A digital corpus resource of authentic anonymized French text messages: 88milSMS—What about transcoding and linguistic annotation?

Rachel Panckhurst
Praxiling, Université Paul-Valéry Montpellier 3, France

Abstract

In 2011, six academics gathered over 90,000 authentic text messages (SMS) in French from the general public, in compliance with French law (http://sud4science.org, Panckhurst *et al.*, 2013). The SMS 'donors' were also invited to fill out a sociolinguistic questionnaire (see Figure A1, Moïse, 2013, Panckhurst and Moïse, 2014). The 'sud4science' project is part of a vast international initiative, entitled 'sms4science' (http://www.sms4science.org/, Fairon *et al.*, 2006, Cougnon and Fairon, 2014, Cougnon, 2015), which aims to build a worldwide database and analyse authentic text messages in different languages. After the 'sud4science' SMS data collection, a pre-processing phase of checking and eliminating any spurious information and a three-step semi-automatic anonymization phase were conducted (Accorsi *et al.*, 2014, Patel *et al.*, 2013). Two extracts were transcoded into standardized French (1,000 SMS) and annotated (100 SMS). The finalized digital resource of 88,000 anonymized French text messages, the '88milSMS' corpus, the extracts and the sociolinguistic questionnaire data are currently available for all to download, from the Huma-Num web service (http://88milsms.huma-num.fr, Panckhurst *et al.*, 2014). The 88milSMS corpus has also recently become available via a Creative Commons Attribution 4.0 International licence on the 'Ortolang' platform (https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1, Panckhurst *et al.*, in Chanier (ed), 2016). In this paper, first the authors briefly situate the project and describe the anonymization process. Then, they focus on why they decided to exclude full 'transcoding' and linguistic annotation in the first version of the final corpus.

Correspondence:
Rachel Panckhurst, Praxiling, UMR5267 CNRS, Université Paul-Valéry Montpellier 3, Route de Mende, 34199 Montpellier cedex 5, France.
E-mail:
rachel.panckhurst@univ-montp3.fr

1 Introduction

The 'sud4science' project (http://sud4science.org; Panckhurst *et al.*, 2013) is part of a vast international initiative, entitled 'sms4science' (http://www.sms4science.org/; Fairon *et al.*, 2006; Cougnon and Fairon, 2014; Cougnon, 2015), which aims to build a worldwide database and analyse authentic text messages in different languages—mainly French, but also Creole, Swiss German, Standard German, Italian, Romansh (Dürscheid and Stark, 2011), and English (Guilbault and Drouin, 2016). Several related SMS

Digital Scholarship in the Humanities, Vol. 32, Supplement 1, 2017.

doi:10.1093/llc/fqw049 Advance Access published on 16 November 2016

data collections have taken place since the initial Belgian one: Reunion Island (20,000 SMS, 2008, http://www.lareunion4science.org/; Cougnon and Ledegen, 2010), Switzerland (24,000 SMS, 2009–10, http://www.sms4science.uzh.ch; Dürscheid and Stark, 2011), Quebec (5,000 SMS, 2010, http://www.texto4science.ca/; Langlais *et al.*, 2012), French Rhône-Alps (22,000 SMS, 2010, http://www.alpes4science.org/; Antoniadis *et al.*, 2011), and British Columbia (14,300 SMS, 2012, http://www.text4science.ca/; Drouin and Guilbault, 2016). The 2011 French text message collection from the general public in the south of France lasted 3 months (http://sud4science.org; Panckhurst *et al.*, 2013), and SMS 'donors' were also invited to fill out a sociolinguistic questionnaire (see Figure A1; Moïse, 2013, Panckhurst and Moïse, 2014).

After the 'sud4science' SMS data collection, a pre-processing phase of checking and eliminating any spurious information (including duplicates, advertisements, messages from telephone operators, etc.) and a three-step semi-automatic anonymization phase were conducted (Accorsi *et al.*, 2014, Patel *et al.*, 2013). Two extracts were 'transcoded' into standardized French (1,000 SMS) and linguistically annotated (100 SMS).

In June 2014, the finalized digital resource of 88,000 anonymized French text messages, the '88milSMS' corpus, the extracts, and the sociolinguistic questionnaire data were made available for all to download, from the Huma-Num web service (http://88milsms.huma-num.fr; Panckhurst *et al.*, 2014). The 88milSMS corpus has also recently become available via a Creative Commons Attribution 4.0 International (CC BY 4.0) licence on the 'Ortolang' platform (https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1; Panckhurst *et al.*, in Chanier (ed) 2016).

In this article, after briefly evoking the anonymization process, I focus on why we decided to exclude full 'transcoding' and linguistic annotation from the final processing of the first version of the 88milSMS corpus. Although anonymization is explained in depth elsewhere (Accorsi *et al.*, 2014, Patel *et al.*, 2013), I have provided a summary of the problems involved here, as well as a list of the anonymization tags used for 88milSMS, since these are similar in structure to those provided for the linguistic annotation, which is described later.

2 Anonymization

Anonymization of private data is crucial and a legal requirement, which was closely monitored by the University's legal specialists. It took eight student internships and 21 months to accomplish the non-trivial three-step semi-automatic anonymization task, involving computational linguistic techniques.

