An Internationally Fair Mediated Digital Discourse Corpus: Improving Knowledge on Reuse

Rachel Panckhurst
Dipralang EA 739
Université Paul-Valéry Montpellier 3
Montpellier, France
rachel.panckhurst@univ-montp3.fr

Francesca Frontini
Istituto di Linguistica Computazionale
"A. Zampolli" - ILC - CNR
Pisa, Italy
& CLARIN ERIC
francesca.frontini@ilc.cnr.it

Abstract

In this paper, the authors present a French Mediated Digital Discourse corpus, (*88milSMS* http://88milsms.huma-num.fr https://hdl.handle.net/11403/comere/cmr-88milsms). Efforts were undertaken over the years to ensure its publication according to the best practices and standards of the community, thus guaranteeing compliance with FAIR principles and CLARIN recommendations with pertinent scientific and pedagogical reuse. Since knowledge on how resources are reused is sometimes difficult to obtain, ways of improving this are also envisaged.

1 Introduction

The adoption of Open Data and Open Science principles is producing important effects in SSH disciplines and has been enhanced by widespread awareness of the internationally ratified FAIR principles, aiming at ensuring that research data should be Findable, Accessible, Interoperable and Reusable[1]. In Linguistics and Natural Language Processing (NLP) the importance of curating Language Resources (LRs) for replicability and reuse has been particularly recognized, with relevant initiatives dating back several decades. Even before the formalisation of the FAIR principles as such, various initiatives have promoted good data management practices in the domain of Language Resources, starting with the FLaReNet[2] action and culminating with the creation of the CLARIN infrastructure. With its network of consortia and centres, CLARIN is making it easier for researchers to adhere to the requirements of the FAIR principles (de Jong et al., 2018), something which is increasingly required by evaluation and funding agencies. In France, over and above its role as CLARIN observer, various national centres are now active under the leadership of the national Huma-Num infrastructure, offering services and promoting the sharing of textual data (among other types) which meet the FAIR principles and the CLARIN best practices.

European Computer-Mediated Communication (CMC) and Mediated Digital Discourse (MDD) corpora initiatives are becoming more visible: Belgian *sms4science*, *Vos Pouces*, (Fairon et al., 2006; Cougnon, 2015; Cougnon and Fairon, 2014; Cougnon et al., 2017); Dutch SoNaR, (Oostdijk et al., 2008); French CoMeRe, (Chanier et al., 2014); German DeRik, (Beißwenger et al., 2013); Swiss *What's up Switzerland?*, (Ueberwasser and Stark, 2017; Frey et al., 2016). These data types are often difficult to process, standardize, analyze, owing to their complex nature, including 'noisy' content (Frey et al., 2019; Poudat et al., 2020).

The objective of this paper is to present *88milSMS*, a French CMC/MDD corpus with a focus on scientific and pedagogical reuse. Over the years the authors of the corpus have undertaken several efforts to ensure its publication according to the best practices and standards of the community, which in turn guarantee compliance with FAIR principles and CLARIN recommendations. However assessing how this translates into an increased usability is a cumbersome task. Knowledge on how resources are reused

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

[1] http://datafairport.org/fair-principles-living-document-menu

[2] http://www.flarenet.eu/

Rachel Panckhurst and Francesca Frontini 2021. An internationally FAIR Mediated Digital Discourse Corpus: Improving Knowledge on Reuse. *Selected papers from the CLARIN Annual Conference 2020.* Linköping Electronic Conference Proceedings 180: 180 185–193.

is sometimes difficult to obtain. After having provided surveys on reuse and analysed data emanating from them, the authors provide ideas on how to improve access to information about what people are doing with corpora when they reuse them.

2 Project & Corpus

The *sud4science* project[3] was part of a vast international initiative, entitled *sms4science*[4], which aimed at building a worldwide database and analysing authentic text messages in different languages — mainly French, but also Creole, German (written in Switzerland and Germany), Italian, Romansh (Dürscheid and Stark, 2011), and English (Drouin and Guilbault, 2016). Many scientific projects analyse authentic data, but ensuing corpora are not always made available for the scientific community and the general public, sometimes owing to legal requirements and commercial isses. However, there is a crucial need for researchers from a wide range of disciplines to have easy access to authentic data, in order to conduct analyses pertaining to their particular research fields. From the onset of the *sud4science* project, the possibility of easy access and reuse of authentic data was of utmost importance to the scientific team.

