Structure TEI-v2 après étiquetage morpho-syntaxique au sein du projet CoMeRe

Version 1

Version 11 mars 2014, Thierry Chanier

Comment citer ce document:

Chanier, T (2014). *Structure TEI-v2 après étiquetage morpho-syntaxique au sein du projet CoMeRe*. Projet CoMeRe (Communication Médiée par les Réseaux), IR Corpus-écrits. [http://comere.org/ , comere-traitement-v1]

Participants : Thierry Chanier, Benoit Sagot, Georges Antoniadis

Contenu

- **Objectif de ce document**

Ce document est destiné à définir la façon dont les corpus CoMeRe, initialement structurés en TEI-CMC (version `tei-v1`), doivent être structurés en format TEI-CMC en sortie de l'analyseur morpho-syntaxique MElt (Denis & Sagot, 2012). Les corpus seront alors en version `tei-v2`.

Il s'agit, pour l'instant d'un document de travail interne au projet CoMeRe permettant aux chercheurs responsables des traitements et à ceux responsables de la structuration en TEI de se mettre d'accord sur les intrants et extrants de la tâche traitements du projet CoMeRe. Chaque personne concernée peut donc librement modifier ce texte à condition d'indiquer en

première page la date et les auteurs de la nouvelle version, puis dans le corpus du texte à rendre les modifications du texte visibles (révision Word, ou coloriage du texte modifiée, ou …).

Ce document présuppose la connaissance du contenu du rapport CoMeRe (Chanier & Jin, 2013).

1. Flux de traitement de la tei-V2

1.1. Groupes de travail

Composition actuelle des groupes de travail :

- **CoMeRe-LRL** comprend : Paul Lotin, Kun Jin et Thierry Chanier.
- **CoMeRe-TEI** comprend : Kun Jin, Linda Hriba, Thierry Chanier
- **CoMeRe-qualité** : Julien Longhi, Gorgeta Cislaru, Ciara Wigham, Gudrun Ledungen
- **CoMeRe-traitements** : Georges Antoniadis, Benoit Sagot
- **CoMeRe-nouvelles-acquisitions** : Céline Poudat, Julien Longhi, puis en plus pour Wikipedia : Natalia Graba, Camille Paloque-Berges et pour les tweets : Claudia Marinica (contributrice).

1.2. Répertoires AJAX

Dans le `serveur http://msh-handle.univ-bpclermont.fr/comere > Mes Fichiers`

- `traitements-comere > tei-v1` : répertoire destiné au dépôt par le groupe Comere-LRL des corpus à traiter, par exemple `sms-lareunion-tei-v1.xml`. Des sous-répertoires peuvent apparaitre (par exemple pour `cmr-getalp_org` avec 80 fichiers XML)
- `traitements-comere > tei-v2` : répertoire destiné au groupe CoMeRe-traitements pour y déposer les fichiers résultats du traitement.

1.3. Modifications entre intrant et extrant

1.3.1. Intrant

Prenons l'exemple du corpus `sms-lareunion`. Etant donné les problèmes de coordination entre les agendas des différents groupes, il est possible que le dépôt du `tei-v1` par CoMeRe-LRL dans `traitements-comere > tei-v1` ne corresponde pas tout à fait à la version définitive de cette version `tei-v1`. Les parties `<teiheader>` sont en effet en cours de vérification par le groupe qualité. Mais peu importe, car les parties `<text>` ou `<body>` ne changeront pas, et c'est sur ces parties là que porteront les travaux d'étiquetage.

1.3.2. Extrant

Le groupe CoMeRe-traitements déposera, comme indiqué, le résultat de son travail dans `traitements-comere > tei-v2.` Il ne changera pas la partie `<teiheader>`[1], mais la recopiera en l'état (de façon à ce que le fichier soit valide au regard de notre schéma RELAX (fichier `tei_cmr.rng`). **Attention, le groupe CoMeRe-traitements vérifiera à l'issue du traitement d'étiquetage que le nouveau fichier est valide**, ici `sms-lareunion-tei-v2.xml` (vérif à faire

[1] Pour ce qui concerne la description des étiquettes POS (voir (2.6)), soit ce sera introduit dans chaque `<teiheader>` à l'emplacement indiqué, soit ce sera fourni dans le bon format XML au groupe CoMeRE-LRL qui l'introduira après coup.

dans un éditeur XML). Clermont ne pourra en effet pas réparer à la main des problèmes de syntaxe qui pourraient porter sur des milliers, voire des millions d'occurrences.

1.3.3. Après le travail du groupe traitements

CoMeRe-LRL récupérera alors dans ce dernier répertoire (`sms-lareunion-tei-v2.xml`) l'extrant puis, dans d'autres répertoires AJAX, il en modifiera le `<teiheader>` (voir annexes §7.1 pour l'organisation de AJAX). Ensuite la nouvelle version sera déposée par CoMeRe-LRL pour vérification dans l'espace du groupe CoMeRe-qualité. Plus tard encore, Linda Hriba réalisera une fiche OLAC pour la version `tei-v2` du corpus de départ.

