The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres

Thierry Chanier, Céline Poudat, Benoît Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi, Djamé Seddah

Chanier, T., Poudat, C., Sagot, B., Antoniadis, B., Wigham, C.R., Hriba L., Longhi,J. & Seddah, D. (to appear, 2014). "The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres". Journal of Language Technology and Computational Linguistics (JLCL). Special Issue : "Building And Annotating Corpora Of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computa-tional Linguistics" (ed. by Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel).

The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres

Chanier, T., Poudat, C., Sagot, B., Antoniadis, B., Wigham, C.R., Hriba L., Longhi, J. & Seddah, D.

Abstract

The CoMeRe project aims to build a kernel corpus of different Computer-Mediated Communication (CMC) genres with interactions in French as the main language, by assembling interactions stemming from networks such as the Internet or telecommunication, as well as mono and multimodal, synchronous and asynchronous communications. Corpora are assembled using a standard, thanks to the TEI (Text Encoding Initiative) format. This implies extending, through a European endeavor, the TEI model of text, in order to encompass the richest and the more complex CMC genres. This paper presents the Interaction Space model. We explain how this model has been encoded within the TEI corpus header and body. The model is then instantiated through the first four corpora we have processed: three corpora where interactions occurred in single-modality environments (text chat, or SMS systems) and a fourth corpus where text chat, email and forum modalities were used simultaneously.

The CoMeRe project has two main research perspectives: Discourse Analysis, only alluded to in this paper, and the linguistic study of idiolects occurring in different CMC genres. As NLP algorithms are an indispensable prerequisite for such research, we present our motivations for applying an automatic annotation process to the CoMeRe corpora. Our wish to guarantee generic annotations meant we did not consider any processing beyond morpho-syntactic labelling, but prioritized the automatic annotation of any freely variant elements within the corpora. We then turn to decisions made concerning which annotations to make for which units and describe the processing pipeline for adding these. All CoMeRe corpora are verified, thanks to a staged quality control process, designed to allow corpora to move from one project phase to the next.

Public release of the CoMeRe corpora is a short-term goal: corpora will be integrated into the forthcoming French National Reference Corpus, and disseminated through the national linguistic infrastructure ORTOLANG. We, therefore, highlight issues and decisions made concerning the OpenData perspective.

1. Introduction: the CoMeRe project

Various national reference corpora have been successfully developed and made available over the past few decades, e.g. the *British National Corpus* (ASTON & BURNARD , 1998), the SoNaR *Reference Corpus of Contemporary Written Dutch* (OOSTDIJK et al., 2008), the DWDS *Corpus for the German Language of the 20th century* (GEYKEN, 2007), the DeReKo *German*

Reference Corpus (KUPIETZ & KEIBEL 2009) and the *Russian Reference Corpus* (SHAROFF, 2006). Despite being in strong demand, no French National Reference Corpus currently exists. Thus, the *Institut de la Langue Française* (ILF) has recently taken the initiatory step lay the groundwork for such a project. The aim is for the national project to both collect existing data and develop new corpora, in order to ensure the representativeness of the final data set.

The French CoMeRe project (CoMeRe, 2014)[1] is an ongoing pilot project whose deliverables will form part of the forthcoming *French National Reference Corpus*. It aims to build a kernel corpus of different Computer-Mediated Communication (CMC) genres with interactions in French as the main language. Three fundamental principles underlie CoMeRe: variety, standards and openness.

"Variety" is one of our key words since we expect to assemble interactions stemming from networks such as the Internet or telecommunication (mobile phones), as well as mono and multimodal, synchronous and asynchronous communications. Our interest covers genres such as text or oral chats, email, discussion forums, blogs, tweets, audio-graphic conferencing systems (conference systems with text, audio, and iconic signs for communication), even collaborative working/learning environments with verbal and nonverbal communication. A variety of discourse situations is also sought; public or more private conversations, informal, learning and professional situations. Part of our (sub)corpora is taken from existing corpora, since partners involved in the project had previously collected almost all the genres mentioned earlier. Other parts, such as Wikipedia talk pages, will be extracted from the Web following the recommendations of the *New Collections* workgroup.

"Standards" is our second key word. It refers to two different aspects of corpus linguistics. Firstly, corpora will be structured and referred to in a uniform way. The Text Encoding Initiative (TEI) format (BURNARD & BAUMAN, 2013) has been chosen, jointly with our European partners, alongside existing metadata formats including the Dublin Core. The TEI is not only a format for corpus structure. First and foremost, it is a model of text. This model needs to be extended in order to encompass the Interaction Space (IS) of CMC multimodal discourse, as we will discuss in Section 2. The European TEI-CMC (2013) special interest group aims to propose such extensions to the TEI consortium.

"Standard" also refers to the uniform basic level of automatic annotations, related to segmentation and Part Of Speech (POS) tagging. This will be applied to all of our CMC genres, and is presented hereafter in Section 3.

The third key word is "openness". At the end of the first stage (2013-2014) of the project, a sample of corpora (including those described in this paper, see Section 3.1) that is representative of CMC genres and that has been organized and processed in standard ways, will be released as open data on the French national platform of linguistic resources ORTOLANG (2013). 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).This openness is motivated, on the one hand, by the fact that CoMeRe will become part of the larger reference corpus for the French language. The latter is expected to become a reference for studies in French linguistics. On the other hand, the wish to release CoMeRe corpora as open data stems from the fact that, although studies on new CMC communication genres draw much attention, there is currently no existing dataset with

significant coverage or that encompasses a variety of genres to form the basis for systematic research. This situation is not specific to the French language as aforementioned languages, which already benefit from reference corpora, also face the same challenge. That being said, a few genre-based corpora are being developed (*e.g.* REHM et al. 2008). This may explain why a common motivation amongst European partners incited, from the outset, the design of a shared framework for the development of models of CMC genres. Indeed, there is a need for open access corpora that can be cross examined in order to exemplify the way models could be instantiated.

This OpenData perspective paves the way for scientific examination, replication and cumulative research. Of course, this type of openness implies specific considerations of licenses, ethics and rights, as discussed in Section 4.2. In order to achieve this goal, CoMeRe is supported by the research consortium *Corpus-Écrits* (2013), a subsection of the national infrastructure HUMA-NUM (2013, cf. Digital Humanities), and ORTOLANG, the French equivalent of DARIAH (2013), the European infrastructure for Humanities.

2. CoMeRe 2013: moving from existing data to models of CMC interaction

The CoMeRe project developed out of collaborations between researchers who had previously collected and structured different types of CMC corpora within their local teams. Once the project was officially underway, it was decided, building upon the SoNaR experience (OOSTDIJK et al., ibid), to organize workgroups (WG) with distinct tasks in the project: *TEI & Metadata , New Collections, Automatic Processing, Quality*.

The present section firstly describes four of the corpora that individual researchers brought to the CoMeRe project (Section 2.1). Secondly, we discuss how these four corpora helped the *TEI & metadata* WG to instantiate a model of CMC interaction, working collaboratively with the TEI-CMC SIG (Sections 2.2 and 2.3). Section 2.4 details how the same WG then structured corpora according to this model. The work of other WGs will be the focus of Sections 3 (*Automatic Processing* WG), 4 (*Quality*) and 5 (*New collections*).

2.1 Gathering existing data

Illustrations in this article will be based on the first four corpora processed by the CoMeRe project in fall 2013. They were collected within the frameworks of national and / or international projects. After their conversion to the new TEI format, they were renamed *cmr-smsalpes*, *cmr-smslareunion*, *cmr-simuligne* and *cmr-getalp_org*.

Our first corpus *cmr-getalp_org* (FALAISE, in print) is a **text chat corpus**, collected from a public Internet Relay Chat (IRC) website. Eighty different discussion channels focusing on a variety of, mainly informal, topics were collected in 2004. The corpus includes more than three million messages. The first version of the corpus (FALAISE, 2005) had been organized using a simple eXtensible Markup Language (XML) structure.

The data we organized in *cmr-smsalpes* (ANTONIADIS, in print) and *cmr-smslareunion* (LEDEGEN, in print) emanated from the international project "Faites don de vos SMS à la science" (FAIRON, KLEIN & PAUMIER, 2006) that began in 2004 and was coordinated by the Institute for Computational Linguistics (CENTAL) of the Catholic University of Louvain (Belgium). The project, named *sms4science*, aims to collect **SMS text messages** worldwide

(PANCKHURST et al., 2013). It regroups researchers from several countries to collaboratively conduct scientific research on a large number of languages with the objective of contributing to SMS message communication studies.

Data from *cmr-smslareunion* were issued between April and June 2008 within the framework of the first French investigation which led to the collection of 12,622 SMS messages sent by 884 participants. The *Laboratoire de recherche dans les espaces créolophones et francophones* (LCF) of the Université de La Réunion was responsible for the local coordination. As described in the project presentation (LaRéunion4Science, 2008), the particularity of the investigation in Réunion is the new scientific dimension that it adds: French-Creole bilingualism, the ludic neographies in SMS messages, and the communication practices of young people that are characterized by multiple alternating languages (French, Creole, English, and Spanish).

Data from *cmr-smsalpes* were collected in 2011 (ANTONIADIS, CHABERT & ZAMPA, 2011). The corpus includes 22,117 messages sent by 359 participants mainly living in the French Alps. The project was coordinated by the *Laboratoire de linguistique et didactique des langues étrangères et maternelles* (LIDILEM) of the Université Stendhal in Grenoble.

For both SMS text message corpora, the harvest of SMS messages required the intervention of technical partners. Indeed, the companies *Orange Informatique* and *Cirrus Private* were responsible for receiving the SMS messages and transferring them to the laboratories concerned. Researchers in charge of compiling data for the two corpora anonymized and structured the messages in different formats: XML for the French Alps corpus and in the form of a spreadsheet for the Réunion corpus. Note that researchers from Réunion also added manual annotations to the messages, providing orthographic transcription and language identification (either pidgin or French), as we will see further on.

