarly in both databases. However, we noticed that a third of the responders did provide a phone number different from those we collected with the SMSs. The questions of the webform are grouped into 4 main categories:

Personal information such as age, gender, postal code, mother tongues, spoken languages, educational level.

Usage of SMSs average number of SMSs send per week, usual places were they send or receive messages, categories of persons (friends, relatives, etc.) to whom most messages are addressed, etc.

Abilities in writing SMSs familiarity of responders with abbreviations and other codes frequent in text messages; their use of such idioms in their production; their tendency to mix several languages in a single SMS.

Technical device kind of device subjects are mostly texting from (12 touchtone pad, qwerty keyboard, tablets, etc.); the use of completion or correction tools, if their device provides such facilities.

3.2.2. Analysis

In the following, we analyze part of the questions asked to the 298 contributors who answered the survey. This analysis is articulated along the four dimensions aforementioned.

Personal information A total of 298 persons responded to the survey (63% females); 209 of them provided a telephone number corresponding to one we collected in the database of SMSs. Those 209 responders represent 58% of the persons who gave their SMSs to the project. The responders aged 27 on average, the youngest person was 12 year-old, the oldest 65. Slightly less than 30% (89) of the responders said they were students. 60% (177) of the responders received a university education (45 of them at a graduate level), and 26% (77) went to college. Most responders (284) have French as their mother tongue, which is not surprising since we targeted SMSs written in Quebec French; 75% (224) persons also mentioned they speak English currently. A few responders live abroad Canada.

Since when are you using SMSs?						
	≤6m	≤1y	≤2y	≤3y	≤4y	>4y
	18	23	65	53	39	36

How many SMSs a week?						
	<5	≤10	≤20	≤50	≤100	>100
in	135	63	46	36	13	5
out	140	58	48	32	15	5

Whom are you writting SMSs to?						
	fam	*friend*	*lover*	*col*	*compet*	*other*
|resp.|	250	291	201	208	79	91
rank 1	38	151	111	12	2	4
avg.	2.7	1.7	1.8	3.2	4.9	4.5

Why are you using SMSs?						
	tel	*cost*	*info*	*app*	*contact*	*chat*
|resp.|	279	223	275	243	253	242
rank 1	145	58	55	36	47	56
avg.	2.2	3.3	2.7	3.3	3.1	3.3

Where are you composing or reading your SMSs?				
	home	*job*	*public*	*transp*
|resp.|	286	268	294	275
rank 1	85	124	89	87
avg.	2.5	2.1	2.1	2.4

Table 5: Five questions regarding the use of SMSs. See the text for more.

We analyzed the distribution[17]. Clearly, many responders are located in Montreal, and a significant part are located downtown. This underlines the difficulty we had at motivating people to donate their SMSs to the project. We are currently conducting a collection of Canadian English SMSs all over Canada.

Usage of SMSs The way responders are using the SMS technology is summarized in Table 5. Only 6% (18) of the responders are new users of SMSs (less than 6 months of usage); while 12% (36) are used to it since at least 4 years. Regarding the number of SMSs received and sent (obviously both figures are highly correlated), a significant portion of responders (47%) are dealing with a few SMSs a week only (less than 5 messages received and sent). Only 5 responders were sending (and receiving) more than one hundred messages a week.
The three other questions detailed in Table 5 were formulated as multiple-choice questions. Each responder could rank each option, a score of 1 being associated to the option the most appropriate and a score of 5 to the less appropriate one. Options not relevant were marked as such. For each of those questions, Table 5 reports three lines: $|resp.|$ indicates the number of responders who marked a specific option as relevant, *rank 1* indicates the number of responders that gave an option the first rank, and $avg.$ indicates

[17] See the Google map at `http://rali.iro.umontreal.ca/texto4science/` over Quebec of our responders