The ten tags used and the number of occurrences of each tag, within the corpus, are as follows: PRE (*PRÉnom* = first name, 10,905), SUR (*SURnom* = nickname, 1,042), NOM (NOM = last name, 785), TEL (telephone number, 123), LIE (*LIEu* = place 102), ADR (address, 85), MAR (*MARque* = brand name, 58), COD (code, 50), MEL (MEL = email address, 27), and URL (13).

A piece of software was especially devised by students to semi-automatically anonymize the first/last names, nicknames, (email) addresses, places, telephone numbers, codes, URLs, tradenames, etc., appearing in the SMS data, collected within the 'sud4science' framework (Accorsi *et al.*, 2014, Patel *et al.*, 2013). Of course, SMS writing is often very creative, rendering the anonymization process highly difficult: first names may (or may not) be capitalized (*Cédric*/*cédric*), characters may be repeated (*Céééééédric*), diminutive/abbreviated forms appear (*Gégé* for *Gérard*, *JP* for *Jean-Pierre*), etc.

The first automatic step meant 72% of the corpus was anonymized (dictionary comparison allowed proper names such as *Cédric* to be anonymized, whereas words such as *crayon* ('pencil', in English) were discarded, since they belonged to one of the 'anti-dictionaries' used: LEFFF (Dictionary of inflected forms of the French language, Sagot, 2010; Dictionary of some SMS writing forms; Dictionary of place names).

The second semi-automatic step required human expert intervention to discriminate between items

Table 1 Anonymization for SMS data (Accorsi *et al.*, 2014, p. 16)

Word processed	In dictionary?	In anti-dictionary?	Label	Treatment
Cédric	Yes	No	Dictionary	Automatically Anonymized
Crayon	No	Yes	Anti-dictionary	Ignored (not to be anonymized)
Pierre	Yes	Yes	Ambiguous	Highlighted (candidate for the semi-automatic phase)
Namrata	No	No	Unknown	Highlighted (candidate for the semi-automatic phase)

requiring anonymization and those that remained unchanged (*Pierre/pierre* corresponds to 'Peter' or 'stone' in French, depending on the context).

All words not contained in either the dictionary or one of the anti-dictionaries (automatic phase), or that had been highlighted as ambiguous candidates (semi-automatic phase), were considered 'unknown' and also highlighted (semi-automatic phase). This is summarized in Table 1.

The third validation phase (conducted by student linguist interns) was important for confirming or modifying previous automatic decisions:

(n° 18307)
grace a lui on comprend trop bien franchement ke kiffe la physique cette **anne** meme si cest bien dur [...]
thanks *to him we really understand frankly I love physics this* ***year*** *even if it's really hard*
Example 1: Validation phase.

In Example 1, 'grace' (as in, first name = 'Grace' versus 'grâce à' = 'thanks to') and 'anne' (first name = 'Anne', 'année' = year) were automatically anonymized, owing to omitted accents. The linguist experts were able to modify this decision by removing anonymization tags during the validation phase, since 'grâce à' ('thanks to') and 'année' ('year'), respectively, were the required words, in this case.

(n° 81793)
C bon tu peux m appeler sur mon fixe **<TEL_10><PRE_4>**
It's ok you can call me on my landline (ten character telephone number, four characters in first name)
Example 2: Anonymized SMS.

The 88milSMS corpus was entirely anonymized before dissemination (*cf.* Accorsi *et al.*, 2014, Patel *et al.*, 2013 for more details).

We have provided a sample of 1,000 'raw' text messages transcoded into standardized French and another sample of 100 linguistically annotated SMS. Why decide to exclude 'full' transcoding and annotation phases in the first version of the final corpus?

3 Transcoding

Transcoding 'raw' text messages into 'standardized' French means morpho-syntactic parsers and other natural language processing (NLP) tools can ultimately analyse them. Concerning the terminology, the 'sud4science' team deliberately chose to use 'transcoding', since it can be defined as converting from one form of coded representation to another. This allows to discriminate between oral speech (to written) 'transcription' techniques and written (to written) 'transcoding' ones, such as SMS data. From a linguistic point of view, one can also use the mainstream 'standardization', a synonym that we indeed used previously, along with 'normalization', which we prefer to use when faced with computational linguistics matters (Lopez *et al.*, 2014). Here, I have maintained 'transcoding'.

Checking spelling and grammar facilitates comprehension, but 'no' supplementary information should be 'injected'.

'Raw' anonymized SMS (n° **22446**):
En fait c rien de spécial, jprends juste un peu de recul et jcomprends pas ce que jfous là, fac, psycho, montpellier, pourquoi simplement je vis, enfin bref rien de grave. Qu'est ce qui cloche chez toi?
Anonymized and transcoded SMS:
En fait **c'est** rien de spécial, **je** prends juste un peu de recul et **je** comprends pas ce que **je** fous là, **fac**, psychologie, **M**ontpellier,

pourquoi simplement je vis, enfin bref rien de grave. Qu'est-ce qui cloche chez toi?
In fact, it's nothing in particular, I'm just stepping back abit and I don't understand what I'm doing here, at uni, psychology, Montpellier, simply why am I alive, you know, nothing dire. What's wrong with you?
Example 3: From a 'raw' anonymized text message to a transcoded one.