Lastly, the *cmr-simuligne* corpus was built out of interaction data resulting from an online language learning course, Simuligne. Data have been extracted from the LETEC (LEarning and TEaching Corpus) Simuligne (REFFAY et al., 2009), a corpora deposited in the MULCE repository (2013), which has its own XML schema. Sixty-seven participants (language learners, teachers, native speakers — i.e., language experts) followed the same pedagogical scenario, but were divided into four groups. All interactions occurred within a Learning Management Systems (LMS), namely WebCT. They include text chat turns (7,000), emails (2,300) and forum messages (2,700). Since the LMS had no export facilities, data were extracted from its internal database by the LETEC corpus compiler, then structured and anonymized.

Such disparities in corpus compilation choices may have represented a major handicap for the CoMeRe project, particularly when in the Linguistics field many individual researchers still pose the question as to whether spending the time to make data shareable and accessible is worthwhile. However, data heterogeneity soon turned into a real asset, favoring exchanges between project participants concerning the data collection contexts and different ways of interpreting the data, as well as increasing our motivation to design a common model and share different areas of expertise.

2.2 Rationales for modelling CMC discourse

Before determining the TEI-compliant structural markup of the corpus, the *TEI & metadata* WG found it necessary to first settle on a common document model that would fit all of our

CMC data as well as new collections of data to be added to the corpus repository in the future. Indeed, annotation is basically an interpretation and the TEI markup naturally encompasses hypotheses concerning what a text is and what it should be. Although the TEI was historically dedicated to the markup of literature texts, various extensions have been developed for the annotation of other genres and discourses, including poetry, dictionaries, language corpora or speech transcriptions.

If one wants to still apply the word "text" to a coherent and circumscribed set of CMC interactions, it is not so much in the sense developed by the TEI. Indeed, it would be closer to the meaning adopted by BALDRY & THIBAULT (2006). These authors consider (ibid: 4) "texts to be meaning-making events whose functions are defined in particular social contexts," following HALLIDAY (1989:10) who declared that "any instance of living language that is playing a role some part in a context of situation, we shall call it a text. It may be either spoken or written, or indeed in any other medium of expression that we like to think of."

Bearing the above in mind, we found it more relevant to start from a general framework, that we will term "Interaction Space" (see next section), encompassing, from the outset, the richest and the more complex CMC genres and situations. Therefore, we did not work genre by genre, nor with scales that would, for instance, oppose simple and complex situations (*e.g.* unimodal versus multimodal environments) - as said, our goal is to release guidelines for *all* CMC documents and not for each CMC genre. This also explains why we did not limit ourselves solely to written communication. Indeed, written communication can be simultaneously combined with other modalities. For these reasons, the CoMeRe model takes multimodality into account and our approach is akin to the one adopted by the French research consortium IRCOM (2013). This consortium rejected the collection and study of oral corpora as self-contained elements and decided that it was preferable for oral and multimodal corpora to be studied within a common framework, before becoming part of the French reference corpus.

2.3 The notion of 'Interaction Space'

2.3.1 Interaction space: time, location, participants

An Interaction Space (henceforth referred to as IS) is an abstract concept, located in **time** (with a beginning and ending date with absolute time, hence a time frame) where interactions between **a set of participants** occur within an **online location** (see Figure 1 for a general overview). The online location is defined by the properties of the **set of environments** used by the set of participants. *Online* means that interactions have been transmitted through networks; Internet, Intranet, telephone, etc.

The set of participants is composed of **individual** members or **groups**. It can be a predefined learner group or a circumscribed interest group. A mandatory property of a group is the listing of its participants.

The range of types of interactions (and their related locations) is widespread. It is related to the environment(s) participants use and their corresponding modes and modalities.

2.3.2 Environment, mode and modality

An environment may be synchronous or asynchronous, mono or multimodal. **Modes** (text, oral, icon, image, gesture, etc.) are semiotic resources which support the simultaneous genesis of discourse and interaction. Attached to this sense of mode orienteered towards communication, we use the term **modality** as a specific way of realizing communication (this sense refers to the Human Computer Interaction field (BELLIK & TEIL, 1992)). Within an environment, one mode may correspond to one modality, with its own grammar that constraints interactions. For example, the icon modality within an audio-graphic environment is composed of a finite set of icons (*raise hand, clap hand, is_talking, momentarily absent*, etc.). In contrast, within an environment, one mode may correspond to several modalities: a text chat has a specific textual modality that is different from the modality of a collective word processor, although both are based on the same textual mode. Consequently, an interaction may be multimodal because several modes are used and/or several modalities (CHANIER & VETTER, 2006 ; see also LAMY & HAMPEL, 2007 for another presentation).

Environments may be simple or complex. On one end of the scale, we find simple types with one environment based on one modality (e.g. one text chat system in the *cmr-get-alp_org* corpus). On the other end of the scale, stand complex environments, such as the LMS of the aforementioned *cmr-simuligne* corpus, where several types of textual modalities are integrated, either synchronous — text chat — or asynchronous— email and forum —), or in 3D environments, where several modes and modalities appear (see hereafter).

An environment offers the participants one or more locations / places in which to interact. For example, a conference system may have several rooms where a set of participants may work separately in sub-groups or gather in one place. In a 3D environment such as the synthetic world *Second Life*, a location may be an island or a plot. A plot may even be divided into small sub-plots where verbal communication (through text chat or audio chat) is impossible from one to another. Hence we say that participants are in the same location / place if they can interact at a given time. Notions of location and interaction are closely related and are defined by the affordances of the environment.

2.3.3 Interaction

As previously described, participants in the same IS can interact (but not necessarily do it, cf. lurkers). They interact through input devices (microphone, keyboard, mouse, gloves, etc.), which let them use the modalities and output devices, mainly producing visual or oral signals. (These however, will not be described in this article). Hence when participants cannot hear nor see the other participants' actions, they are not in the same IS. Of course, participants may not be participants during the whole time frame of the IS. They can enter late, or leave early. Note that an IS may have a recursive structure: in an online course when the same participants interact over several weeks, different ISs will be created, correspondingly to different occurrences of interaction sessions.

In an IS, actions occur between participants. Let us call the trace of an action within an environment and one particular modality an "**act**". Acts are generated by participants, and sometimes by the system. Some of them may be considered as directly communicative (e.g. verbal acts in synchronous text or oral modalities). Others may not be directly communica-

tive but may represent the cause of communicative reaction / interaction (e.g. when participants write collaboratively in an online word processor and comment on their work). Participants see and hear what others are doing. These actions may represent the rationale for participants to be there and to interact (produce something collectively). Hence the distinction between acts that are directly communicative, or not, is irrelevant.

A verbal act may be realized as an *en bloc* message or as an adaptive one. For example, there are situations where a participant does not plan an utterance as a one-shot process before it is sent as an *en bloc* message to a server, which in turn displays it to the other participants as an non modifiable piece of language (e.g. as a text chat act which corresponds to what is generally called a chat turn) (BEISSWENGER et al., 2012). However, a participant's utterance (e.g. in an audio chat act) can also be planned, then modified in the throes of the interaction while taking into account other acts occurring in other modalities of communication (see WIGHAM & CHANIER, 2013 as an example).

If all the environments, corresponding to the four first corpora that we have processed, form the basis of our current presentation and even all these corpora correspond to messages sent *en bloc*, our IS model needs to take into account other corpora where this does not hold true. Within other multimodal environments from which we have already collected data and which we are currently processing, verbal (speech, text chat) and nonverbal acts occur simultaneously. The main purpose of transcriptions is then to describe inter-relations amongst acts and within acts.

2.4 Describing the interaction space within TEI

Since TEI was the format adopted by national research networks (*Corpus-écrits* and IRCOM) and by the European TEI-CMC SIG, the challenge faced by the *TEI & Metadata* WG was to firstly find out how information related to the IS could be described within the TEI header, and secondly, decide how, within the corpus body, verbal acts could be coded in such a way that all information included in the original version of each corpus be kept.

The choice to adopt TEI was also motivated by two different research interests that members of CoMeRe shared: research on NLP models and research on Discourse. The focus of these may appear quite different and although analysis work will only start once the CoMeRe corpora have been disseminated, it was important for the *TEI & Metadata* WG to keep both perspectives in mind when making TEI coding decisions.

One interest of CoMeRe members is to study linguistic idiolects occurring in different CMC genres. NLP algorithms are an indispensable prerequisite for this. However, it should be noted that NLP models may be developed solely on the contents of the verbal acts, whilst ignoring the rest of the IS. However, for other CoMeRe members interested in completing studies on Discourse, the IS is fundamental. This especially holds true if members want to later study research questions such as: how does discourse organization vary from one situation to another? What type of interaction supports or hinders discourse amongst participants? What features of participant groups influence online interactions? What are the relationships between discourse organization and language complexity? These are current topics investigated by researchers in fields such as Computer Supported Collaborative Learning (CSCL) and Computer-Assisted Language Learning (CALL).

The difference in the importance attributed to the IS when adopting one or other of these research perspectives seems, however, to be dialectical. Indeed research in CSCL and CALL may take advantage of linguistic annotations, which they previously have never considered, possibly because they had not been available to scientists in these fields.

We now move on to illustrate how the *TEI & metadata* WG encoded the IS in TEI in the four corpora (*cmr-smsalpes*, *cmr-smslareunion*, *cmr-getalp_org* and *cmr-simuligne*) whilst taking the above research perspectives into account. Figure 1 illustrates the different concepts we introduced and which have to be described in TEI. Note that element <u>, used in speech transcriptions and the new (not yet present in TEI) element <prod> used in non-verbal transcriptions will not be presented, because they do not occur in the corpora used here as examples.