2. Principes généraux d'encodage des analyses

2.1. Macro-structure et segmentation

2.1.1. Corps des énoncés, du texte

Les parties à analyser seront contenues dans les éléments `<post>` ou `<u>` (pour transcription audio). Un <post> contient un ou plusieurs paragraphes `<p>`. L'analyse conservera cette organisation. Donc les segments analysés seront tous remis dans le paragraphe de départ.

Lorsque l'analyseur traitera un paragraphe, il pourra y détecter un ou plusieurs segments correspondant à des "phrases" (terminologie douteuse compte tenu du type de modalité de communication que nous avons à traiter, mais on s'en contentera). Donc un `<p>` sera segmenté en un ou plusieurs éléments `<s>` (correspondant à *sentence*).

Lorsque l'analyseur traitera un énoncé <u>, s'il détecte plusieurs sous-parties correspondant à des phrases, alors il étiquettera chaque sous-partie avec <s> aussi. Si aucune sous-partie n'est détectée, alors la segmentation en *tokens* / graphies sera laissée dans <u>, sans utiliser de <s>.

- **Cas du <lb/>**

Cet élément apparait dans certains corps de messages (courriels, forums ou blogues). Lors du traitement des messages (passage en XML), un certain nombre de mises en forme ont été supprimées. D'autres ont pu être préservées (voir 2.3). Une que nous avons tenue à sauvegarder correspondant à des retours à la ligne successifs. Nous avons alors traduit ce ou ces retours par l'élément `<lb/>`. Cet élément peut être un indice de segmentation. Il peut être redondant avec un signe de ponctuation (un point, par exemple), ou le remplacer. Il peut indiquer une fin de phrase, voire de paragraphe (mais impossible de trancher dans ce dernier cas). Mais il peut aussi correspondre à une simple mise en emphase d'un morceau de phrase.

❖ Question : cet élément <lb/> peut-il ou non servir d'indicateur lors de la segmentation automatique ? Il devra de toute façon être conservé dans la sortie du traitement.

2.1.2. Titre et autres informations de niveau méta

Dans de nombreuses modalités le message ou énoncé est accompagné d'un titre (message de forum, de courriel, de blogue, etc.), voire d'étiquettes (message de blogue). Ces éléments (`<title>, <label>`) devront être traités. Le résultat de l'analyse sera mis dans un élément `<phr>` (pour *phrase*).

Ces éléments sont contenus dans la partie `<head>` du `<post>`. (2.1) donne un exemple de message / billet de blogue et du commentaire associé à ce message. Le billet a un titre et une étiquette / catégorie et le message également un titre. Ces éléments sont facultatifs.

```
(2.1)
<post xml:id="cmr-blog-a2" synch="#T2" who="#P2" type="blog-message">
     <head>
      <title>Présentation de ma personne</title>
     <label>étapeE1 ; </label>
    </head>
<p>Bon soir à tous!<lb/> Maintenant, je vais commencer avec les
présentations..... <lb/> Je pense que vous avez vu que je m'appelle <name
ref="#P2">Kerstin</name> . J'ai 22 ans. Mon nom est un nom suédois qui est très
fréquent en Allemagne. Comme  vous savez peut-être, on a commencé nos études en
master cette semaine. <lb/> Ma famille - mes parents et mes deux soeurs -habite
à Osnabrueck. C'est une ville qui
  […] <lb/> A bientôt, <name ref="#P2">Kerstin</name></p>
 </post>

<post xml:id="cmr-blog-a3" synch="#T3" who="#P3" type="blog-comment" ref="#cmr-
blog-a2">
    <head>
      <title>Hallo Kirstin! J'ai lu que tu as fait des stages e...</title>
    </head>
<p>Hallo<name ref="#P2">Kirstin</name> ! J'ai lu que tu as fait des stages en
Suisse francophone ! Où exactement car j'habite près de la frontière suisse (à
1h de Lausanne !)! Je pense qu'on aura l'occasion d'en reparler ! Bis Bald </p>
</post>
```

Donc l'analyse de (2.1) donnera (2.2).

```
(2.2)
<post xml:id="cmr-blog-a2" synch="#T2" who="#P2" type="blog-message">
   <head>
     <title><phr>analyse du titre au niveau micro</phr></title>
     <label><phr> analyse étiquette au niveau micro</phr></label>
   </head>
<p> analyse du paragraphe au niveau micro</p>
 </post>

<post xml:id="cmr-blog-a3" synch="#T3" who="#P3" type="blog-comment" ref="#cmr-
blog-a2">
    <head>
      <title><phr>analyse du titre au niveau micro</phr></</title>
    </head>
<p> analyse du paragraphe au niveau micro</p>
</post>
```

Dans le message de courriel (2.3), on voit apparaître au sein d'un paragraphe <p>, des balises de mises en forme du message (`<hi>`), ainsi qu'un tableau. Ce qui a été mis en emphase peut parfaitement constituer avec les mots suivants une phrase complète.