In Example 3 above, the French negation 'ne' is not re-inserted (*ce n'est rien, je* ***ne*** *comprends pas*), since in oral forms, this is quite common and the negation 'pas' is sufficient for a parser. Prepositions/articles (« *à la* fac », « *en* psychologie », « *à* Montpellier ») are not 'reinjected' either, since automatic processing is possible without them. However, for abbreviated and agglutinated forms ('c' = > 'c'est'; 'jprends' = > 'je prends') transcoding into standardized French is necessary, so that a morpho-syntactic parser can automatically process the sentence. The apocope 'fac' (instead of 'faculté', for University) has not been modified since the researchers decided to validate the transcoding in relation to the online French *Petit Robert* (PR, 2014) dictionary. If a lexical item appears therein, it is not transcoded in the corpus. Here, 'psycho' is transcoded into 'psychologie' because it does not appear as such in the dictionary. The PR includes certain popular forms, such as 'frérot' (brother), foreign words: 'week-end', acronyms: 'lol', French inverted forms ('verlan'): 'relou' (lourd/that's a pain), etc. These are not transcoded into standardized French. Typographical norms are also re-inserted; in this example, a space before the question mark in French and a capital 'M' for the city of Montpellier.

What if a texter tries to simulate a certain form of oral French, for instance, by using an apostrophe, or through agglutination ('j'sais'='je sais', 'chuis'='je suis') as shown above? Should these items be transcoded or not? What about punctuation, often absent in text messages? Should one re-introduce this systematically? Example three shows how difficult the transcoding process can be.

Researchers may well have differing theoretical viewpoints on these matters. In November 2011, the Montpellier team invited researchers involved in previous sms4science data collections to a two-day workshop to exchange views on harmonization/standardization techniques related to anonymization, transcoding, and annotation for processing SMS written data. Over and above compulsory anonymization, some teams had either partially or entirely transcoded their SMS 'raw' data into standardized French and conducted linguistic annotation. Others had not. It is extremely difficult to agree on standardized ways to proceed, owing to varying theoretical views, or (pluri)disciplinary positions. For instance, in one of our seminars, two psychologists, Goumi and Bernicot (2011), presented some of their transcoded data. One of the 'raw' SMS examples they provided was as follows: 'Lèa t c se kil i a fair en techno'. This more or less translates to the following: 'Léa, do you know what we have to do in technology?' The example was transcoded—following their specifications—so as to maintain 'oral forms' ('Léa t'sais ce qu'il y a à faire en techno') and 'a formal academic normed transcription' was then provided ('Léa sais-tu ce qu'il y a à faire en technologie?'). In this case, they chose to radically transform the original SMS, with, among other aspects, questions with subject pronoun + verb inverted forms ('t c'/'t'sais'/'sais-tu'), contractions or apocopes ('techno'/'technologie'), phonetic variations, ellipsis, etc. ('kil i a fair'/'qu'il y a à faire'). These transcodings may suffice for psychologists, but they would most certainly cause debate for linguists, who would be inclined to have differing views on acceptable transcodings, from oral/written/computational linguistics perspectives. I actually set up a transcoding exercise with my colleagues to check these differences. I chose a sample of 1,000 text messages and submitted it to them: there are two computer scientists involved with NLP, one computational linguist (CL), two discourse analysis linguists, and one sociolinguist. The conclusion was radical: we had transcoded the extract depending on our discipline areas. For those involved in NLP and CL, it was important to take into account the fact that the sample could be processed by a machine, therefore 't' from the above example would need to be transcoded into 'tu', whereas for a linguist who is used to working with oral transcriptions, this is unjustified and perceived

as 'injecting' an interpretation which is initially absent. The list goes on and on.

Even though manual transcoding is not a viable option for standardization of subsequent versions of the 88milSMS corpus, normalization using automated NLP techniques has been researched by our team (see Section 5, Lopez *et al.*, 2014).

4 Annotation

Another issue is linguistic annotation of the corpus (Ide and Pustejovsky, forthcoming). For example, the 'raw' SMS 'je met tout ça de coté et peux tout encaisser juste pour toi.' (*I'm leaving all of that aside and I can bear it all just for you.*) could be transcoded into standardized French as follows: '**Je mets** tout **ça** de **côté** et **je** peux tout encaisser juste pour **toi.**' It could then be linguistically annotated with information of interest to researchers, among other items: spelling, grammatical information, emoji insertion, code-switching, typography, missing accents, voluntary modification, etc. Therefore, I define linguistic annotation of SMS data for the 88milSMS corpus, as 'interpretative' linguistic information indicated via appropriate tags (see below), related to the difference between a 'raw' text message and its transcoded equivalent in standardized French. I do not include in this definition, lemmatization, or part-of-speech (POS) tagging (see Section 5), which do indeed also correspond to other methods of linguistic annotation (based mainly on providing lexico-morpho-syntactic information).