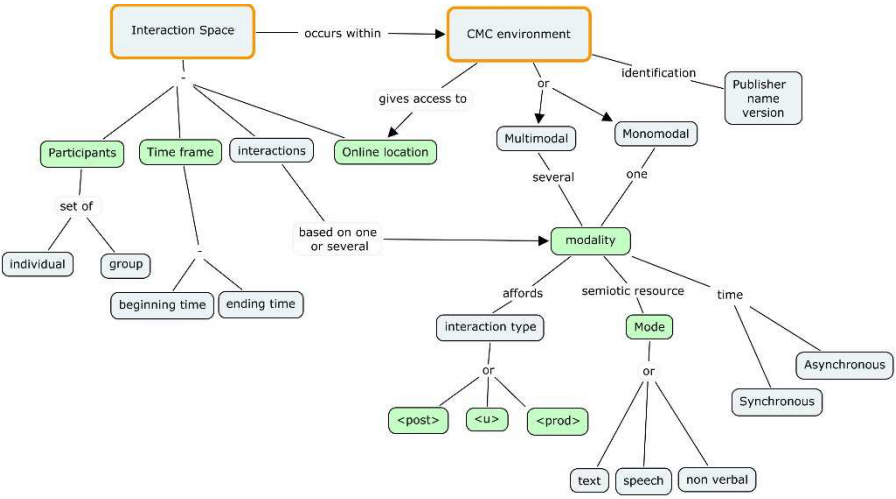

Figure 1: Description of concepts related to the Interaction Space.

2.4.1 Environments and affordances

The first step when describing an environment is to define within the <teiHeader> the general features attached to the overall environment type to which it belongs (e.g., IRC text chat systems). However, this needs to be refined in order to elicit specific features of the system. For example, Figure 2, (2a) describes, in TEI, the general text chat modality where inside one public channel[2] every connected participant may interact with the other participants in a spontaneous way through discussions held in informal settings, contrastingly to educative or professional discussions. Example (2b), however, details the affordances related to the specific IRC system used in *cmr-getalp_org*. This simplified extract displays the three main types of chat actions (message, command, and event), and part of the subtype of events. Relationships between this definition of the environment in the <teiHeader> and its actual

use by participants in interactions, described in the <body> part of the TEI file, will appear through the attribute @type of the <post> element (see next section).

```
(2a)
<textDesc xml:lang="en-GB">
    <channel mode="w" xml:lang="en-GB"><term
ref="#texchat-epiknet">text
chat</term></channel>
    <constitution>Messages typed by partici-
pants inside EpikNet IRC Channels and then
collected by Botstats.com </constitution>
    <derivation type="original"/>
    <domain type="public"/>
    <factuality type="fact"/>
    <interaction type="complete" ac-
tive="plural" passive="many"/>
    <preparedness type="spontaneous"/>
    <purpose degree="high"><note>Informal
discussion</note></purpose>
  </textDesc>
```

```
(2b)

<classDecl>
 <taxonomy>
   <category xml:id="texchat-epiknet" />
    <category xml:id="chat-message"/>
    <category xml:id="chat-command"/>
    <category xml:id="chat-event">
      <category xml:id="connexion" />
      <category xml:id="deconnexion"/>
      <category xml:id="changementpseudo"
/>
         [...]
```

Figure 2: TEI description of a text chat environment in the <teiHeader>

Figure 2 illustrates a mono-modal environment. Distinctively, when the environment is complex, such as the one related to *cmr-simuligne* corpus where interactions happened in ISs based on text chat, email, or forum modalities (cf. section 2.1), it is described in the same way into the <teiHeader> thanks to a more complex taxonomy: one category per modality and each category having its own text description (<textDesc>). Here again, each category corresponds to a type of message appearing in the body of the corpus.

Besides its multimodal environment, the *cmr-simuligne* corpus has another more complex organization. On the LMS platform, there were four distinct interaction spaces where groups of participants completed the same activities. The participants within one group could only communicate with members of that group. These top level ISs have been encoded as distinct TEI texts, and all of them included within a <teiCorpus> file. Every TEI text in *cmr-simuligne* is organized around sets of learning activities that are either simple or complex. A learning activity may include one or several modalities (email, chat or forums). The organization here is strikingly different to that adopted in other corpora. In *cmr-smsalpes*, *cmr-smslareunion*, and *cmr-getalp_org,* all messages are included within one division (element), whereas in *cmr-simuligne* there is one division per modality and a division may be nested several times.

2.4.2 A common post element

As agreed upon in the TEI-CMC SIG, we decided to use a common main new element, called "post" in order to encode all verbal acts produced by a participant in a textual monomodal environment, prepared in advance by its author and sent *en bloc* to the server. The macro-structure[3] of the post may vary from one modality to another. Every structure is detailed in the header of the TEI files and is accompanied by comments that are of foremost importance because they describe constraints that researchers will have to take into account when conducting future analyses.

Figure 3 provides a simplified extract of the <teiHeader> that describes the structure of a SMS message, specifying how time events and participants' identifications should be interpreted.

```
<tagsDecl>
      [...]
      <post>one post corresponds to one SMS.
        @xml:id  ID of the posting.
        @when corresponds to the date of the message collected by the system.
It depends on the date the participant sent to the system, but not the date of
the conversation. Accordingly, one participant may have sent his/her messages
to his/her correspondent at different times, but may have assembled her mes-
sages and sent them together to the server.
        @who is the anonymized telephone number. Hence one ID identifies one
participant over the whole corpus. If messages sent by the same participant
(sender) may be studied, it should be noted that we have no information about
the receiver.
       [...]</post>
</tagsDecl>
```

Figure 3: Simplified structure of the <post> element for an SMS message as described in the <teiHeader>

Figure 4, extracted (and simplified) from *cmr-simuligne*, describes the structure of email and forum messages. In the latter case, each message / <post> in a thread is either the first message of the thread or a response to another message within the thread. The difference is simply made by the XML attribute @ref: A message without a @ref opens a thread whereas a message which has a @ref is an answer to another message and is consequently included in a thread[4]. It has a title (<Title>), may have an attached file (<trailer>) and may also include a list of addressees (<listPerson>). When the message has been read (i.e. opened), this is noted within the structure (@type=Read). The name(s) of the reader(s), as well as the time(s) at which the message was read appear. The latter information is important when studying networks of participants interacting in a group (see, as an example, a CSCL analysis based on *Social Network Analysis* in (REFFAY & CHANIER, 2003).

```
<tagsDecl>
        [...]
  <post>one post corresponds to one email message or one forum message or one
text chat act.
          @xml:id ID of the post.
          @when date of the message when created, given by the system
          @who id of the author of the message.
          @type type of the post cf. taxononomy.
          @ref reference to the post ID to which the current post responded to
(for email and forum)
          <head> contains all the rest of the structure of the post, which can-
not be described as TEI elements.
             <title> Title of the forum, or subject of an email.
             <listPerson> list of people who received / read the post.
                    @type=SendTo addressee(s) of an email
                    @type=Read who opened (read?) an email or a forum message?
                   [...]
         <trailer>At the end of a post when there is an attached file
 </tagsDecl>
```

Figure 4: Simplified presentation of the structure of an email or a forum message in the <teiHeader>

2.4.3 Locations and time frame

Locations and time frame are also components of the IS. Different notions of locations need to be distinguished: the server location where data was collected firsthand; locations attached to a modality (e.g., distinct chat rooms or channels); locations of participants (leaving areas, see hereafter). Information on time is given at the level of the IS and also with every post. It is an indispensable component of the data, not only for studying interactions within one IS, but also for the study of group or individual activities within the overall corpus (see for example tools for displaying discussion forum time lines (CALICO, 2013). For space reasons, we shall not detail here how locations and time frames have been encoded in TEI.

2.4.4 Participants

Since CoMeRe has collected different CMC genres, we have a large variety of participant description types — types which highly constrain further research analysis. On the one hand, in *cmr-smsalpes* and *cmr-smslareunion*, the only information we have about each participant is her/his identification number and information on his/her location given at a regional level (respectively in French Alpes or Réunion). On the other hand, in *cmr-simuligne*, we have access to detailed information about participants (individuals and groups), as shown in Figure 5. An individual female learner, aged 51, who is affiliated to The Open University and who has adopted the alias *Alba* is detailed, as well as information about a learner group.

```
<particDesc>
  <listPerson>
    <person role="learner" xml:id="Gl1">
       <sex>female</sex><age value="51"/>
       <residence>United Kindom</residence>
       <affiliation>The Open University</affiliation >
       <persName><addName type="alias">Alba</addName></persName>    </person >
                 [other participants]
    <personGrp role="learnerGroup" xml:id="Simu-g-Ga">
        <persName><addName type="alias">Gallia</addName></persName></personGrp>
              [other groups]
     <listRelation corresp="#Simu-g-Ga">
        <relation type="social" name="tutor" active="#Gt"/>
        <relation type="social" name="native" active="#Gn1 #Gn2"/>
        <relation type="social" name="learner" active="#Gl1 #Gl2 #Gl3 #Gl4 #Gl5
#Gl6 #Gl8 #Gl9 #Gl10"/>
        <relation type="social" name="researcher" active="#Tm"/></listRelation>
```

Figure 5: Description of one participant, one group and relationships within a group

A common requirement in corpus linguistics is to associate each individual with a single identification code throughout the corpus. In CMC corpora, this is not always easy to achieve. On the one hand, in corpora built from experiments with a limited number of participants, such as *cmr-simuligne*, it was a tedious process to identify each participant every time s/he was named in a post (see example in a message forum in Figure 9). On the other hand, in a public chat channel, it may be difficult to identify participants due to constant changes in their alias names. In one case, analysis of individual contributions, activities, language level, lexical diversity, etc. can become an object of study. In the latter case, it is the variation in alias names which may be interesting to study: see Figure 6 taken from *cmr-getalp_org*

where one participant uses suffixes attached to her/his alias in order to reflect different states of mind or activities (e.g. sport, school, busy, away, etc.).

```
<person xml:id="cmr-get-c027-p4215">
  <persName>
      <addName type="alias">Farlin</addName>
      <addName type="alias">Farlin[AuStade]</addName>
      <addName type="alias">Farlin[AwAy]</addName>
      <addName type="alias">Farlin[IRL]</addName>
      <addName type="alias">Farlin[Lycee]</addName>
      <addName type="alias">Farlin[OqP]</addName>
      <addName type="alias">Farlin[Oral]</addName>
      <addName type="alias">Farlin[PALA]</addName>
       [...]
```

Figure 6: Variety of aliases chosen by one participant in a text chat

2.4.5 Examples of posts

Let us now consider examples of messages sent through different modalities. Whereas affordances of the Interaction Space were described previously in the <teiHeader>, here we discuss corpora bodies (element <body>).