In 2011, over 88,000 authentic French text messages were collected during a 13-week period from the general public in Montpellier, France (Panckhurst et al., 2014; Panckhurst et al., 2016b) and SMS 'donors' were also invited to fill out a sociolinguistic questionnaire (Panckhurst and Moïse, 2014). An anonymization phase was conducted (Patel et al., 2013), owing to legal requirements for data-protection of private data (Ghliss and André, 2017). This involved anonymizing names, telephone numbers, places, brand names, addresses, codes, URLs[5]. In 2014, the finalised largest digital resource of 88,000 'raw' anonymized French text messages, the specific *88milSMS* corpus, two samples (1,000 transcoded SMS, 100 annotated SMS), and the sociolinguistic questionnaire data were made available for all[6] to download. The researchers chose the Huma-Num web service[7], then in 2016, they made a TEI/XML version of *88milSMS* available under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence on the 'Ortolang' platform, which provided the corpus with a citable persistent identifier[8]. Contributions to DARIAH and ELRA were also made in 2015, and *88milSMS* therefore has an ISLRN. These initiatives preceded the FAIR principles but are in strict alignment with them.

3 Towards Being FAIR

Findability refers to initiatives aimed at ensuring long-term preservation of LRs by depositing in a specialized data centre (Ortolang repository, France), documented by a rich and standardized set of metadata, which in turn can be harvested by international meta-catalogues (CLARIN Virtual Language Observatory), allowing international visibility. Thanks to the deposit on the Ortolang repository, the *88milSMS* corpus is **Findable** from the VLO[9], where it will appear by performing free text searches such as "cmc corpus" or "SMS corpus" and filtering by language. In addition to the visibility within the VLO, the corpus is indexed on Google, searchable on the ELRA catalogue[10], as well as on Isidore[11].

Accessibility is gained by adopting a clear set of licences and promoting open access within copyright limits and data protection regulations; single-sign-on technologies allow researchers to gain access to resources based on their institutional identifier. *88milSMS* is fully **accessible**, despite its sensitive content, thanks to the thorough anonymization and verification work (Ghliss and André, 2017) carried out with the help of the University's legal advisors; a short mandatory form needs to be completed to download from

[3] `http://sud4science.org`; (Panckhurst et al., 2016b).

[4] `http://www.sms4science.org`; (Fairon et al., 2006; Cougnon and Fairon, 2014; Cougnon, 2015).

[5] By default, first names, surnames and any data which enable identifying information are anonymized. It is of course frustrating for linguists and other scientists to feel that anonymization causes loss of information, which will not be able to be retieved at a later stage, but it is a stringent legal requirement.

[6] Both the scientific community and the general public.

[7] `http://88milsms.huma-num.fr` (Panckhurst et al., 2014), .

[8] `https://hdl.handle.net/11403/comere/cmr-88milsms` (Panckhurst et al., 2016a).

[9] `https://vlo.clarin.eu`

[10] `https://catalogue.elra.info`

[11] Isidore is a French search engine for documents and resources in SSH `https://isidore.science/document/http://hdl.handle.net/11403/COMERE/V3.3/CMR-88MILSMS`

the bilingual (French/English) Huma-Num web interface, with a user free-of-charge licence. However, no authentication or form completion are needed to download the corpus from Ortolang, where it is available via the Creative Commons CC BY 4.0 licence.

Interoperability and **re-usability** for LRs are particularly important, and crucially enabled by the use of standard annotation formats and common best practices, allowing researchers to exploit data from different projects. At the time of the data collection, similar initiatives took place at the international level (*cf.* § 1) thus making the overall philosophy of the corpus attuned to that of these other datasets. Indeed, other authentic data collections projects followed on in more recent years (Ueberwasser & Stark 2017; Cougnon *et al.* 2017). From the point of view of encoding, initial formats of the corpus were .ods spreadsheets and *ad hoc* .xml. The use of utf-8 was crucial at the time (2011), in particular to ensure the preservation of a subset of SMS containing the first instances of emoji (Panckhurst and Frontini, 2020). The work to make *88milSMS* fully **interoperable** was carried out later, with the inclusion within the CoMeRe initiative, where the project adopted a common TEI format. Thanks to the aforementioned efforts, *88milSMS* has been **reused** beyond the initial scope of the project, boasting 1,067 downloads from 52 countries (as of 01/01/2021), and with a broad spectrum of multidisciplinary applications (to name a few: language sciences, computational linguistics and text-mining processing initiatives, geographical place name identification, psychology case studies).