After much scholarly debate about previous experiences with other sms4science members (e.g. the Quebec team had used eighteen tags which were very difficult for their annotators to discriminate between and apply easily), eight tags were chosen for linguistic annotation of 88milSMS:

(1) <TYP> (typography: punctuation, mathematical symbols, accents, numbers, hours, &, < >, (), upper and lower case, page formatting); (2) <MOD> (modification by reduction, increase, character substitution, abbreviations, acronyms, character/phonetic repetition, interjections and onomatopoeia. . .): *ht* (*acheter*), *pr* (*pour*), c (*s'est, c'est, ces. . .*), dcd (*décider*), etc.; (3) <GRA> (grammar: grammatical agreement: *il viens* (*il vient*), syntax, etc.); (4) <EMO> (emoji, emoticons: :) ^^ :p ;) :d <3 :-) xd :(:/; (5)<ABS> (absence/ellipsis: negation, pronouns, easily identifiable missing items); (6) <LAN> (language: words borrowed from other languages, regionalisms, neologisms, French 'verlan', slang, etc.); (7) <ORT> (spelling: typing mistakes, inverted characters, etc.); and (8) <DIV> (diverse: if no other tag is appropriate).

Examples of these tags appear in Table 2 (note that only one type of tag appears per SMS to facilitate reading).

As for transcoding, if items appear in the PR2014 (e.g. 'ah', 'boum', 'ben', 'bah', 'bouh', 'ouais', 'frérot', 'lol', 'relou', 'prof', 'sympa', 'papi', 'cool', 'box-office'), then tags are not applied.

When used, some tags seem relatively unambiguous:

n° 7063, Ahah t'es drôle <TYP_missing space>! Samedi matin<TYP_missing space>?
Ha ha you're funny! Saturday morning?
n° 43927, Pk tu t <LAN> fighter avec 1 mec a midi
Because you had a fight with a guy at midday
n° 6887, elle est trop bien cette prof, chui amoureux d'elle <EMO> ^^
this teacher is great, i'm in love with her ^^
n° 4671, <DIV> Ffghoeksjclfpzozkdkfoeeogrjz jglelsjloe
Example 4: Unambiguous tags.

In n° 7063, <TYP> indicates a missing space before punctuation (necessary in French). In n° 43927, <LAN> refers to a word which is borrowed from English ('fight'). The emoticon in n° 6887 is easy to recognize.

Annotation involving double (or more) tags may also be necessary in some situations:

n° 5409, T'y vas à quelle heure? Nous on y est dans 10 minutes <EMO_TYP_missing space>^^
What time are you going there? We'll be there in ten minutes^^
n° 43818, Oww emm gee <MOD_LAN> neighb !! La saison 3 de vampire diaries est juste incroyable!

Table 2 Linguistic annotation

SMS	Tags	SMS after tagging	Translation
n° 6885	TYP	Zorro est <TYP_arrivé> arrive, sans s'presse [. . .]	Zorro arrived, without hurrying [. . .]
n° 4360	MOD	[. . .] Oui, <MOD_j'y> j <MOD_suis> sui <MOD_allé> zalé ! [. . .]	Yes, I went there ! [. . .]
n° 5536	GRA	Cc tu <GRA_vas> va mieux. Mam ma <GRA_dit> dis ke tèté retmbè malade. Et bb ? Bisx	Hi are you better. Mum told me that you were sick again. And baby? Kisses
n° 6887	EMO	[. . .] <EMO> ☺ elle est trop bien cette prof, chui amoureux d'elle <EMO> ^^	☺ This teacher is fantastic, I'm in love with her ^^
n° 19621	ABS	[. . .] je met tout ça de coté et <ABS_je> peux tout encaisser juste pour toi. [. . .]	I'm putting all of that aside and I can support it all just for you.
n° 43133	LAN	<LAN> if(ce_soir == film) { <LAN> get_commande;} <LAN> else { <LAN> set_tagueule;} <LAN> return "bisous"	If(this evening == film) {get command;} else {set_shut up;} return "kisses"
n° 19621	ORT	[. . .] notre couple sera tel un <ORT_roseau> rosau à jamais se casser [. . .]	Our couple will be like a reed, never to be broken
n° 4671	DIV	<DIV> Ffghoeksjclfpzozkdkfoeeogrjzjglelsjloe	

OMG neighbour!! Season 3 of Vampire Diaries is just incredible!
n° 49721, C est pas TOI le pb le pb <TYP_ORT>c edt le groupe!
It's not YOU the pb the pb was the group!
Example 5: Unambiguous double tags.

The emoticon in n° 5409 has a missing space before it; thus, <TYP> is also a necessary tag. In n° 43818, 'neighbour', which appears in English <LAN>, has been shortened to 'neighb', thus justifying the <MOD> tag. In n° 49721, 'c edt' (*c'est*) has a missing apostrophe <TYP> and a typing mistake <ORT>.

In other situations, however, it might be difficult to decide which tag(s) to choose:

n° 49808, <MOD? > <ORT?>bone journè
Have a nice day
n° 11682, Il <GRA? > <MOD?>es rentrer a 22h30 et jai eu ldroii au : jsui fatiguer, jai mal a la tete jvai me coucher.
He came home at 10.30pm and I got to hear: I'm tired, I have a headache, I'm going to bed
Example 6: Tag choice.