Text chat

One of the interests of assembling heterogeneous corpora is to be able to step back from some forms of oversimplification. One such idea is that on the Internet there is one language, often called Netspeak (CRYSTAL, 2001). Figure 7 shows two messages uttered in the same modality, text chat: (7a) is an extract from *cmr-simuligne* and (7b) from *cmr-getalp_org*. Whereas the author of (7b) types as if s/he were sending an SMS - writing some words such as 'vé' phonetically and not using the plural 's', for example in 'les equation' - the author of (7a), a learner of French, seeks to type full sentences. In the latter message, well-formedness is only endangered by a lack of knowledge in the target language or by the speed of typing which may cause typos e.g. 'hueres' in (7a) rather than 'heures'. (7a) is prototypical of CALL interactions where topics such as lexical or grammatical diversity can be studied in comparison to the target language spoken offline. Whether (7b) is prototypical of text chat or only reflects an idiosyncratic behavior is a research question in itself.

(7a)	(7b)
<post xml:id="cmr-Simu-Chat_Lugdunensis_Room1_S47_00528" when-iso="2001-05-11T12:30:13" who="#cmr-Simu-L18" type="chat-message"> Le bateau est ammare a St Helier dans un marina qui s'ouvre seulement trois hueres avant la maree </post>	<post xml:id="cmr-get-c043-a21693" when-iso="2004-03-18T14:09" who="#cmr-get-c043-p39174" alias="cortex_taff" type="chat-message"> Apres je vé faire ma physique c aussi les equation bilan </post>

Figure 7: Linguistic diversity in text chat acts

SMS

Idiosyncratic ways of communication within a specific modality have also been identified within our SMS corpora. Messages (8a) and (8b) were sent by the same author, who regularly

introduces spaces into her/his message, whereas (8c) and (8d) come from another author following a serious conversation with her/his correspondent. As we will later see in Section 3, both the whitespaces in (8a) and (8b) and the abbreviations and agglutination-abbreviations in (8c) and (8d) will pose issues for the process of automatic annotation of the corpora.

```
(8a) <post xml:id="cmr-slr-c001-a11644" when-iso="2008-06-16T11:59:00"
who="#cmr-slr-c001-p868" type="sms">
      <p>à k e l a d r e s e n v o y e r d e s f l e u r s ?</p>
    </post>
     […]
(8b) <post xml:id="cmr-slr-c001-a11647" when-iso="2008-06-16T12:00:39"
who="#cmr-slr-c001-p868" type="sms">
      <p>e n f o n t d e n t i s t e</p>
    </post>
     […]
(8c) <post xml:id="cmr-slr-c001-a00011" when-iso="2008-04-14T10:17:11"
who="#cmr-slr-c001-p010" type="sms">
      <p>é@??$?Le + triste c ke tu na aucune phraz agréabl et ke tu va encor me
dir ke c moi ki Merde par mon attitu2! Moi je deman2 pa mieu ke klke mot agré-
abl échangé</p>
      […]
(8d)  <post xml:id="cmr-slr-c001-a00304" when-iso="2008-04-15T20:23:59"
who="#cmr-slr-c001-p010" type="sms">
      <p>.2 te comporter comme ca avec moi. Je ve bien admettr mes erreur kan
j'agi vraimen mal comm hier mé fo pa exagérer. Si t pa d'accor c ton droi. Si
tentain.le rest c à dirreposer dé question sur 1 sujet déjà expliké c pa 1 rai-
son valabl pr ke tu te monte contr moi.pr moi ossi ca suffi.</p>
```

Figure 8: Different composition of graphemes and lexical items between two authors of SMS messages.

Forum

As shown in Figure 9, the structure of a forum message is more complex. The example in this figure is taken from *cmr-simuligne*. The author of the message is a native speaker of French who is replying to a post made by a learner of French. Each person mentioned has been identified in the message structure (author, list of readers -here shortened-) and in its contents (signature of the author). This information may lead to other types of research on discourse and group interactions. For example, who takes the position of a leader, or an animator within a group? Can subgroups of communication be traced within a group, thanks to an analysis of clusters, cliques (REFFAY & CHANIER, ibid)?

```
<post xml:id="cmr-Simu-Gall_e2a2_hymne-234" when="2001-06-06T08:17:00"
who="#cmr-Simu-Gn2" type="forum-message" ref="#cmr-Simu-Gall_e2a2_hymne-209">
  <head>
     <title>constitution des groupes</title>
     <listPerson>
        <person corresp="#cmr-Simu-Gt">
           <event type="Read" when="2001-06-06T08:17:00">
           <label>Read</label>  </event> </person>
           [….] </head>
   <p><name ref="#cmr-Simu-Gl4" type="person"><forename>Nick</forename></name>,
Est ce que c'est de l'humour anglais? Tu risques de le regretter Amicalement
<name ref="#cmr-Simu-Gn2" type="person"><forename>Laurence</fore-
name></name></p></post>
```

Figure 9: Message posted in a forum

Encoding manual annotations

Finally, Figure 10 illustrates another challenge faced by the CoMeRe editors when elaborating the TEI schema: the inclusion of manual annotations by researchers within the corpus. In *cmr-smslareunion,* a large number of SMS mix French from France (French-fra) with French pidgin from Réunion (pidgin-cpf). The content of the post before the <reg> element (which is a standard element belonging to the core TEI) corresponds to the actual message sent. The contents of the <reg> includes the researcher's manual annotations as s/he tries to identify, with various degrees of certainty (cf. @cert), whether part of the message is in French-fra or pidgin-cpf, and who, at the same time, transposes various segments into a more standard orthography.

```
<post xml:id="cmr-slr-c001-a2860" when="2008-05-01T09:49:36" who="#cmr-slr-
c001-p000424" type="sms">
   <p>Oui ver20h mc do st benoit vu ke mi mange la ba. tu mange avan de venir?
Tu me sone kan t la?</p>
   <reg type="transortho"><seg xml:lang="fra" cert="medium">Oui vers 20h Mac Do
Saint Benoît vu que </seg> <seg xml:lang="cpf">mi manj la ba.</seg> <seg
xml:lang="fra">Tu manges avant de venir ? tu me sonnes quand t'es
là ?</seg><add type="F"><seg xml:lang="cpf" cert="low"> Wi vèr 20h Mac Do Sin
Benoi  vu ke</seg> </add> <add type="trad"> <seg xml:lang="fra">je mange là-
bas</seg> </add> </reg>
 </post>
```

Figure 10: Annotation of a SMS

The challenge here was to find out how the researcher's annotations, contained within a spreadsheet, could be kept and coded into TEI. The next challenge is to measure the extent to which these manual annotations will correspond to automatic annotations made during the next phase of our project.

3.Automatic corpora annotations

Drawing on previous NLP experience applied to various types of linguistic data issued from social media, the *Automatic Processing* WG is in charge of processing the first layer of annotations on TEI-compliant corpora. This project stage will begin in Spring 2014. In this section, we present our motivations for applying an automatic annotation process to the CoMeRe corpora (Section 3.1) before turning to the decisions made concerning which annotations to make for which units (Section 3.2) and to a description of the processing pipeline for adding such annotations to the CoMeRe data (Section 3.2).

3.1 Motivations

If the usefulness of corpora has already been proven in numerous studies and applications, the real value of these corpora relies, most of the time, on the quantity and quality of the information that has been added to them. This information (as annotations) allows content characteristics that are useful (and often essential) for operational use to be highlighted. For example, knowing the grammatical nature of the "words" of a text chat or SMS corpora allows the syntactic structure of each element of the corpora to be identified, as well as the possibility to calculate the vocabulary used and analyze the syntactic or semantic context of a word or class of words, etc.

Depending on the nature of annotations, they can be added automatically, when possible, or manually with the help of appropriate interfaces. The often high cost of manual annotations represents a real handicap for their elaboration. Most of the time, only automatic annotations are used, due to limited budgets that cannot allow for manual and better descriptive ones. The CoMeRe corpora do not overcome this constraint. Provided by the project partners (corpus compilers), the corpora can contain annotations added by the compilers (as detailed in Section 2.4). One goal of the CoMeRe project is to automatically add additional annotations that will prove useful to improve the operational use of the CMC corpora.

Our starting point for this automated annotation processing is based on anonymized initial corpora that partners brought to the CoMeRe project. Anonymisation of the corpora had previously been completed by the compilers. However, the anonymisation rules were often different from one corpus to the next, and have therefore been made consistent across corpora.

The automated processing of annotations that was performed concerns the textual corpora (or part of them), regardless of the text's form (standard French, text chat, SMS, etc.). Its purpose is to split the interactions into minimal textual units and associate each of them with a label representing their membership to specific morphosyntactic classes as well as additional information, for example, the lemma associated with each unit. This processing is based on automated language processing procedures and techniques.

If the CoMeRe annotated corpora are to be used by any researcher for his/her own personal research questions (see Section 2.2), the set of morphosyntactic labels used (as well as the associated information) must be as "consensual" as possible. Ideally, they must be able to be projected/transformed into the specific model the researcher wants to use, without requiring extensive work and calculation. Even though such a configuration currently seems quite difficult to determine (does it even exist?), our goal is to get as close to this as possible, using a set of labels and "generic" associated information, susceptible to be understood, used and transformed at a minor cost. We especially think that the association of a lemma to each minimal unit should allow for easier "customization" for researchers to conduct future studies on the contents of the CoMeRe repository.

This need for generic annotations led the *Automatic Processing* WG not to consider any processing beyond morphosyntactic labelling. Therefore, even though syntactic annotations (components, dependencies, etc.) could be considered, the diversity and specificity of existing syntactic analysis model undermines our concern for "genericity" and substantially handicaps any use/adaptation of such annotated corpora.

