**PANCKHURST — 88mil**

<0 images>

**‘88milSMS’: A New Digital Corpus Resource of French Text Messages: Why We Chose to Exclude Full Transcoding and Standardised Tagging**

Panckhurst, R.

In 2011, six academics gathered over 90,000 authentic text messages in French from the general public, in compliance with French law.1 The SMS ‘donors’ were also invited to fill out a sociolinguistic questionnaire (http://sud4science.org; Panckhurst et al.*,* 2013). The project is part of a vast international initiative titled sms4science (http://www.sms4science.org/; Fairon et al., 2006; Cougnon and Fairon, 2014; Cougnon, 2015) that aims to build a worldwide database and analyse authentic text messages. After the sud4science SMS data collection, a pre-processing phase of checking and eliminating any spurious information and a three-step semi-automatic anonymisation phase were conducted (Accorsi et al., 2014; Patel et al., 2013). Two extracts were transcoded into standardised French (1,000 SMS) and annotated (100 SMS). The finalised digital resource of 88,000 anonymised French text messages, the ‘88milSMS’ corpus, the extracts, and the sociolinguistic questionnaire data are currently available for all to download, via a user free-of-charge licence agreement, from the Huma-Num web service (http://88milsms.huma-num.fr; Panckhurst et al., 2014).

Why decide to *exclude* full transcoding and annotation tagging phases?

Transcoding ‘raw’ text messages into ‘standardised’ French means that morpho-syntactic parsers and other natural language processing tools can ultimately analyse them. Checking spelling and grammar facilitates comprehension, but *no* supplementary informationshould be ‘injected’. 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’)? Should these items be transcoded or not? What about punctuation, often absent in text messages? Should one re-introduce this systematically? Researchers may have differing theoretical viewpoints.

Another issue is tagging the corpus. After much scientific debate about previous experiences with other sms4science members, eight tags were chosen for ‘88milSMS’: TYP(ography), MOD(ificiation), GRA(mmar), BIN(ettes, smileys/emojis), ABS(ence), LAN(guage), ORT(hography, spelling), DIV(erse). Like the previous transcoding phase, annotation is a source of theoretical disagreement. To highlight this, it may be difficult to decide which tag to use, and double tagging may be necessary: *Bone journé*. The ‘scriptor’ may have voluntarily modified the two words (‘Bonne journée’ [have a niceday]) or may have lacked spelling knowledge. So should ‘MOD’ and/or ‘ORT’ be used? In another example, in the statement, ‘Il 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), ‘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). It is extremely difficult to provide satisfactory standardised tagging.

We decided to limit the processing to two extracts. Our (rare) choice to exclude full transcoding and tagging is a theoretical position: annotation is far from neutral. It is directly linked to an interpretative framework. A true consensus on how to standardise the transcoding and annotation does not exist, owing to differing/varying theoretical, (pluri)disciplinary, and scientific stances. We believe that no additional mark-up initiatives should be imposed upon researchers (other researchers disagree; see Chanier et al., 2014); 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.

The 88milSMS resource will provide inspiration for many years to come. Our corpus can be used to analyse contemporary mediated electronic discourse, build knowledge on SMS writing forms (Panckhurst, 2009), and let algorithms learn from this: alignment methods for facilitating automatic transcoding are currently being explored (Lopez et al., 2014), as are methods for classifying ‘unknown’ items for use in automatically identifying lexical ‘creativity’ within 88milSMS and also to improve electronic dictionary approaches. The resource also sheds light on ‘corpus-driven’ and ‘corpus-based’ approaches (Panckhurst 2013; Panckhurst et al., 2015). XML encoding means that the resource will be eligible for long-term archiving with the CINES (https://www.cines.fr/). Perhaps in the future, people will look back and explore these ‘snapshot’ resources and understand more about the evolution of scriptural practices and usages in the 21st century.

**Note**

1. Many thanks to my colleagues, Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche, Bertrand Verine; our 13 students, who all contributed to sud4science; Nicolas Hvoinsky, legal expert (CIL); and MSH-M, CNRS, and DGLFLF for funding.

**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: John Benjamins, pp. 11–28.  
**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. *JLCL (Journal of Language Technology and Computational Linguistics),* special issue on Building and Annotating Corpora of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics, http://www.jlcl.org/2014\_Heft2/Heft2-2014.pdf, 1–31.

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

**Cougnon, L.-A. and Fairon, C.** (eds). (2014). *SMS Communication. A Linguistic Approach.* John Benjamins, Amsterdam.**Fairon, C., Klein, J.-R. and Paumier S.** (2006). *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation.* Presses universitaires de Louvain, Louvain-la-Neuve; Manual + CD-Rom, http://www.smspourlascience.be/.

**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, 26–31 May 2014, pp. 2833–38.

**Panckhurst, R.** (2009). Short Message Service (SMS): typologie et problématiques futures. In Arnavielle, T. (coord.), *Polyphonies, pour Michelle Lanvin*, Université Paul-Valéry 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: 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., Roche, M. and Lopez C.** (2015). Données authentiques: un grand corpus de SMS en français. In *Proceedings, SHESL-HTL 2015 Colloquium,* Corpus et constitution des savoirs linguistiques, Paris, 30–31 January 2015, pp. 33–35.

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