In n° 49808, the 'scriptor' may have voluntarily modified the two words ('Bonne journée') or may have lacked spelling knowledge. So should <MOD> and/or <ORT> be used? In n°11682, 'rentrer' ('Il est rentré') could be either a grammatical mistake <GRA> or the scriptor may have preferred using an 'r' <MOD> instead of pressing the 'e' to access the acute accent (on a smartphone).

Sometimes, researchers may well disagree with the choice of tags. In Example 7, below, should one indicate that a subject pronoun is 'missing'? The 'absence' or 'ellipsis' notion may not be relevant for certain researchers. For instance, for a CL, in Example 7, the subject pronoun 'je' (I) is missing, and may be categorized as an 'ellipsis'. For other linguists, for instance, those working on oral forms, the ellipsis/absence idea is irrelevant because one should merely accept the example, as it was spoken/written in the first place—from this point of view, nothing is 'missing', as such. Punctuation and typography are also an important issue. To what extent should they be 'reintroduced' if absent? This is a highly frequent situation in text messages.

je met tout ça de coté et <ABS_je> peux tout encaisser juste pour toi <TYP_.>
I'm putting all of this aside and I can put up with it all just for you
Example 7: 'missing' items, 'ellipsis', punctuation/typography.

From the above examples, one can perceive that it is extremely difficult to provide satisfactory standardized linguistic annotation. As in the previous transcoding phase, annotation may therefore become a source of theoretical disagreement.

To provide insight into these issues, we have provided an online sample of 100 annotated text messages. No tags were required for 70% of the 100 SMS sample. The percentages of each tag used for the remaining 30% are as follows: <TYP> 43.6%, <MOD> 28.5%, <GRA> 8.2%, <EMO> 8%, <ABS> 5.1%, <LAN> 3.7%, <ORT> 2.9%, <DIV> 0.1%. Thirty-three instances of double tags were used: <MOD_ORT> 40%, <MOD_LAN> 15%, <TYP_GRA> 15%, <TYP_ORT> 9%, <MOD_GRA> 6%, <GRA_ORT> 6%, <TYP_LAN> 6%, <TYP_EMO> 3%. The most common double tag <MOD_ORT> tends to indicate that spelling variation is intentional.

5 Conclusion

We decided to limit the processing to two extracts. Our (rare) choice to exclude full transcoding and tagging is a theoretical position: linguistic annotation of SMS data (as we have defined it, *cf*. Section 4) is far from neutral. It is directly linked to an interpretative framework. A true consensus on how to standardize the transcoding and linguistic annotation does not exist, owing to differing/varying theoretical, (pluri)disciplinary, and scientific stances. McEnery and Hardie (2012) comment on the two sides of the coin, weighing up the pros and cons of corpus annotation:

> Arguments against annotation are largely predicated upon the purity of the corpus texts themselves, with the analyses being viewed as a form of impurity. This is because they impose an analysis on the users of the data, but also because the annotations themselves may be inaccurate or inconsistent [...]. Such claims are interesting because, as has been noted, corpus annotation is the manifestation within the sphere of corpus linguistics of processes of analysis that are common in most areas of linguistics. To identify problems with accuracy and consistency, in corpus annotation is, in principle at least, to identify flaws with analytical procedures across the whole of linguistics. It is because of the issues of accuracy and consistency, in particular, that some linguists prefer to use unannotated corpora. But this does not mean to say that such linguists do not analyse the data they use; rather, it means that they leave no systematic record of either their analysis or their errors which can easily and readily be tied back to the corpus data itself. (McEnery and Hardie, 2012, p. 14)

We believe that mark-up initiatives should not be imposed upon researchers; it seems more relevant to let them conduct their own annotation bearing their specific scientific questioning in mind, without being trapped within a unique theoretical framework.

Another alternative is that researchers may of course prefer to provide both 'raw' and tagged corpora: 'Dissemination will take two different forms: one version of a corpus with the "raw" text without any tokenization and annotation (v1), and a second version of the same corpus with the annotations (v2).' (Chanier *et al.*, 2014, p. 2). For instance, Riou and Sagot (2016) present morpho-syntactic tagging of a specific corpus within the French CoMeRe corpora repository (v2), following on from a previous version without it (v1).

The 88milSMS digital corpus resource will provide inspiration for many years to come. Our corpus can be used to analyse contemporary mediated electronic discourse, from a (pluri)disciplinary perspective (linguists, communications specialists, psychologists, sociologists, computer data specialists, etc.), build knowledge on SMS writing forms (Panckhurst 2009, Roche *et al.*, forthcoming), and let algorithms learn from this: alignment methods for facilitating automatic transcoding/standardization/normalization are currently being explored (Lopez *et al.*, 2014, following Aw *et al.*, 2006, Beaufort *et al.*, 2008, Guimier de Neef and Fessard, 2007, Kobus *et al.*, 2008), as are methods for classifying 'unknown' items for use in automatically identifying lexical 'creativity' within 88milSMS and also to improve electronic dictionary approaches (Lopez *et al.*, 2015). If normalization techniques can be truly implemented for processing 88milSMS, then lemmatization and POS-tagging may also be envisaged, since the latter currently include a high error ratio (if tools are used on 'raw'

text messages). In Lopez *et al.* (2016) we specified the following as our next step:

> In order to refine automatic normalisation techniques for initially non-standard texts in French, the next logical step is to compare our resource with different types of instant media (i.e. SMS, forums, tweets). Firstly, a new typology of the detected 'mistakes', based on existing typologies, will be elaborated. Secondly, automatic normalisation techniques—focussing on the most frequent errors—will be proposed. These will then be confronted with traditional automatic translation (Vilariño *et al.*, 2012), speech recognition (Kobus *et al.*, 2008) and spelling/grammatical checker principles (Beaufort *et al.*, 2010). Finally, the approach should enable comparison between different types of instant media. (Lopez *et al.*, 2016).

