PLIN 2016: "Language and the new (instant) media"

12 May 2016, Louvain-la-Neuve, Belgium.

"Non-standard texts: from theoretical positions to Natural Language Processing normalisation"

Cédric Lopez*, Mathieu Roche**, Rachel Panckhurst**

*R&D, Viseo, Grenoble cedric.lopez@viseo.com **UMR TETIS,Cirad, Irstea, AgroParisTech, Montpellier; LIRMM, UMR 5506 CNRS & Université Montpellier mathieu.roche@cirad.fr ***Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3 rachel.panckhurst@univ-montp3.fr

Domain Keywords: Mediated discourse analysis, Normalisation, Natural Language Processing. **Medium Keywords**: SMS.

Abstract

A finalised digital resource of 88,000 anonymised French text messages, the *88milSMS* corpus, two extracts (1,000 SMS transcoded into standardised French and 100 linguistically annotated SMS) and sociolinguistic questionnaire data were released in June 2014 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). The *sud4science* project (http://sud4science.org, Panckhurst *et al.* 2013), enabling authentic text message collection from the general public by a group of academics, is part of a vast international initiative (http://www.sms4science.org/, Fairon *et al*. 2006, Cougnon and Fairon, 2014, Cougnon 2015), to build a worldwide database and analyse authentic text messages in different languages.

We decided to exclude full transcoding and annotation tagging in the final corpus. This is a theoretical position, since annotation is far from neutral, and is invariably linked to an interpretative framework. Owing to varying theoretical disciplinary and scientific stances, it seems that a true consensus on how to standardise the transcoding and linguistic annotation tagging does not exist (Panckhurst, 2015). Other researchers may disagree and prefer to provide both 'raw' and fully tagged corpora (Chanier *et al*. 2014).

This theoretical position does not exclude exploring Natural Language Processing (NLP) investigation techniques, which could then be implemented in real-life applications. Examples of investigation techniques are indicated as follows: 1) Our corpus can be used to analyse current mediated electronic discourse, and help build knowledge on different SMS writing forms (Roche *et al*. 2015). 2) Algorithms may be used to learn from this: alignment methods for facilitating automatic transcoding have been explored (Aw *et al.* 2006, Beaufort *et al*., 2008, Guimier de Neef and Fessard, 2007, Kobus et al, 2008, Lopez *et al,* 2014). 3) We have devised a method for classifying 'unknown' items within text messages, which may help to automatically identify lexical 'creativity' within *88milSMS* and improve electronic dictionary approaches (Lopez *et al*. 2015).

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 — focusing 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.

References

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

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

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

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), 1-31. http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf

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

Cougnon, L.-A., Fairon, C., (éd., 2014). SMS Communication. A linguistic approach, Amsterdam/Philadelphia : John Benjamins.

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

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

Kobus, C., Yvon, F., & Damnati, G. (2008), « Transcrire les SMS comme on reconnaît la parole. », Proceedings, TALN 2008, 128-138. https://perso.limsi.fr/yvon/publications/sources/Kobus08transcrire.pdf

Lopez C., Bestandji R., Roche M., Panckhurst R. (2014) « Towards Electronic SMS Dictionary Construction : An Alignment-based Approach », Proceedings, LREC, Reykjavik, Iceland, 26-31 May, 2833-2838, http://www.lrec-conf.org/proceedings/lrec2014/pdf/753_Paper.pdf

Lopez C., Roche M., Panckhurst R. (2015), « Classification des items inconnus de 88milSMS : aide à l'identification automatique de la créativité scripturale », Travaux neuchâtelois de linguistique, 2015, 63, 71-86, https://www2.unine.ch/files/content/sites/islc/files/Tranel/63/71-86_lopez_al_corr.pdf

Panckhurst, R., Détrie, C., Lopez, C., Moïse C., Roche, M., 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, pp. 107-138, https://hal.archives-ouvertes.fr/file/index/docid/923618/filename/panckhurst_detrie_lopez_moise_roche_verine_v16.pdf

Panckhurst, R., Détrie, C., Lopez, C., Moïse C., Roche, M., Verine, B. (2014), '88milSMS. A corpus of authentic text messages in French', produit par l'Université Paul-Valéry Montpellier III et le CNRS, en collaboration avec l'Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8.

Panckhurst R. (2015), « '88milSMS, a new digital corpus resource of French text messages : why we chose to exclude full transcoding and standardised tagging. », Proceedings, Digital Humanities, Sydney, 29 June-3 July, http://dh2015.org/abstracts/

Roche M., Verine B., Lopez C., Panckhurst R. (2015), « La néographie dans un grand corpus de SMS français : 88milSMS », Proceedings, Cineo 2015, Salamanca, 22-24 October.

Vilariño, D., Pinto, D., Beltrán, B., León, S., Castillo, E., & 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