<orig>Scorect.. Tu tdoute ke sa ma vrm fait de koi paske jtai pas recrit depuis alors ke shu kk1 ki oublie facilement dhabitude.. Avant kon passe Par-dessus jvx juste te dire ke si sa ma autant fait de koi c paske oui je le sais ke c vrm un probleme sam mnuit vrm d fois cpour sa ke jvx Vrm le regler, pis ski ma fait chier c ktu mdise sa au moment ou jtai dit ke javais fais d efort dernierement pis ksa sameliorait.. Mais cte Soir la jdevais pas feeler en partant, jc pas pk jy ai tellement pense.. Javais vrm envie pleurer pis dparler a kk1.. Tk jtaime fort fort xX< /orig>
<norm>**C'est correct**... Tu te doutes que ça m'a vraiment **fait de quoi** parce que je ne t'ai pas récrit depuis, alors que je suis quelqu'un qui oublie facilement, d'habitude. Avant qu'on **passe par-dessus**, je veux juste te dire que si ça m'a autant **fait de quoi**, c'est parce que oui, je le sais que c'est vraiment un problème. Ça me nuit vraiment des fois, c'est pour ça que je veux vraiment le régler, **pis** ce qui m'a fait chier, c'est que tu me dises ça au moment où je t'ai dit que j'avais fait des efforts dernièrement **pis** que ça s'améliorait. Mais ce soir-là, je ne devais pas **feeler en partant**, je ne sais pas pourquoi j'y ai tellement pensé... J'avais vraiment envie de pleurer **pis** de parler à quelqu'un... En tout cas, je t'aime fort fort. xX< /norm>

Figure 4: The message with the highest number of annotations in the texto4science database. Forms in bold are typical of Quebec French.

the average score per option (not counting the options that were judged irrelevant).

For instance, 291 responders (that is, most of them) indicated that they are sending SMSs to their friends, while 151 persons ranked this usage first; the average rank of this option is rather high: 1.7. All those figures contribute to indicate that responders in great part are sending SMSs to their friends or lovers, which is not entirely surprising.

When asked about their motivations for using the SMS technology, most responders mentioned they are using it mainly as a replacement of emails and telephone calls (tel). 75% (223) of them indicated that reducing the cost of their mobile phone bill was a concern ($cost$), although this was ranked only 3.3 on average; only 19% mentioned this was their first motivation. Exchanging information ($info$), fixing appointments (app), keeping contact with friends ($contact$) and chatting ($chat$) were options that were mostly ranked by responders.

Last, it is interesting to note that people are reading or writing SMSs in various places, such as home, job (or school), transportation and public places, and that 42% (124) of them are doing so preferably at their job.

Abilities in writing SMSs Only half of the responders (151) mentioned that it was easy for them to understand the abbreviations typically encountered in SMSs. Code switching is a common practice among responders: 73.5% (219) of the responders mentioned they switch from one language to another from time to time. Although English is the language they switch to most frequently, other languages are being used as well, among which Spanish, German, and Arabic are the most popular ones.

Technical device Technical aspects related to the way people write SMSs are summarized in Table 6. First, it is noticeable that our responders are not making a great use of dictionary facilities: 16% (47) only are making use of suggestions proposed by such tools, but often, many do not or only occasionally. 17% (50) even do not know about this technology (those marked *??*). This is certainly related to the kind of technical device they are using. In fact, 50% (148) are using a standard 12-key keypad (*regul*), 31% (92) are using a qwerty keypad and only 15% (44) are using tactile tablets. The kind of device used for texting is likely evolving fast, and the impact this evolution has on the quality of the SMSs produced deserves some investigations in which our corpus will likely be useful. We have noticed that several responders mentioned that typing accents on tablets is difficult (often, it requires to switch the keyboard), and that typing with a QWERTY keyboard reduces the use of SMS-like idioms.

Are you using a device with a dictionary?					
no	*??*	*always*	*often*	*sometime*	*never*
83	50	11	36	44	74

Which keyboard are you using?				
regul	*qwerty*	*tact*	*stylet*	*other*
148	92	44	3	11

Table 6: Technical aspects of SMS writing. See the text for more.

4. Discussion

We have presented an overview of the texto4science project and its database which is freely available for download at URL:

http://rali.iro.umontreal.ca/texto4science/

Facilities for navigating online trough the database are currently being built and will be available as well. We are currently working on several projects which are making use of this database. First, we are developing two translation engines, which current state is available online. The first

one transforms a French text into a SMS-like text. The second one normalizes SMSs according to a statistical translation engine we trained on the texto4science database. This statistical engine is hybridized with rules that are designed to handle specific phenomenons such as agglutinations, e.g. Ca'c'peupa (this cannot be true). Also, we observed that SMSs are often used for scheduling appointments. Therefore we developed a system for recognizing appointments in SMSs and extracting their pertinent information (date, places, etc.).