The resource also sheds light on 'corpus-driven' and 'corpus-based' approaches (Panckhurst *et al.*, forthcoming). We produced and submitted an XML encoding of 88milSMS, within the Dariah initiative in 2015 (Digital Research Infrastructure for the Arts and Humanities: Dariah-fr, http://www.dariah.fr/). A 2016 version of 88milSMS, which has been produced respecting XML, TEI guidelines and allows more widespread access, due to a CC BY 4.0 licence on the Ortolang platform (https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88-milsms-tei-v1; Panckhurst *et al.*, in Chanier (ed), 2016), is another major step forward. This is indeed a further form of (shareable) annotation, which could be of use to the community. Thierry Chanier conducted an XML-TEI transfer for this v2 version of 88milSMS, including additional encoded metadata with detailed information on the project, the corpus, and the questionnaire.

I also hope—thanks to the two most recent XML, TEI initiatives—that the resource will be eligible for long-term archiving with the CINES (Centre Informatique National de l'Enseignement Supérieur, https://www.cines.fr/). This would mean that in the future, people could look back and explore these 'snapshot' resources and understand more about the evolution of scriptural practices and usages in the 21st century.

Acknowledgements

I would like to thank two anonymous reviewers for their valuable and thought-provoking remarks. Any remaining mistakes are of course my own.

This work was 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 the CNRS (PEPS ECOMESS, HuMaIn). The SMS data described in this article was collected within the framework of the sud4science LR (http://www.sud4science.org) project. 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. In particular, we thank Cédrick Fairon, Louise-Amélie Cougnon, and Hubert Naets (CENTAL), for their support, during our project. Many thanks to my colleagues, Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche, Bertrand Verine. The SMS project, Sud4science LR, would never have taken place had my colleagues decided not to join me in the adventure. We are very grateful to our 'Informatique et Libertés' (data protection legislation) legal advisor, Nicolas Hvoinsky, and his director, Stéphanie Delaunay (DAJI, Université Paul-Valéry Montpellier 3), who accompanied and legally advised our team throughout the project. We thank our student interns: Anthony Stifani (Master's student in Information and Communication, Université Paul-Valéry Montpellier 3), who manually analysed many of our text messages, thus allowing evaluation of the anonymization system; Pierre Accorsi and Namrata Patel (Master's students in Computer Science at the Université de Montpellier), who developed the 'Seek&Hide' software, used to anonymize the corpus; Michel Otell, Camille Lagarde-Belleville, Frédéric André, and Yosra Ghliss (Master's students in Language Sciences, Université Paul-Valéry Montpellier 3) who performed the online manual anonymization with 'Seek&Hide' and verified the automatic anonymization of the corpus; Aghiles

Lounes, Tarik Zaknoun, Zakaria Mokrani, Reda Bestandji, Takfarinas Sider, Ahmed Loudah (Master's students in Computer Science, Université de Montpellier) who worked on an automatic transcoding system.

References

Accorsi, P., Patel, N., Lopez, C., Panckhurst, R., and Roche, M. (2014). Seek&Hide: anonymising a French SMS corpus using natural language processing techniques, In Cougnon, L.-A., and Fairon, C. (eds), *SMS Communication. A Linguistic Approach*. Amsterdam/ Philadelphia: John Benjamins, pp. 11–28.

Antoniadis, G., Chabert, G., and Zampa, V. (2011). *Alpes4science: Constitution d'un corpus de SMS réels en France métropolitaine, talk*, May 9–10, Sherbrooke: 79th Acfas colloquium.

Aw, A., Zhang, M., Xiao, J., and Su, J. (2006). A phrase-based statistical model for SMS text normalization. In *Proceedings of the COLING/ACL on Main conference poster sessions*, Sydney, pp 33–40. Association for Computational Linguistics. http://anthology.aclweb.org/P/P06/P06-2.pdf#page=43.

Beaufort, R., Roekhaut, S., and Fairon, C. (2008). Définition d'un système d'alignement SMS/français standard à l'aide d'un filtre de composition. *Proceedings, JADT*, 2008: 155–66.

Beaufort, R., Roekhaut, S., Cougnon, L.-A., and 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, pp. 770–779.

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba L., Longhi J., and Seddah D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres, Special issue on Building And Annotating Corpora Of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics, *JLCL (Journal of Language Technology and Computational Linguistics)*, 29(2): 1–31. http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf.

Cougnon, L.-A. and Ledegen, G. (2010). C'est écrire comme je parle. Une étude comparatiste de variétés de français dans l'écrit sms. *Les voix des Français. Modern French Identities*, 2(94): 39–57.