4 Towards Scientific and Pedagogical Reuse

4.1 Initial 2017 Survey

Three years after providing *88milSMS* for public download and dissemination, a survey on scientific usage of the corpus was conducted[12]. Results have shown a strong disciplinary tendency towards language sciences and computing including NLP, text mining and corpus linguistics research, mainly from higher education establishments. In terms of dissemination, 50% of the research cited was successfully circulated in Master's theses, PhDs, habilitations, books, articles, proceedings, etc. (Panckhurst et al., 2020).

4.2 Follow-up 2019 Survey

In 2019, an update survey was conducted in order to find out if colleagues had cited/used *88milSMS*. Responses were unfortunately minimal but they do indicate that the corpus is being used in language sciences, as is to be expected, but also in other disciplines:

1. Language Sciences and NLP:
   - university courses for 2nd-year students; identifying and improving spelling mistakes (Poitiers University); discourse genres (Lorraine University);
   - recent PhDs: French as a foreign language and how to include SMS-writing in didactic situations; linguistic analysis of French SMS-writing; SMS communication: NLP and information extraction;
   - qualitative comparative analysis between differing corpora, related to morphosyntactic French question-form usage and interactional aspects comparing SMS and oral language.
2. Geography: identification of place names and interpretation of variations (Master's 2 internship subject, 2019, IGN-Paris & Paris-Est Marne-la-Vallée University).
3. Psychology: digital communication and teenagers (relational, emotional romantic aspects, 12-16 year-olds, Master's 1 thesis 2019, Toulouse Jean-Jaurès University).

[12] Researchers and the general public who had signed up to an optional scientific newsletter were contacted.

4.3 Final Questionnaire

A further (third) survey was proposed to the scientific community in November-December 2020 (cf. Appendix for questions) during a three-week period, via three channels:

1. Optional scientific newsletter which had been used for the previous 2019 survey;
2. French natural language list, LN [13];
3. North American Linguist [14] list.

The respondents' countries were varied across the continents, although only a handful (8 countries out of 52 cited in the download forms) were mentioned: Australia, China, France, Germany, India, Serbia, Switzerland, Uruguay, with an understandably high percentage from France (47%). Students make up 38% of the respondents (undergraduate and graduate 23%, doctoral 15%); another 54% are teachers/researchers from a public institution, and 8% are engineers. The *88milSMS* corpus was used for the following research fields[15]:

- Linguistics (75%);
- NLP (25%);
- Psychology (8%);
- Other[16] (8%).

Within research fields, six types of tasks were mentioned[17]:

- Linguistic analyses (60%);
- Psychological analyses (10%),
- Statistical studies (10%);
- Corpus linguistics studies (40%);
- Resource building (30%);
- Data mining (20%).

In 50% of cases, the research has been disseminated[18] in internship reports (20%), PhD manuscripts (10%), scientific publications, including articles, chapters, books (20%). In another 40%, the research has not been disseminated, and finally, 30% mention 'other' without specifying any further information. When respondents indicate that they have cited the *88milSMS* corpus, it is in pedagogical situations and/or directly in the bibliography of some of their publications, via the official citation format either mentioning (Panckhurst et al., 2014) or (Panckhurst et al., 2016a). Interestingly enough, a very high percentage (78%) of the respondents did not know that a second version of the corpus (XML/TEI) had been released on Ortolang[19] in 2016! This seems to indicate that there is a lack of information on new versions of LRs being published, which would be important to address. Among those who knew about the latter version, reasons for downloading it were stipulated: "I want to conduct the study of linguistics

[13] `https://www.atala.org/liste_ln`

[14] `https://linguistlist.org/`

[15] Respondents could cite more than one field, hence percentages indicate more than 100% overall.