As part of the CoMeRe corpora consists of text with freely variant spelling (see examples in 2.4), the robustness of the processing tools used is an important factor for their choice. Indeed, they must allow us to automatically process (annotate) any element of these corpora, regardless of the level of variation: misspelling, agglutination (e.g., "*cp*" instead of *je ne sais pas* 'I do not know'), phonetic spelling (e.g. "*2m1*" instead of *demain* 'tomorrow'), shortened elements (e.g. "biz" instead of *bises* 'kisses'), etc. These occurrences are present in several, if not all, of the CoMeRe corpora. Furthermore, they often represent the majority of the interactions within a corpus, e.g. the *cmr-smsalpes* corpus. Tools with such robustness are currently quite rare for morphosyntactic processing of French texts; they are close to non-

existent (in the form of complete and autonomous tools) for syntactic processing/annotation. This aspect is an important reason for not processing and annotating the CoMeRe corpora beyond a morphosyntactic level.

3.2 Which annotations for which units?

One of the first parts of processing consists of marking off the "processing units" (most of the time equivalent to a sentence) of the text, in order to apply the same processing to each of them. If splitting "normal" texts into these units does not pose any major problem (except for some specific cases), things are a bit different when it comes to CMC data. These corpora include interactions that only contain partial punctuation, if any. Moreover, it is usually based on punctuation elements that the splitting into units is done. Based on this observation, the processing hypotheses and the processing itself that we apply to each type of corpora are different. For corpora with punctuation that is often missing (SMS, text chat, tweets) our processing unit will be each post; no splitting will be performed. Each SMS, tweet or text chat message will be considered the final unit. For the rest of the corpora (email, forum messages, etc.), content will be split into processing units akin to a sentence and annotated accordingly. We are aware that the absence of clear unit delimitation marks can result in troubles with the processing of further elements of these corpora, for example syntactic analysis.

Apart from the definition of the processing unit, the type of processing/annotations that we apply to the corpora (morphosyntactic annotations) requires the definition of the typographic unit, to which annotations can be associated. The targeted annotations being linguistic, they can only be obtained by relying on the linguistic notion of lexical unit (lexeme), which is, however, hard to automate due to the variety of possible ambiguities. For standard texts, these lexical units are often assimilated to units defined purely typographically, units that we will call *tokens.* These tokens are simply defined as a sequence of characters (excluding punctuation and spaces) preceded and followed by a space a punctuation mark. The morphosyntactic taggers thereby consider the tokens as lexical units based on which language calculations can be performed to select the correct labels. The same goes for the lemmatizers.

This purely typographic approximation of the splitting into lexical units is very simple to obtain automatically. However, this process will not suffice for corpora that contain non-standard text. Indeed, putting aside the partial or complete absence of punctuation, other phenomena, for example abbreviations ("*bis*" or "*biz*" for *bises* 'kisses') or agglutination-abbreviations ("*chépa*" for *je ne sais pas* 'I do not know,' "*mdr*" for *mort de rire* 'LOL', "*ct*" or "*c t*" for *c'était* 'it was'), prevent any identification of lexical units and tokens, even in an approximative way. Following (partially or totally) the approach used in similar work (Fairon & Paumier, 2006; Cook & Stevenson, 2009; Chabert et al, 2012), the *Automatic Processing* WG decided upon the following: the tokens will receive the annotations but these annotations will provide as much information about the underlying lexical units as possible. As a consequence, "*chépa*" or "*ct*" will be considered as tokens, but will need to be annotated, through linguistic information describing the complexity of their correspondence to the lexical units to which they are linked.

In order to obtain such annotations, some kind of mapping between tokens and (an approximation of) lexical units is required, as only the sequence of lexical units could be successfully tagged by existing POS taggers. This raises a new question: what kind of lexical units should we try and associate with observable tokens? Today, the answer to this question results from the following fact: virtually all POS taggers are trained on edited corpora (often journalistic data). This means that for now, the easiest way to get an acceptable POS-tagging and lemmatization accuracy on CMC data is to *temporarily* transform the data so that it appears as "edited" (as journalistic) as possible - in order for the POS tagger and the lemmatizer to be applied, and then to project the resulting information on the original text.

3.3 Processing pipeline

The processing pipeline used in CoMeRe implements the ideas presented in Section 3.2. It has previously been applied to CMC data in two different ways: as a pre-annotation tool on French (SEDDAH *et al*., 2012a) and as a pre-parsing processing tool on English (SEDDAH *et al.*, 2012b). It can be summarized in the following steps, which we criticize and illustrate below:[5]

- **Pre-processing** step: We first apply several regular-expression-based grammars taken from the SxPipe shallow analysis pipeline (SAGOT & BOULLIER, 2008) to detect smileys, URLs, e-mail addresses, Twitter hashtags and similar entities, in order to consider them as one token even if they contain whitespaces.
- **Tokenization** step: The raw text is tokenized (i.e., split into typographic units) and segmented into processing units which play the role usually devoted to sentences (see above), using the tools included in SxPipe.
- **Normalization** step: We apply a set of 1,807 rewriting rules,[6] together with a few heuristics that rely on a list of highly frequent spelling variations (errors or on-purpose simplifications) and on the L*efff* lexicon (SAGOT, 2010). The number of "corrected tokens" obtained by applying these rules might be different to the number of original tokens. In such cases, we use 1-to-*n* or *n*-to-1 mappings. For example, the rule *ni a pa* → *n'_y a pas '[there] isn't'* explicitly states that *ni* is an amalgam for *n'* and *y* (negative clitic and locative clitic, which will be POS-tagged and lemmatized as two distinct lexical units), whereas *a* should be left unchanged in this context (the lexical unit matches the typographic unit), and finally *pas* is the correction of *pa* (negative adverb, approx. 'not').
- **Annotation** step: Lexical units are POS-tagged and lemmatized using standard tools — in our case, the standard French model from the MElt tagger (DENIS & SAGOT, 2012) and the associated lemmatizer. This POS-tagging model was trained on the French TreeBank (FTB; ABEILLÉ *et al.*, 2003), "UC" version (FTB-UC), and on the L*efff* lexicon (see DENIS & SAGOT (2012) for details).
- **Post-annotation** step: We apply a set of 15 generic and almost language-independent manually-crafted rewriting rules that aim to assign the correct POS to tokens that belong to categories not found in MElt's training corpus, i.e., in FTB; for example, all URLs and e-mail addresses are post-tagged as proper nouns whatever the tag provided by MElt; likewise, all smileys get the POS for interjections.

- **Denormalization** step: We assign POS tags and lemmas to the original tokens based on the mappings between "normalized" lexical units and original token. If a unique lexical unit is associated with more than one original token, all tokens except the last one are assigned the tag *Y* and an empty lemma. The last token receives the tag of the lexical unit and its lemma. If more than one corrected tokens are mapped to one original token (non-standard contraction), it is assigned a tag obtained by concatenating the tags of all the lexical units, separated by the '+' sign. The same holds for lemmas. This convention is consistent with the existing P+D and P+PRO tags, which correspond to standard French contractions (e.g., *aux* 'to the(plur)', contraction of *à* 'to' and *les* 'the(plur)'). If the mapping is one-to-one, the POS tag provided by MElt for the lexical unit is assigned to the corresponding token.

We shall now illustrate this process by way of three examples; first, a single (contracted) token, then a simple non-standard compound and, finally, a whole sentence. Let us first consider the token *chépa* 'dunno'. Steps one and two (pre-processing, tokenization) have no particular effect on it. Step three normalizes this token by associating it with four lexical units, namely *je ne sais pas* 'I do not know.' Steps four and five POS-tags and lemmatizes these lexical units, thus producing, for example, the output *je*/CLS/*je ne*/ADV/*ne sais*/V/*savoir pas*/ADV/*pas*.[7] Then step six denormalizes this output by associating these POS tags and lemmas on the single input token, thus producing the following output: *chépa*/CLS+ADV+V+ADV/*je+ne+savoir+pas*.

Let us now consider the sequence *l'après midi*. It contains three tokens, *l'*, *après* and *midi*. The underlying lexical units are *l'* 'the' and *après-midi* 'afternoon'. In other words, the two last tokens are a non-standard compound. The result of step three is *l' après-midi* thanks to an adequate normalization pattern, and step five outputs *l'*/DET/*le après-midi*/NC/*après-midi*. Then step six applies the convention mentioned above for compounds while denormalizing: *l'*/DET/*le* is unchanged, the token *après* receives the special tag Y and no lemma, and the last token of the compound, *midi*, gets the tag of the corresponding lexical unit, NC, and the full lemma *après-midi*. Hence the final output: *l'*/DET/*le après*/Y/ *midi*/NC/*après-midi*.

Before moving on to the last example, it is important to be aware of the following three points concerning this approach. First, there is no clear-cut way of deciding what should be normalized and what should not. Second, normalization can be sometimes achieved in different ways. For example, *chépa* could be normalized as *je sais pas* (informal) or *je ne sais pas* (standard, formal, would be used in journalistic data). For these two points, the answer is the same: as the normalization is only temporary (just for the POS-tagger and lemmatizer to work), the general guideline is to "normalize" everything that departs from standard (journalistic) French in such a way that it matches as closely as possible standard (journalistic) French. The third point worth mentioning is that the mapping between tokens and lexical units can be very strange. For example, let us consider the sequence *c t*. This sequence can be interpreted by actually pronouncing the name of both letters, which produces /setɛ/, the valid pronunciation of *c'était* 'it was,' which is composed of two lexical units, *c'* 'it' and *était* 'was'. Note that this mapping means that the token *c* corresponds to *c'é-* whereas the token

t corresponds to *-tait*. There is therefore no direct correspondence between the original tokens and the underlying lexical units that are to be POS-tagged and lemmatized. In such a situation, we consider that there is no other way but to consider both tokens as forming a *de facto* compound *c_t* that is itself the (nonstandard) contraction of *c'* and *était*. As a result, we tag and lemmatize it as *c/*Y*/ t/*CLS+V*/ce+être*.

Keeping this in mind, we can move on to our last example, a (simplified) sentence from the *French Social Media Bank*, found on a forum from the website DOCTISSIMO (2013) that provides health-related information: "*sa fé o moin 6 mois qe les preliminaires sont sauté c a dire qil yen a presk pa* Foreplay has disappeared for at least 6 months, that is there is almost none." Table 1 illustrates the whole process by providing the output of steps three, five and six together with the tokenized input (output of step two).