Cougnon, L.-A. and Fairon, C. (eds) (2014). *SMS Communication. A linguistic Approach*. Amsterdam/ Philadelphia: John Benjamins.

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

Drouin, P. and Guilbault, C. (2016). *De ''Viens watcher la partie avec moi'' à ''Come regarder the game with me''*. Louvain-la-Neuve, Belgium. Abstracts, PLIN 2016, 12 May, http://www.plindayucl.com.

Dürscheid, C. and Stark, E. (2011). SMS4science: an international corpus-based texting project and the specific challenges for multilingual Switzerland. In: Thurlow, C. and Mroczek, K. (eds), *Digital Discourse, Language in the New Media*. Oxford: Oxford University Press, pp 299–320.

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

Goumi, A. and Bernicot, J. (2011). Un corpus de SMS produits par de jeunes adolescents: méthode de recueil et premières données, invited seminar, sud4science LR project, MSH-M, 15/3/2011.

Guilbault, C. and Drouin, P. (2016). Pratiques liées aux alternances de code dans un corpus anglais et français au Canada, Talk, Cercle linguistique Belge, 13 May 2016, Louvain-la-Neuve, Belgium.

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

Ide, N. and Pustejovsky, J. (eds) (forthcoming). *Handbook of Linguistic Annotation*. Berlin: Springer.

Kobus, C., Yvon, F., and Damnati, G. (2008). Transcrire les SMS comme on reconnaît la parole. In *Proceedings, TALN 2008*, Avignon, pp. 128–38. https://perso.limsi.fr/yvon/publications/sources/Kobus08transcrire.pdf (accessed 4 September 2016).

Langlais, P. Drouin, P., Paulus, A., Rompré Brodeur, E., and Cottin, F. (2012). Texto4Science: a Quebec French Database of Annotated Short Text Messages. *Proceedings, LREC*, Istanbul, pp.1047–54.

Lopez, C., Bestandji, R., Roche, M., and Panckhurst, R. (2014). Towards Electronic SMS Dictionary

Construction: an Alignment-based Approach. *Proceedings, LREC (Language Resources and Evaluation Conference)*, Reykjavik, Iceland, May 26-31, pp 2833–8.

Lopez, C., Roche, M., and Panckhurst, R. (2015). Classification des items inconnus de 88milSMS: aide à l'identification automatique de la créativité scripturale. *Travaux neuchâtelois de linguistique (revue TRANEL)*, 63: 71–86.

Lopez, C., Roche, M., and Panckhurst R. (2016), *Non-standard texts: from theoretical positions to Natural Language Processing normalisation*. Louvain-la-Neuve, Belgium, Abstracts, PLIN 2016, 12 May, http://www.plindayucl.com.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Moïse, C. (2013). Lol non tkt on ta pas oublié. Rapports à la norme et valeurs de la faute dans l'écriture Sms (projet et corpus Sud4science). Réflexions sociolinguistiques, Plenary, Colloquium Si j'aurais su, j'aurais pas venu! Linguistique des formes exclues: description, genre, épistémologie, Université Libre de Bruxelles, Montpellier for Panckhurst, June 20-22.

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

Panckhurst, R., Détrie C., Lopez C., Moïse C., Roche M., and 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*, University Korea – Centre for Applied Cultural Studies, Seoul, pp. 107–38.

Panckhurst R., Détrie C., Lopez, C., Moïse, C., Roche, M., and Verine B. (2014). 88milSMS. A corpus of authentic text messages in French. Produced by the University Paul-Valéry Montpellier and the CNRS, in collaboration with the Catholic University of Louvain, funded with support from the MSH-M and the Ministry of Culture (General Delegation for the French language and the languages of France) and with the financial participation of Praxiling, Lirmm, Lidilem, Tetis, Viseo. http://88milsms.huma-num.fr/ ISLRN: 024-713-187-947-8.

Panckhurst, R. and Moïse, C. (2014). French text messages. From SMS data collection to preliminary analysis. In Cougnon, L.-A. and Fairon, C. (eds), *SMS Communication. A Linguistic Approach*. Amsterdam/Philadelphia: John Benjamins, University Korea - Centre for Applied Cultural Studies, Seoul, pp. 141–68.

Panckhurst, R., Roche, M., and Lopez C. (2015). Données authentiques: un grand corpus de SMS en français, Abstracts, SHESL-HTL 2015 colloquium, Corpus et constitution des savoirs linguistiques, Paris, 30-31 January 2015, pp. 33–35.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., and Verine, B. (2016). ''88milSMS. A corpus of authentic text messages in French'' (new version of the ISLRN 024-713-187-947-8 corpus). In Chanier T. (ed), *CoMeRe corpora repository*. Nancy: Ortolang. [cmr-88milsms-tei-v1; https://hdl.handle.net/11403/comere/cmr-88milsms/cmr-88milsms-tei-v1].

Panckhurst, R., Roche, M., Lopez, C., Verine, B., Détrie, C., and Moïse, C. (forthcoming), De la collecte à l'analyse d'un corpus de SMS authentiques: une démarche pluridisciplinaire, H.E.L., 38(2). http://www.hel-journal.org/fr/.

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

Petit Robert de la langue française (2014). Digital Multisupport Version of the Dictionary.