[16] Respondents did not necessarily mention which other research field was pertinent.

[17] Again, several boxes could be ticked, hence a total of more than 100% is obtained. One item indicated in the survey, "Sociological analysis", was not used for any tasks.

[18] The percentages also go beyond 100% owing to checking several boxes, and two items were not chosen at all: "Yes, through software" and "Yes, with the publication of a resource or dataset".

[19] `(https://hdl.handle.net/11403/comere/cmr-88milsms)`

area, and I need more up-to-date information", "I'm going to download the xml-tei version to be able to use it with TXM"[20]. The final survey question was added for any further information respondents wished to submit. Answers were given by 24% and mainly stipulated pedagogical usage by teachers ("I also use it for pedagogical purposes as an example of data constitution. I hope to have more opportunities to use it also in my publications.") and students ("Thank you very much for the linguistic corpus, it was very useful for my presentation on the French language.").

4.4 What Are the Difficulties Linked to Obtaining Reuse Information?

Unfortunately, as in 2019, only a minimal number of questionnaires were completed (corresponding, in terms of figures, to just under 5% of those having signed up for the scientific *88milSMS* newsletter). Firstly, there is no way of knowing which of the three channels was preferred for accessing the survey. Secondly, the worldwide COVID pandemic combined with end-of-the-year obligations may well have negatively affected the number of responses. Finally, it is difficult to assess whether respondents are, on the one hand, actually just not interested in providing this information to resource producers, or, on the other hand, if they really are too overloaded with many more urgent matters. In the case of the *88milSMS* corpus, there are two more factors which may contribute to low numbers of questionnaire responses:

1. Only 37% of those researchers having downloaded the first version of the corpus from the HumaNum platform have signed up to the optional scientific newsletter, which is the main contact method [21];

2. Downloading the second version of the corpus from the Ortolang platform does not require authentication or mandatory form completion.

4.5 How Do We Gain More Accurate Knowledge on Reuse?

Knowledge on resource reuse is, at times, difficult to obtain, as indicated above. Yet it is essential for resource producers to have access to this, in order to tailor future productions and provide useful person-machine interfaces for users. So the authors definitely need to improve the way in which corpus reuse information is obtained. But how can this be done? If the corpus is reused on another online platform, it might be possible to collaborate with software engineers to gain information (all the while respecting legal aspects) on how often the resource is used and what are the types of formulated queries.

Currently, tools such as the LRE (Language Resources and Evaluation) map have tried to systematically document resource citations in papers[22], and proposal for a Language Resource Impact Factor was made some years ago (Mariani and Francopoulo, 2015). However such approaches are incomplete, since they measure the impact of papers only on a limited number of sources. For instance the LRE map only gathers data from a set of conferences (COLING, IJCNLP, Interspeech, LTC, ACLHT, O-COCOSDA, RANLP) and one journal (the LRE Journal) and mentions of resources outside this scope are not referenced. As a consequence, for instance no reference to *88milSMS* can be found on the LREmap, which is surprising, given that the corpus has existed on the ELRA website since February 2015[23].

In the long term, infrastructures such as CLARIN could provide systematic information to researchers concerning the reuse and citation of their data, for instance by promoting data citation practices and systematically collecting information from publications and other sources; of course, practices should also be harmonised across disciplines, within the SSH and beyond, so that reuse beyond disciplinary boundaries can be detected.

In the current panorama, various initiatives are on-going to promote and systematise data citation[24].

[20] TXM a corpus exploration tool widely used in France in the field of SSH`http://textometrie.ens-lyon.fr/`

[21] Download numbers of the *88milSMS* 2014 corpus from the HumaNum platform, as of 1/1/2021: 1,067 total downloads: 675 persons (63%) having not enrolled to receive the newsletter; 392 persons (37%), on the contrary, having enrolled to receive it.

[22] `https://lremap.elra.info/`

[23] `https://catalog.elra.info/en-us/repository/browse/ELRA-W0082/` ISLRN: 024-713-187-947-8 ID: ELRA-W0082

[24] See for instance the Declaration on Data Citation principles `https://www.force11.org/datacitationprinciples`

Within the SSHOC project, which aims at building the Social Sciences and Humanities Open Cloud, work is under way on how to identify, describe and collect data citations (Larrousse et al., 2019).