Within the CoMeRe project, this processing pipeline has already been tested and improved (for instance, the pre-annotation pipeline (used for developing the *French Social Media Bank*) used 327 instead of 1,804 normalization rewriting rules). There is still room for improvement, and applying it systematically to the various CoMeRe corpora will certainly lead to further modifications and improvements. Note that CoMeRe will use this processing pipeline in a way that is similar to its use for developing the *French Social Media Bank*, i.e., as a pre-annotation tool. In other words, because the goal will be to have the best possible annotations on a well-defined set of data, we intend to improve our normalization rules and maybe MElt's (DENIS & SAGOT, 2012) training data by exploiting the very data we want to annotate. This is very different from a standard setting where one expects to process new data as well as possible, thus preventing target data from being used in any way during the development or training of the tools.

Tokenized text (output of step two)	Normalized text (output of step three) *whitespaces denote additional lexical unit boundaries*	POS-tagged and lemmatized normalized text (output of step five)	Final output (after step six)
sa	*ça*	*ça*/PRO/*ça*	*sa*/PRO/*ça*
fé	*fait*	*fait*/V/*faire*	*fé*/V/*faire*
o	*au*	*au*/P+D/*à+le*	*o*/P+D/*à+le*
moins	*moins*	*moins*/ADV/*moins*	*moins*/ADV/*moins*
6	*6*	*6*/DET/*6*	*6*/DET/*6*
mois	*mois*	*mois*/NC/*mois*	*mois*/NC/*mois*
qe	*que*	*que*/PROREL/*que* (erroneous POS tag, should be CS)	*qe*/PROREL/*que*
les	*les*	*les*/DET/les	*les*/DET/les
preliminaires	*preliminaires* (the missing acute accent on the first *e* has not been restored)	*preliminaires*/NC/*preliminaire* (despite the missing acute accent, the POS tag is correct, but not the lemma)	*preliminaires*/NC/*preliminaire*
sont	*sont*	*sont*/V/*être*	*sont*/V/*être*
sauté	*sautés*	*sautés*/VPP/*sauter*	*sauté*/VPP/*sauter*
c	*c'est-à-dire*	*c'est-à-dire*/CC/*c'est-à-dire*	*c*/Y/
a			*a*/Y/
dire			*dire*/CC/*c'est-à-dire*
qil	*qu' il*	*qu'*/CS/*que il*/CLS/*il*	*qil*/CS+CLS/*que+il*
yen	*y en*	*y*/CLO/*y en*/CLO/*en*	*yen*/CLO+CLO/*y+en*
a	*a*	*a*/V/*avoir*	*a*/V/*avoir*
presk	*presque*	*presque*/ADV/*presque*	*presk*/ADV/*presque*
pa	*pas*	*pas*/ADV/*pas*	*pa*/ADV/*pas*

Table 1: Automatic correction and annotation (POS tags, lemmas) for a very noisy sentence extracted from the *French Social Media Bank* (SEDDAH *et al.*, 2012a). Errors produced by the pipeline are mentioned.

The way the processing pipeline described above shall be used in CoMeRe is twofold:

- A fully automatic setting: the whole pipeline will be applied. The resulting annotations might be kept as such or might be manually corrected afterwards.

- A semi-automatic setting: for some corpora, such as *cmr-smsalpes* (ANTONIADIS, 2013), manual normalization was performed, in a way that is approximately compatible with the objectives of step three. In such a setting, the manually normalized data is provided as an input to steps one and two, step three is skipped, steps four and five (tagging and lemmatization) are applied, and step six is replaced by an *a posteriori* alignment step, in order to dispatch the resulting annotations in the original data (before manual normalization). This alignment step has not yet been developed. However, we believe we can achieve it based among others on the set of normalization rewriting rules used by step three.

CoMeRe's automatic annotation process raises several issues, especially important on noiser corpora (SMS, text chat, etc.), which will be mentioned in the conclusion.

4. Quality control and dissemination

All the data collected in the CoMeRe data bank (CoMeRe Repository, 2014), as well as annotations added to the CMC corpora detailed in Section 3, are verified by the *Quality* WG before the public release of the corpora and their dissemination at the end of 2014. In this current section we detail these two processes. Firstly, in Section 4.1 CoMeRe's staged process of quality control that allows a corpus to move from one project phase to the next. Secondly, in Section 4.2, we describe the planned dissemination of CoMeRe which is scheduled for the end of 2014. We also highlight questions this raised for members of the *TEI and metadata* WG concerning the acknowledgement of individual researchers' work in both the metadata and corpus reference, as well as the need for appropriate licenses for our corpora.

4.1. Corpus quality control process

For the production of any corpus, quality control is an essential aspect, particularly when a corpus undergoes format conversions. As REYNAERT *et al.* state, quality control should "take place all along the production timeline of the resource, rather than being put as a final check at the very end of corpus completion" (2010:2697). Within the CoMeRe project, quality control is a staged process that allows a corpus to move from one phase of the project to the next.

A first validation step occurs when the corpus compiler deposits the original corpus in the CoMeRe repository. The nomenclature for this version is *corpusname-v0.* At this stage, a member of the *Quality* workgroup checks that the information concerning the corpus license, the corpus size, the context in which data was collected and descriptions of any previously performed anonymisation processes has been supplied, as well as the legibility of corpus files. Requests for additional information from the compiler are handled. Once these criteria met, the corpus moves on to the TEI conversion phase.

Once the corpus converted into TEI, it is deposited in the *corpusname-v0* server space under the nomenclature *corpusname-tei-v1.* The corpus then undergoes a second quality control process during which the metadata in the TEI header is firstly validated in relation with the information provided by the corpus compiler. At this stage, the corpus description in both English and French is checked alongside the bibliographic reference for the corpus and the encoding of different participant roles and the description of the corpus license. Secondly, the description of the anonymisation process is then compared to the information

supplied by the corpus compiler and the identification of the corpus' interaction participants is verified. In a third step, the quality workgroup then proceed by randomly selecting a certain number of <post> elements with the original contents in *corpusname-v0* in order to check that no information has been lost in the TEI conversion process. After any back and forth exchange between the corpus compilers and data inputters the corpus is then validated. The validated version moves into the *corpusname-v1* server space and the automatic annotations phase is set in motion.

Once automatic annotations have been completed, a final quality control occurs during which the version *corpusname-tei-v1* and the post annotation corpus version are compared to ensure that no information has been lost. That the person who performed the annotations has been correctly cited in the metadata and that the annotation process has been included in the corpus description are verified. Again, a selection of <post> elements are chosen and compared between the two versions in order to ensure that no interaction information has been lost. This validation is also directed towards the correctness of the annotations. Once this final quality control validated, the corpus moves into the *corpusname-tei-v2* server space and both the versions *corpusname-v1* and *corpusname-v2* are then deemed ready for dissemination and are deposited on the national server, ORTOLANG.

At the time of writing, the first stage is achieved where the four corpora previously mentioned are concerned. The *Quality* group has started its work in order to assess the version *corpusname-tei-v1,* before the automatic annotations scheduled for the upcoming months.

4.2 The dissemination of CoMeRe

As mentioned, CoMeRe corpora will be released at the end of 2014. Meanwhile, new corpora from the *New Acquisitions* WG are under process (see the next section for details) and will be integrated into the CoMeRe repository hosted by ORTOLANG.

ORTOLANG (2013) is a new national infrastructure network for whom the objective is, firstly, to allow linguistic data in French (lexicons, corpora, dictionaries) and NLP tools to be disseminated amongst the international community of researchers in Linguistics. Secondly, a selection of these data will be saved permanently by another national infrastructure (CINES) who has been mandated to save top-priority French research data in all scientific fields. This data storage is expensive: notably because files need to be converted into different formats regularly, as certain current formats may soon become obsolete.

The dissemination of CoMeRe corpora in open-access formats imposes some specific constraints because our corpora will join other corpora deposited in ORTOLANG that have been prepared within other national projects. All corpora deposited in ORTOLANG will be structured in TEI and made accessible through an interface that is still under development. The latter will allow users to perform linguistic queries using concordancers, lexicometric and morphosyntactic tools, similar to the one found on the query interface of the German DWDS (2013) corpus. Variations in TEI formats within the range of corpora deposited in ORTOLANG are foreseen. This requires every project to document, in detail, the specific TEI structures used to format their corpora, particularly if any further conversions need to be made to facilitate corpora incorporation into the query interface. Releasing corpora in open access formats also requires the provision of specific information for each corpus concerning the protection of author rights and that future users circumscribe to ethical reuse of the corpora.

Where the CoMeRe project is concerned, we have made some progress towards meeting ORTOLANG's requirements. Firstly, our IS model has been carefully documented in the header of every TEI file, as previously explained. Other metadata were added, detailing how data was collected as well as how ethics and rights were respected. Secondly, in order to encourage data reuse, following the philosophy of OpenData (2013), we have decided to release our corpora under Creative Common licenses or others that are closely related. This includes possibly accepting terms for commercial use (i.e., discarding the Creative Commons' NC option) and the creators waiving their intellectual property rights (CC0 license). We therefore had to ensure that all members' work was given scientific acknowledgement; both within corpus metadata and by way of a specific bibliographic reference attributed to the corpus.