Riou, S. and Sagot, B. (2016). Étiquetage morpho-syntaxique du corpus FAVI [corpus]. D'après Yun, H. and Chanier, T. (2014). Corpus d'apprentissage FAVI (Français académique virtuel international) [cmr-favi-tei-v1]. Banque de corpus CoMeRe. Ortolang.fr: Nancy. [https://hdl.handle.net/11403/comere/cmr-favi/cmr-favi-tei-v2].

Roche, M., Verine, B., Lopez, C., and Panckhurst, R. (forthcoming). La néographie dans un grand corpus de SMS français: 88milSMS. *Proceedings, CINEO 2015*. Salamanca.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings, LREC, 2010*, Valletta, Malta. http://hal.inria.fr/inria-00521242/; Project Alexina: https://gforge.inria.fr/projects/alexina/.

Vilariño, D., Pinto, D., Beltrán, B., León, S., Castillo, E. and Tovar, M. (2012). A machine-translation method for normalization of SMS. In Pattern Recognition (pp. 293–302). Berlin/Heidelberg: Springer. http://www.cs.buap.mx/~dpinto/research/MCPR2012/MCPR2012_Vilarino.pdf.

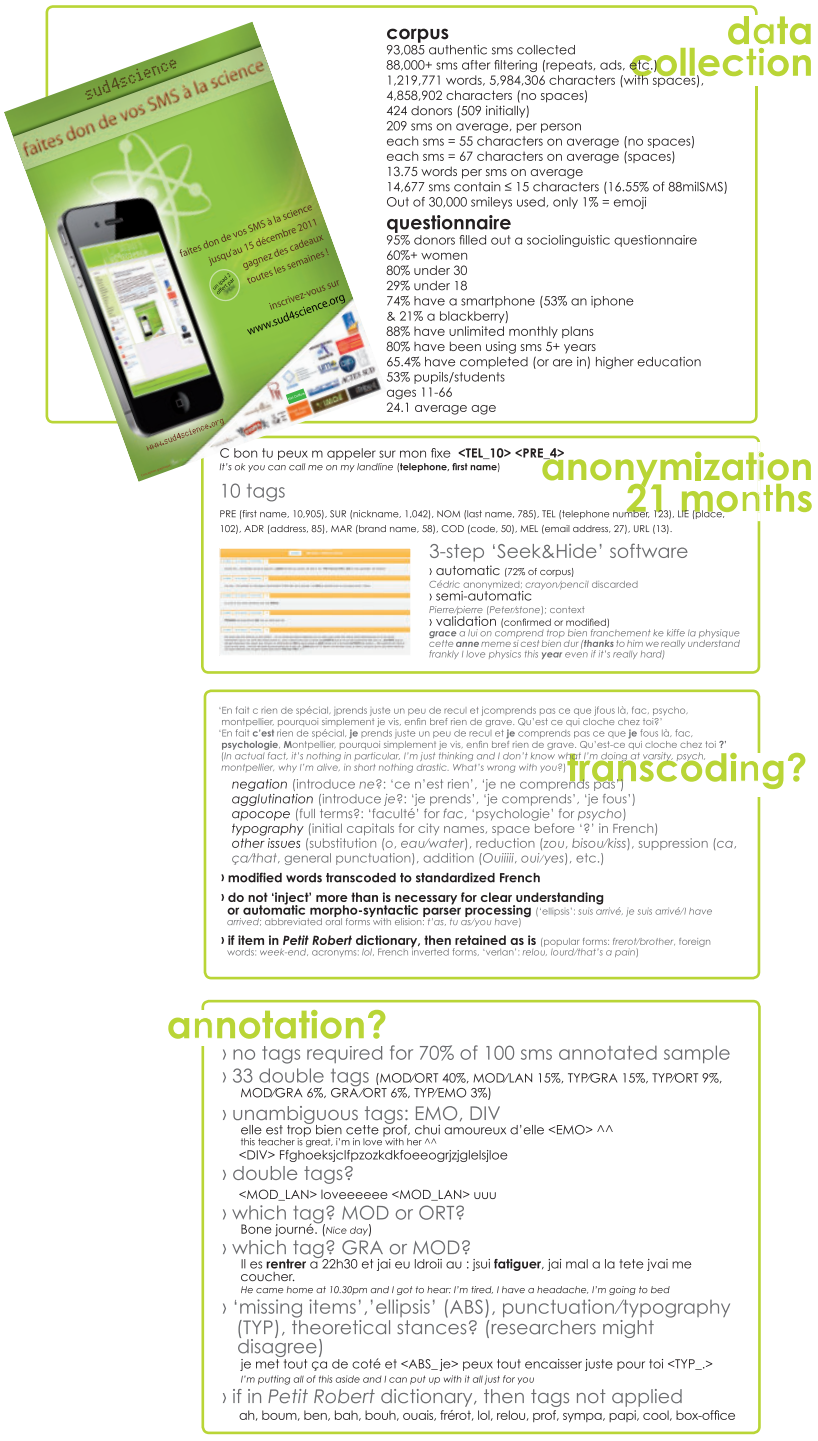

Fig. A1 88milSMS: data collection, anonymization, transcoding, annotation