Given that standards for data citation are still under construction, efforts by individual researchers or researcher teams to manually collect evidence of data reuse can provide ground truth (such as the surveys mentioned above), indicating what type of information researchers are more keen to obtain.

5 Conclusion and Future Work

In this paper, the authors indicated how the French Mediated Digital Discourse *88milSMS* corpus complies with the four FAIR principles (findability, accessibility, interopeability and reusability), also taking into account CLARIN recommendations. In particular, since reusability is an important goal of open and collaborative research, we concentrated on addressing the reusability factor (§4) by analysing various survey results and providing more in-depth discussion about difficulties and knowledge linked to reuse. FAIR principles, which were considered in §3, could be indeed adapted after future research is conducted on other CMC corpora.

Even though the corpus is fairly widely consulted, downloaded and used across the scientific community and beyond, it remains difficult to have sufficient access to other researchers' scientific and pedagogical reuse, despite implementation of an optional scientific newsletter and surveys.

The authors' own pedagogical usage at Université Paul-Valéry Montpellier 3 (under-graduate and post-graduate levels) for Language Science students includes studying discourse analysis and NLP techniques with contemporary instant messaging authentical data such as the *88milSMS* corpus, which has recently been incorporated on the widely consulted Sketch Engine[25] platform, thus allowing online analysis via a user-friendly interface without mandatory downloading[26]. The advantage of this sort of integration is to provide ever-increasing interdisciplinary scientific and pedagogical reuse possibilities.

In this sense, the collaboration on CMC corpora which has already started within CLARIN is crucial[27] for the harmonization of formats across international projects, for the identification of common technical solutions for browsing interfaces, and finally for the implementation of a Federated Content Search. Moreover, the centralised and curated access to different projects on similar themes which the CMC 'Resource Family' provides, allows to easily find comparable data from other projects. For instance, it would be interesting to compare *88milSMS/sms4science/CoMeRe*, (Panckhurst et al., 2014; Panckhurst et al., 2016b; Fairon et al., 2006; Cougnon, 2015; Cougnon and Fairon, 2014; Chanier et al., 2014) with the French sub-corpus of WhatsApp messages from *What's up Switzerland?* (Ueberwasser and Stark, 2017; Ueberwasser, 2017)[28], or with the FaceBook, Viber, WhatsApp messages from *vos pouces* (Cougnon et al., 2017), or even with daily writing during WW1 in *Corpus14* (Praxiling - UMR 5267, 2019), in order to see how communication has evolved over time and medium.

Finally, an important step for the preservation of *88milSMS* will be its long-term archiving, together with other FAIR corpora from French CLARIN repositories, at the National Computing Center for Higher Education (CINES)[29], thus providing insight on digital textuality usage for future generations.

[25] http://www.sketchengine.eu/

[26] One of the authors used Sketch Engine in an introductory NLP undergraduate course during the autumn 2020 semester, and the students were very enthusiastic about the easy-to-use interface, which they immediately adopted and preferred, compared to other corpus linguistics tools. In actual fact, she presented the platform rather hastily during an online synchronous class during COVID lockdown, so she was pleasantly surprised by the quality of the data analysis the students provided in their end-of-term assignments on *88milSMS* queries and regular expression usage.

[27] See for instance the past thematic event https://www.clarin.eu/event/2017/clarin-plus-workshop-creation-and-use-social-media-resources as well as the CMC 'Resource Family' entry https://www.clarin.eu/resource-families/cmc-corpora

[28] The latter project has recently provided open access to the corpus. See here for a general overview of the Swiss CMC-corpora initiative: https://cmc-corpora.ch/

[29] https://www.cines.fr/en/

References

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., and Storrer, A. 2013. DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 28(4):531–537. Publisher: Oxford Academic, https://academic.oup.com/dsh/article/28/4/531/1077484.

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 for Language Technology and Computational Linguistics*, 29(2):1–30. https://halshs.archives-ouvertes.fr/halshs-00953507.