The need to acknowledge the time spent by researchers in compiling and structuring corpora is a well-known, if not always respected, issue in corpus linguists. In order to acknowledge the contributions made by different members of the CoMeRe project, the *TEI and Metadata* WG chose to use standard and precise terminology to encode participants' roles in each corpus. The OLAC format was adopted for this. This format is an overlayer of the Dublin Core, an ISO standard that is made up of 15 generic tags that, if need be, can be refined. Figure 11 is an extract of the *cmr-smslareunion* corpus' OLAC metadata card. It illustrates the encoding roles (JOHNSON, 2006). These roles can also easily be encoded, as metadata, in the TEI header.

```
<dc:creator>LEDEGEN Gudrun</dc:creator>
<dc:creator>CHANIER Thierry </dc:creator>
<dc:contributor xsi:type="olac:role" olac:code="compiler">LEDEGEN Gudrun</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="editor">CHANIER Thierry</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="depositor">CHANIER Thierry</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="data_inputter">JIN Kun</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="data_inputter">HRIBA Linda</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="developer">LOTIN Paul</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="participant"/>
<dc:contributor xsi:type="olac:role" olac:code="sponsor"> [...]
```

Figure 11: Examples of OLAC encoding roles

Whilst acknowledging different participants' contributions to a corpus is one issue, referring to a corpus as a global entity and to its creator is another. A specific way of referencing corpora must be adopted when citing and referencing the work; much the same way as bibliographic references are constructed and used within scientific publications. Bearing in mind the CODATA/ITSCI (2013) recommendations, CoMeRe decided to encode bibliographic reference to corpora as shown in Figure 12.

```
<dcterms:bibliographicCitation>Ledegen, G.(2014). Grand corpus de sms SMSLa
Réunion [corpus]. In Chanier T. (ed.) Banque de corpus CoMeRe. Ortolang.fr :
Nancy. [cmr-smsalpes-tei-v1 ; http://handle.net/xxx/cmr-smslareunion-tei-v1]
</dcterms:bibliographicCitation>
```

Figure 12: Corpus citation

In the "Dublin Core - OLAC" metadata set, the bibliographic reference is integrated into the tag <BibliographicCitation>. The contents of this element will be displayed on the Internet interface developed by ORTOLANG for corpus consultation and access. Following the CoMeRe example of how to form a bibliographic reference for a corpus, ORTOLANG have

taken the decision to ask every corpus depositor to elicit this reference. This is a step in the right direction where standardized citation procedures are concerned.

Within a corpus citation, the permalink is an essential part of the reference[8]. In the same way that a Digital Object Identifier (DOI) allows a user to obtain direct access to the abstract of a scientific publication, the permalink will be a permanent link to the corpus metadata. The latter both allows users to search the ORTOLANG corpus access interface but will also be compliant with harvesting protocols including the *Open Archives Initiative Protocol for Metadata Harvesting (*OAI-PMH). The advantage of this is that every corpus will be easily searchable on the Web. Moreover, each CoMeRe corpus will have an OLAC form (also converted inside the corpus' TEI header), allowing automatic harvesting by European servers since ORTOLANG is a representative of CLARIN.

5.Conclusion and perspectives

The present article presented a general overview of the ongoing French CoMeRe project. Our ultimate goal is to build a kernel corpus of different CMC genres that is structured in TEI. At the time of writing, the CoMeRe repository comprises eight corpora, (out of which four served as examples in this paper) representing different CMC genres: text chats (more than 3 million), SMS (44,000), emails (2,300), forum messages (2,700), and Tweets (34,000).

Standardization is one of the key principles of the project and all CoMeRe corpora will be TEI-compliant. With this in mind, the CoMeRe project is involved in the European TEI-CMC SIG to design and write TEI guidelines for the markup of CMC data. The four corpora were marked-up in TEI under a format that is now part of the draft proposal of the TEI-CMC SIG. As explained above, we found it more adequate to first design a more general framework, termed "Interaction Space", that would fit the richest and the more complex CMC genres and situations. In doing so, the model developed encompasses multimodality. This is particularly important as new data will soon be added to the repository including, for example, MULCE corpora which comprise data coming from audio-graphic conferencing systems. Each CMC genre was then described through its interaction space and the TEI markup was determined regarding the IS.

Several of our TEI-compliant corpora are currently being tagged. The *Automatic Processing* WG has presented its motivations for applying an automatic annotation process to the CoMeRe corpora before turning to the decisions made concerning which annotations to make for which units and to a description of the processing pipeline for adding these to the CoMeRe data.

CoMeRe's automatic annotation process raises several issues, which are especially important where noisier corpora are concerned (SMS, text chat, etc.). Ongoing work[9] aims to better understand the phenomena that cause such data to depart from standard language corpora, in order to improve their automatic processing. As a first step, the *Automatic Processing* WG will focus on improving its tokenization and normalization scheme. This will require an explicit definition of the scope of the normalization process and a definition of the notion of *noisy token*.

The "genericity" of CoMeRe's POS and lemma annotation is a baseline that makes sense only if it can serve as input for various transformations, in order to be used in various types of linguistic and NLP uses of the CoMeRe corpora. Further work is now required to study the

balance between our annotations and requirements of the various uses of CoMeRe corpora. This might lead the WG to develop tools for converting annotations from its generic (FTB-UC) tagset - widespread in the French NLP community - into various other tagsets, more adequate for downstream uses.

Finally, the ideas, methods and tools described above have been designed and deployed on a few types of CMC corpora in two languages (French, English), including for the development of the *French Social Media Bank* (SEDDAH et al, 2012a), which will soon become part of CoMeRe.[10] Including new types of French CMC corpora within CoMeRe may require improvements and modifications of the approach and pipeline of the group, and even new strategies and tools. We are aware that a small set of gold standard annotations have to be produced and a formal evaluation of the tagging process be conducted. This may not be possible before the end of 2014; the concluding date of the first phase of the CoMeRe project.

Additional corpora are currently being processed by the *New Collections* WG. The **Twitter team** has developed a corpus of political tweets, *cmr-polititweets,* which reflects new political genres (LONGHI, 2013:31) in the framework of a more general research project on lexicon. The corpus aims to gather the most influential French political statements. More than 34000 tweets coming from 206 accounts have been collected and organized in our TEI format. The **Wikipedia team** is focusing on controversial talk pages in the fields of sciences and technologies. The corpus of talk pages, *cmr-wikiconflits*, will ultimately reflect different oppositions, such as controversial vs consensual, people vs objects. The team endeavors to examine four types of talk pages: (i) pages signaled on the Wikipedia mediation page; (ii) pages listed in the category *Neutral point of view: dispute*,[11] (iii) talk pages of articles having a pertinence controversy; and (iv) protected and semi-protected pages, *i.e.* pages subject to individual restrictions, temporarily or permanently limiting their editing. Data have already been collected and their transformation into our TEI format is in its final stages. Let us add that the Wikipedia team plan to conduct two types of analysis on the data and will concentrate both on the linguistic characteristics and the structure of the discussion pages.

These corpora, besides a selection of MULCE multimodal ones, will increase the representativeness and the variety of the CoMeRe repository, which will be released by the end of 2014. It will be the first milestone in the forthcoming French National Reference Corpus and we assume that the efforts we undertook will meet the strong demand for open and standard data within our community.

6.References

ABEILLÉ, A., CLÉMENT, L. & TOUSSENEL, F. (2003). *Building a Treebank for French*. Kluwer: Dordrecht.

ANTONIADIS, G (in print). „Corpus de SMS réels dans les Alpes, smsalpes" [corpus]. In Chanier T. (ed.) *Banque de corpus CoMeRe*. Ortolang.fr : Nancy. [cmr-smsalpes-tei-v1; *http://www.ortolang.fr/*]

ANTONIADIS, G., CHABERT, G. & ZAMPA V. (2011). „Alpes4science : Constitution d'un corpus de SMS réels en France métropolitaine". *TEXTOS conference: dimensions culturelles, linguistiques et pragmatiques*. Annual conference of ACFAS, 9-10 May 2011, Sherbrooke, Canada.

ASTON G. & BURNARD L. (1998). *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

BALDRY, A. & THIBAULT, P-J. (2006). *Multimodal Transcription and Text Analysis*. Equinox: London.

BEIßWENGER, M., ERMAKOVA, M., GEYKEN, A., LEMNITZER, L. & STORRER, A. (2012). „A TEI Schema for the Representation of Computer-mediated Communication". In *Journal of the Text Encoding Initiative* (jTEI), 3, *http://jtei.revues.org/476* ; DOI : 10.4000/jtei.476.

BEIßWENGER, M., CHANIER, T. ,CHIARI, I., ERMAKOVA, M., VAN GOMPEL, M.,HENDRICKX, I, HEROLD, A., VAN DEN HEUVEL, H,LEMNITZER, L. & STORRER, A. (2013). „Computer-Mediated Communication in TEI: What Lies Ahead". Special Topic Panel, *TEI Conference and Members Meeting 2013,* 2-5 Octobre 2013, Rome, Italy.

BELLIK Y. & TEIL D. (1992). „Définitions terminologiques pour la communication multimodale". *Conference Interaction Humain-Machine IHM'92*, Paris. http://perso.limsi.fr/bellik/publications/1992_IHM_1.pdf

BURNARD, L. & BAUMAN, S. (2013). *TEI P5: Guidelines for electronic text encoding and interchange* [Document] . TEI consortium, tei-c.org. *http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf*

CALICO (2013). *Online Tools for analysing and visualising discussion forums* [website] *http://www.stef.ens-cachan.fr/calico/outils/outils.htm*

CHABERT, G., ZAMPA, V., ANTONIADIS, G. & MALLEN, M. (2012). *Des SMS Alpins*. Éditions de la Bibliothèque départementale des Hautes-Alpes: Gap.

CHANIER, T. & VETTER, A. (2006). „Multimodalité et expression en langue étrangère dans une plate-forme audio-synchrone". *Apprentissage des Langues et Systèmes d'Information et de Communication* (ALSIC), 9. DOI: 10.4000/alsic.270, *http://alsic.revues.org/270*

CODATA/ITSCI Task Force on Data Citation (2013). „Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation". *Data Science Journal* 12, pp 1-75, DOI: 10.2481/dsj.OSOM13-043

CoMeRe (2014). *Communication Médiée par les Réseaux*, project documentation [website], *http://comere.org*

CoMeRe Repository (2014). Repository fo the CoMeRe corpora [website], *http://hdl.handle.net/11403/comere*

COOK, P. & STEVENSON, S. (2009). „An Unsupervised Model for Text Message Normalization". In Feldman, A. & Lönneker-Rodman, B. (Ed.). *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pp. 71–78. *http://aclweb.org/anthology/W/W09/W09-2000.pdf*

Corpus-écrits (2013). *Consortium Corpus-écrits* [website]. *http://corpusecrits.corpus-ir.fr*

DARIAH (2013). *Digital Research Infrastructure for Arts and Humanities* [website]. *http://www.dariah.eu/*

DENIS, P. & SAGOT, B. (2012). „Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging". In *Language Resources and Evaluation*, 46(4), pp. 721–736.