On top of those applications, there is a number of issues that we plan to address. First, we want to extend the markup language we used to annotate the SMSs in order to account for phenomenon that should be handled, such as Quebec French expressions. This would ease the comparison of the texto4science database with other databases for the French language. Preliminary investigations are indicating that they are numerous in our database and deserve specific annotations. Second, we are still receiving SMSs that lack annotation. It is our intention to update the current database with those new messages, possibly by semi-automatically annotating them thanks to the data we already annotated. Finally, we observed a number of arguable annotation choices we would like to correct in future versions of this database.

Acknowledgments

We are deeply indebted to the *Adenyo Télécom Mobile Inc* who kindly provided us at no cost the platform we used for collecting our SMSs. We are also grateful to Fabrizio Gotti, Benoît Robichaud and Annie Desnoyers for their implication in the texto4science project and to Aurélie Picton and Fabienne Venant for their involvement in publicizing it. This work has been made possible thanks to the monetary and infrastructural support of the Multidisciplinary Center in Emergent Technologies (CITÉ) from University of Montreal.

5. References

D. Ahvazi. 2004. Text messaging and its effects on everyday life.

J. Anis. 2001. *Parlez-vous texto ? Guide des nouveaux langages du réseau*. Le Cherche Midi.

A. Aw, M. Zhang, J. Xiao, and J. Su. 2006. A phrase-based statistical model for sms text normalization. In *COLING/ACL on Main conference poster sessions*, pages 33–40.

F. Bamba and S. J. Barnes. 2007. Sms advertising, permission and the consumer: a study. *Business Process Management Journal*, 13:815–829.

N. S. Baron. 2008. *Always On: Language in an Online and Mobile World*. Oxford University Press.

R. Beaufort, S. Roekhaut, L.-A. Cougnon, and C. Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing sms messages. In *48th Annual Meeting of the Association for Computational Linguistics*, pages 770–779, Uppsala.

T. Chen and M.Y. Kan. 2011. Creating a live, public short message service corpus: The nus sms corpus. *CoRR*, abs/1112.2468.

M. Choudhury, R. saraf, V. Jain, A. Mukherjee, S. Sarkar, and A. Basu. 2007. Investigation and modeling of the structure of texting language. *International journal on document analysis and recognition*, 10:157–174.

L.-A. Cougnon and R. Beaufort. 2010. Ssld: a french sms to standard language dictionary. In *ELexicography in the 21st century. New challenges, new applications*, Cahiers du Cental, 7. Presses Universitaires de Louvain. To appear.

L.-A. Cougnon. 2010. Orthographe et langue dans les sms. *Ela. Études de lingusitique appliquée*, 160:397–410.

P. Drouin, E. Rompé Brodeur, A. Paulus, and P. Langlais. 2010. Guide d'annotation: Projet texto4science. Technical report, Université de Montréal, June.

C. Fairon, J.-R. Klein, and S. Paumier. 2006. A translated corpus of 30 000 french sms. In *LREC*, pages 351–354, Genova.

Y.-C. Ju and T. Paek. 2010. Using speech to reply to sms messages while driving: An in-car simulator user study. In *ACL 2010 Short paper session*, pages 313–317, Uppsala.

L. Leung. 2006. Unwillingness-to-communicate and college student's motives in sms mobile messaging. *Telematics and Informatics*, 24:115–129.

W.D. Lewis. 2010. Haitian creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *EAMT*, Saint-Raphael.

S. MacKenzie, H. Kober, D. Smith, T. Jones, and E. Skepner. 2001. Letterwise: Prefix-based disambiguation for mobile text input. In *ACM Symposium on User Interface Software and Technology*, pages 111–120, New-York.

R. Munro. 2010. Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation*, Denver.

D. Reid and F. Reid. 2004. Insights into the Social and Psychological Effects of SMS Text Messaging. University of Plymouth Working Paper.

E. Scornavacca, S. Huff, and S. Marshall. 2007. Developing a sms-based classroom interaction system. In *Conference on Mobile Learning Technologies and Applications*, pages 47–54, Auckland.

S. So. 2009. The development of a sms-based teaching and learning system. *Journal of Education Technology Development and Exchnge*, 2(1):113–124.

J. Wajcman, M. Bittman, P. Jones, L. Johnstone, and J. Brown. 2007. The impact of the mobile phone on work/life balance. Technical report, Australian Research Council.

R. Wei, H. Xiaoming, and J. Pan. 2010. Examining user behavioral response to sms ads: Implications for the evolution of the mobile phone as a bona-fide medium. *Telematics and Informatics*, 27(1):32–41.

F. Yvon. 2010. Rewriting the orthography of sms messages. *Natural Language Engineering*, 16(2):133–159.