Cougnon, L.-A. and Fairon, C., editors. 2014. *SMS Communication*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Cougnon, L.-A., Maskens, L., Roekhaut, S., and Fairon, C. 2017. Social media, spontaneous writing and dictation. Spelling variation. *Journal of French Language Studies*, 27(3):309–327. https://www.cambridge.org/core/journals/journal-of-french-language-studies/article/div-classtitlesocial-media-spontaneous-writing-and-dictation-spelling-variationdiv/9574CD6BF604BD8F866A270E1EC909A3.

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

de Jong, F., Maegaard, B., De Smedt, K., Fišer, D., and Van Uytvanck, D. 2018. CLARIN: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1515.

Drouin, P. and Guilbault, C. 2016. De 'Viens watcher la partie avec moi' à 'Come regarder the game with me'. In *Abstracts, PLIN 2016*, Louvain-la-Neuve, Belgium.

Dürscheid, C. and Stark, E. 2011. sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland. In Crispin, T. and Mroczek, K., editors, *Digital Discourse. Language in the New Media*. Oxford University Press. ISBN: 9780199795437, https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199795437.001.0001/acprof-9780199795437-chapter-14.

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. Manuel.CD-Rom., Louvain-la-Neuve.

Frey, J.-C., Glaznieks, A., and Stemle, E. W. 2016. The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In Corazza, A., Montemagni, S., and Semeraro, G., editors, *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016. 5-6 December 2016 Napoli*, pages 157–161, Torino. Academia University Press. https://bia.unibz.it/handle/10863/8949.

Frey, J.-C., König, A., and Stemle, E. W. 2019. How FAIR are CMC corpora? In Longhi, J. and Marinica, C., editors, *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora19), Cergy-Pontoise University, France, 9-10 September 2019*, pages 26–31. https://bia.unibz.it/handle/10863/11294.

Ghliss, Y. and André, F. 2017. Après la collecte, l'anonymisation : enjeux éthiques et juridiques dans la constitution du corpus 88milSMS. In Ciara R. Wigham, G. L., editor, *Corpus de Communication Médiée par les Réseaux*, pages 71–84. L'Harmattan, Paris. https://hal.archives-ouvertes.fr/hal-01722169.

Larrousse, N., Broeder, D., Brase, J., Concordia, C., and Kalaitzi, V. 2019. SSHOC D3.2 Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning, December. Final version - Approved by the European Commission.

Mariani, J. and Francopoulo, G. 2015. Language Matrices and a Language Resource Impact Factor. In Gala, N., Rapp, R., and Bel-Enguix, G., editors, *Language Production, Cognition, and the Lexicon*, pages 441–471. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-08043-7_25.

Oostdijk, N., Reynaert, M., Monachesi, P., Noord, G. V., Ordelman, R., Schuurman, I., and Vandeghinste, V. 2008. From D-Coi to SoNaR: a reference corpus for Dutch. In *Proceedings of LREC 2008*, Marrakech, Morocco. ELRA. http://www.lrec-conf.org/proceedings/lrec2008/pdf/365_paper.pdf.

Panckhurst, R. and Frontini, F. 2020. Evolving interactional practices of emoji in text messages. In Thurlow, C., Dürscheid, C., and Diémoz, F., editors, *Visualizing Digital Discourse. Interactional, Institutional and Ideological Perspectives*, pages 81–103. De Gruyter Mouton.

Panckhurst, R. and Moïse, C. 2014. French text messages. From SMS data collection to preliminary analysis. In Cougnon, L.-A. and Fairon, C., editors, *SMS Communication. A Linguistic Approach*, pages 141–168. John Benjamins. https://hal.archives-ouvertes.fr/hal-01485595.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., and 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, https://hal.archives-ouvertes.fr/hal-01485560.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., and Verine, B. 2016a. 88milSMS. A corpus of authentic text messages in French. In Chanier, T., editor, *Banque de corpus CoMeRe*. Nancy, France. Ortolang, https://hdl.handle.net/11403/comere/cmr-88milsms.

Panckhurst, R., Roche, M., Lopez, C., Verine, B., Détrie, C., and Moïse, C. 2016b. De la collecte à l'analyse d'un corpus de SMS authentiques : une démarche pluridisciplinaire. *Histoire Epistémologie Langage*, 38(2):63–82. https://hal.archives-ouvertes.fr/hal-01485577.