DOCTISSIMO (2013). *Discussion forum linked to the website Doctissimo, general public welfare and health care* [webservice]. Lagardère Active : doctissimo.fr. *http://forum.doctissimo.fr/*

DWDS (2013). *Das Digitale Wörterbuch der deutschen Sprache* [website] *http://www.dwds.de/*

FAIRON, C. & PAUMIER, S. (2006). „A translated corpus of 30,000 French SMS". In *Proceedings of LREC 2006*, 22-28 May 2006, Genova, Italy. *http://www.lrec-conf.org/proceedings/lrec2006/*

FALAISE, A. (in print). „Corpus de français tchaté getalp_org" [corpus] . In Chanier T. (ed) Banque de corpus CoMeRe *Banque de corpus CoMeRe*. Ortolang.fr : Nancy. [cmr-getalp_org-tei-v1 ; *http://www.ortolang.fr*]

FALAISE, A. (2005). „Constitution d'un corpus de français tchaté". In *Actes de RECITAL 2005*, 6-10 June, Dourdan, France. *http://hal.archives-ouvertes.fr/hal-00909667*

GEYKEN, A. (2007). „The DWDS-Corpus: A reference corpus for the German language of the 20th century". In C. Fellbaum (Ed.). *Collocations and idioms: linguistic, lexicographic, and computational aspects.* London: Continuum Press.

HUMA-NUM (2013*). French Infrastructure for Digital Humanities* [website]. *http://www.huma-num.fr/*

IRCOM (2013). *Consortium Corpus Oraux et Multimodaux* [website]. *http://ircom.corpus-ir.fr*

KUPIETZ, M. & H. KEIBEL (2009): „The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research".*Working Papers in Corpus-based Linguistics and Language Education*, No. 3, pp. 53–59. Tokyo: Tokyo University of Foreign Studies (TUFS).

LAMY, M-N. & HAMPEL, R. (2007). *Online Communication in Language Learning and Teaching*. Basingstoke: Palgrave Macmillan.

LaRéunion4Science (2008). *Site of the project sms4science located in La Réunion* [website] *http://www.lareunion4science.org/*

LEDEGEN, G. (in print). „Grand corpus de sms SMSLa Réunion" [corpus]. In Chanier T. (ed.) *Banque de corpus CoMeRe*. Ortolang.fr : Nancy. [cmr-smsalpes-tei-v1; http://www.ortolang.fr/]

LEDEGEN, G. (2010). „Contact de langues à La Réunion : « On ne débouche pas des cadeaux. Ben i fé qoué alors ? »". *Langues et Cité, 'Langues en contact'*, vol.16, pp. 9-10. *http://www.dglf.culture.gouv.fr/Langues_et_cite/LC16.pdf*

LONGHI, J. (2013). „Essai de caractérisation du tweet politique". *L'Information Grammaticale*. vol.136, pp.25-32.

MULCE repository (2013). *Repository of learning and teaching (LETEC) corpora* [webservice] . Clermont Université : MULCE.org. *http://repository.mulce.org*

OOSTDIJK, N., REYNAERT, M., MONACHESI, P., VAN NOORD, G., ORDELMAN, R., SCHUURMAN I., & VANDEGHINSTEV. (2008). „From D-Coi to SoNaR: A reference corpus for Dutch". In *Proceedings of LREC*, 28-30 May, Marrakech, Morocco. *http://www.lrec-conf.org/proceedings/lrec2008/index.html*

JOHNSON, H. (2006). *OLAC Role Vocabulary* [document]. Open Language Archive Community (OLAC). *http://www.language-archives.org/REC/role.html*

OpenData (2013) *Principles for "openness" in relation to data and content* [document]. Open Kowledge Foundation : http://okfn.org/. *http://opendefinition.org/od/*

ORTOLANG (2013). *Open Resources and TOols for LANGuage* [website]. ATILF / CNRS - Université de Lorraine : Nancy, *http://www.ortolang.fr*

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 sciencessociales appliquées, 9: Des usages numériques aux pratiques scripturales électroniques, 107-138. *http://hal.archives-ouvertes.fr/hal-00923618*

REFFAY, C. CHANIER, T. LAMY, M.-N. & BETBEDER, M.-L. (2009). (editors). *LETEC corpus Simuligne* [corpus]. MULCE.org : Clermont Université. [oai:mulce.org:mce.simu.all.all ; *http://repository.mulce.org*]

REFFAY, C. & CHANIER, T. (2003). „How social network analysis can help to measure cohesion in collaborative distance-learning". In *Proceedings of Computer Supported Collaborative Learning Conference (CSCL'2003)*. June 2003, Bergen, Norway. Kluwer Academic Publishers : Dordrecht, pp. 343-352. *http://edutice.archives-ouvertes.fr/edutice-00000422*

REHM, G. ET AL. (2008). „Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems". In *Proceedings of LREC*, 28-30 May, Marrakech, Morocco. *http://www.lrec-conf.org/proceedings/lrec2008/index.html*

REYNAERT, M., OOSTDIJK, N., DE CLERCQ, O., VAN DEN HEUVEL, H., & DE JONG, F. (2010). „Balancing SoNaR: IPR versus Processing Issues in a 500-million-Word Written Dutch Reference Corpus". In, *Seventh conference on International Language Resources and Evaluation*, LREC '10, 19-21 May 2010, Malta. *http://doc.utwente.nl/72111/1/LREC2010_549_Paper_SoNaR.pdf*

SAGOT, B. & BOULLIER, P. (2008). „SxPipe 2 : architecture pour le traitement présyntaxique de corpus bruts". *Traitement Automatique des Langues*, 49(2), pp. 1-35.

SAGOT, B. (2010). „The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French". In Calzolari, N., et al. (Ed.). *Proceedings of LREC'10*, 17-23 May, Valetta, Malta. *http://lrec-conf.org/proceedings/lrec2010/index.html*

SEDDAH, D., SAGOT, B., CANDITO, M., MOUILLERON, V. & COMBET, V. (2012a). „The French Social Media Bank: a Treebank of Noisy User Generated Content". In Kay, M. & Boitet, C. (Ed.). *Proceedings of CoLing 2012: Technical Papers*, 8-15 Decembre 2012, Mumbai, India, pp. 2441-2458. *http://aclweb.org/anthology/C/C12*

SEDDAH, D., SAGOT, B. & CANDITO, M. (2012b) „The Alpage Architecture at the SANCL 2012 Shared Task: Robust Pre-Processing and Lexical Bridging for User-Generated Content Parsing". In *Notes of the first workshop of Syntactic Analysis of Non Canonical Languages* (SANCL'2012), in conjunction with NAACL'2012, 3-8 June 2012, Montreal, Canada.

SHAROFF, S.(2006). „Methods and tools for development of the Russian Reference Corpus". In Wilson, A., Archer, D. & Rayson, P. (Ed.). *Language and Computers, Corpus Linguistics Around the World*. Rodopi: Amsterdam, pp.167-180. *http://npu.edu.ua/!e-book/book/djvu/A/iif_kgpm_Corpus%20Linguistics.pdf*

TEI-CMC (2013). *TEI Special Interest Group on Computer-Mediated Communication* [website] *http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication*

WIGHAM, C.R. & CHANIER, T. (2013). „Interactions between text chat and audio modalities for L2 communication and feedback in the synthetic world Second Life". *Computer Assisted Language Learning*, DOI: 10.1080/09588221.2013.851702.

[1] CoMeRe stands for "Communication Médiée par les Réseaux" an updated equivalent to Computer-Mediated Communication (CMC) or Network-mediated Communication.

[2] The TEI term „channel“, which here corresponds to the environment, should not be confused with channels of the Internet Relay Chat (IRC) environment, where every channel correspond to a particular location oft he IS.

[3] In this article, we only discuss text (taken in the HALLIDAY (ibid) sense) macrostructure: IS structure, message / <post> structure (its title, elements which include its contents, relationships with other messages, addressees, etc). The micro-structure of the text refers to the type of elements found in the actual contents of the message / <post>, for example interaction words, emoticons, hash code, etc. See (BEISWENGER et al., 2012) for linguistic consideration on the micro-structure.

[4] This simple description of the structure of a forum (also used in analysis tools of forums based on XML structures such as (CALICO, 2013) is a sufficient one. Describing the structure of the modality forum should not be confused with the visual description of a forum a participant can adjust when using it: threads of discussions visualized as a sequence of indented messages, or as messages ordered accordingly the date of posting, etc. Our structure include all the information required for every specific visual display.

[5] During the whole process, XML annotations in the corpus are protected and ignored (but preserved).

[6] These rules were forged as follows: first, we extracted from various development corpora (the development part of the French Social Media Bank, parts of the CoMeRe data) *n*-gram sequences involving unknown tokens or occurring at an unexpectedly high frequency; then we manually selected the relevant ones and provided them manually with a corresponding “correction.”

[7] This tagged and lemmatized example is given in the MElt format, an extension of the Brown Corpus format, in which the “word”, its POS tag and its lemma are separated by slashes. A whitespace is a word-separator, and each sentence (i.e., each unit of treatment) is in one line. The tagset used here is the tagset used in the *French Social Media Bank*, which extends the so-called FTB-UC tagset (see SEDDAH *et al.*, 2012a and references therein); CLS is the POS tag for subject clitics, V for finite non-imperative verbs and ADV for adverbs, including for negative adverbs such as *pas* and (maybe surprisingly) for the negative clitic *ne*.

[8] Note that in Figure 12, the corresponding URL of the Handle type will be obtained when the corpus is deposited.

[9] Among other, these issues are the main topic of a PhD funded by the Région Rhône-Alpes about the study and exploitation of SMS French

[10] The other use case is the 2012 SANCL shared task organized by Google on “non-canonical” English parsing, a task based on the English Google WebBank (see SEDDAH *et al.*, 2012b and references therein).

[11] Signaling articles that for which the neutral point of view is controversial, *i.e.* articles deemed to be non-neutral. This is one of the major subjects of dispute on Wikipedia