Panckhurst, R., Lopez, C., and Roche, M. 2020. A French text-message corpus: 88milSMS. Synthesis and usage. *Corpus [online]*, (20). http://journals.openedition.org/corpus/4852.

Patel, N., Accorsi, P., Inkpen, D., Lopez, C., and Roche, M. 2013. Approaches of anonymisation of an SMS corpus. In *Proceedings of CICLING 2013, LNCS*, pages 77–88, March 24-30, 2013, University of the Aegean, Samos, Greece. Springer-Verlag. https://hal-lirmm.ccsd.cnrs.fr/lirmm-00816285.

Poudat, C., Wigham, C. R., and Liégeois, L. 2020. *Corpus complexes. Traitements, standardisation et analyse des corpus de communication médiée par les réseaux*. Corpus (20).

Praxiling - UMR 5267. 2019. Corpus 14. ORTOLANG (Open Resources and TOols for LANGuage. https://hdl.handle.net/11403/corpus14/v2.

Ueberwasser, S. and Stark, E. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 84(5). https://bop.unibe.ch/linguistik-online/article/view/3849.

Ueberwasser, S. 2017. What's up, switzerland?: Challenges of a large, multilingual cmc corpus. CLARIN-PLUS workshop "Creation and Use of Social Media Resources", Kaunas, Lithuania. https://www.clarin.eu/sites/default/files/SimoneUeberwasser.pdf.

Appendix

Bilingual French-English 2020 Survey on *88milSMS* Reuse

Retour d'expérience sur l'utilisation du corpus 88milSMS. Pour pouvoir approfondir les données concernant les utilisations scientifiques et pédagogiques de 88milSMS, Rachel Panckhurst et Francesca Frontini vous seraient très reconnaissantes de prendre 5 minutes pour répondre à quelques questions. Nous vous remercions pour votre aide.
Feedback on usage of the 88milSMS corpus. In order to gain further insight into scientific and pedagogical usage of 88milSMS, Rachel Panckhurst Francesca Frontini would be very grateful if you could take 5 minutes to answer a few questions. We are very grateful for your help.

1. Nom de famille / Last name + Prénom / First name

2. Affiliation, institution, autre/other

3. Pays / Country

4. Quel est votre profil ? / I am a:

5. Dans quel(s) domaine(s) d'étude avez-vous utilisé le corpus *88milSMS* ? / I have used the *88milSMS* corpus for the following research fields:

6. Pour quelle(s) tâche(s) avez- vous utilisé 88milSMS? / I have used 88milSMS for the following task(s):

7. Les travaux ont-ils été valorisés ? / Has the research been disseminated?

8. Avez-vous cité le corpus *88milSMS* dans la bibliographie de vos travaux ? Si oui, pourriez-vous indiquer ci-dessous la citation utilisée ?

   Have you cited the *88milSMS* corpus in the bibliography of some of your publications? If so, can you insert the citation format you used below?

9. Le corpus *88milSMS* a été téléchargé plus de 1050 fois depuis sa mise à disposition en 2014 (`http://88milsms.huma-num.fr/`) Saviez-vous qu'une deuxième version du corpus (XML/TEI) a été mise sur Ortolang (`https://hdl.handle.net/11403/comere/cmr-88milsms`) en 2016 ? The *88milSMS* corpus has been downloaded more than 1050 times since its release in 2014 (`http://88milsms.huma-num.fr/`). Did you know that a second version of the corpus (XML/TEI) was released on Ortolang (`https://hdl.handle.net/11403/comere/cmr-88milsms`) in 2016?

10. Si vous avez répondu oui à la question précédente, avez- vous également téléchargé *88milSMS* depuis Ortolang? Pour quelle(s) raison(s) ? Si vous avez répondu non à la question précédente, pourriez-vous nous dire si vous comptez télécharger la version de 2016 à l'avenir ? Pourquoi ?

    If you answered yes to the previous question, have you also downloaded 88milSMS from Ortolang? For what reason(s)? If you answered no to the previous question, could you tell us if you plan to download the 2016 version in the future? Why?

11. Complément d'information que vous souhaiteriez apporter, notamment si vous avez répondu ”autre” à l'une des questions précédentes.

    Further information you wish to submit, including if you answered ”other” in any previous question.