- ❖ Question : Allons-nous pour autant essayer de reconstituer cette phrase, ou simplement étiqueter chaque partie avec <phr> ? Rappelons que les balises TEI doivent être conservées (donc ici le <hi>)

Pour les contenus des cellules, dans cet exemple ils ne correspondent pas à des phrases complètes et devraient donc être balisés avec <phr>. S'il s'agit d'un ou plusieurs phrases, MElt arrivera –t-il à le repérer et à baliser avec un ou plusieurs <s> ?

(2.3)

```
     <post xml:id="cmr-email-a1" synch="#T1" who="#P1" type="email-message">
        <head>
          <title>Evaluation de synthese</title>
          <listPerson>
[…]          </listPerson>
        </head>
        <p>
          <lb/>
          <hi style="#000088 ; +1">Une question avait été posée dans le forum
Interculture.</hi>
          <lb/>
          <hi style="#000088 ; +1">Cette question était :</hi>
          <lb/>Si vous aviez à évaluer les messages postés dans Interculture. Quels
seraient vos différents critères et leur importance relative ? Un bon message dans
Interculture est un  message qui...<lb/>
          <hi style="#000088 ; +1">La synthèse proposée par <name ref="#P2"
type="person"><surname><fs type="anonymisation">
                  <f name="origfrom"><string>Depositor</string></f>
                  <f name="anonyString"><string>[_surname_]</string></f>
               </fs></surname></name> est :</hi>
          <lb/>Selon l'avis général (peu éloquant), on peut dire qu'un bon message
dans InterCulture est d'abord un message qui permet de faire réagir l'autre, mais
aussi d'ajouter des éléments au sujet, ou de converger vers une synthèse au moins
partielle du sujet. Je pense de surcroît que le ton employé et la dose d'humour ou
de tact sont d'une grande importance  pour amener les autres à réagir sans jamais
dévaloriser leurs contributions. Un bon message doit aussi être inséré au bon
moment, au bon endroit et avec les mots à la fois les plus simples et les plus
justes pour permettre à l'autre de les saisir sans ambiguïté.<lb/>
          <hi style="#000088 ; +1">Que pensez-vous de cette synthèse ?</hi>
          <lb/>
          <table>
            <row>
              <cell> répond parfaitement à la question</cell>
              <cell> clair, précis, expression correcte</cell>
            </row>
            <row>
              <cell> répond partiellement à la question</cell>
              <cell> confus ou imprécis</cell>
            </row>
            <row>
              <cell> ne répond pas correctement à la question</cell>
              <cell> nombreuses fautes d'orthographe</cell>
            </row>
            <row>
              <cell> ne répond pas du tout à la question</cell>
              <cell> très mal exprimé</cell>
            </row>
          </table>
          <lb/>
          <hi style="#000088 ; +1">Précisions éventuelles :</hi>
          <lb/>
        </p>
        <trailer>
          <name ref="#F1"/>
        </trailer>
      </post>
```

2.2. Micro-structure

2.2.1. Exemple

Partons de celui figurant dans (Chanier & al., 2014), à savoir (2.4)

```
(2.4)
sa fé o moin 6 mois qe les preliminaires sont sauté c a dire qil yen a presk pa
```

Une première sortie pourrait être telle qu'en (2.5) et, par ailleurs, dans le <teiheader> devra figurer un listing exhaustif des étiquettes sous un format tel que (2.6). `<interGrp>` sera intégré dans `<teiheader>/<encodingDesc>/<editorialDecl>/<interpretation>/<p>.` Il parait raisonnable de ne pas chercher à traduire les informations provenant de la phase "Normalization" du traitement. En effet cela entraînerait une complexification de la structure TEI (avec des choix entre version de départ et version "corrigée"), certains regroupements (cf. c'est-à-dire). L'interprétation ne peut être garantie comme correcte (cf. propos des auteurs). Il vaut mieux laisser cette normalisation pour une étape ultérieure de traitement, hors projet CoMeRe.

```
(2.5)
<w ana="#PRO" lemma="ça">sa</w> <w="#V" lemma="faire">fé</w> <w ana=#P+D"
lemma="à+le">o</w> [...]<w ana="#Y">c</w> <w ana"#Y">a</w> <w ana="#CC"
lemma="c’est-à-dire">dire</w> […] <w ana="ADV" lemma="pas">pa</w>
```

```
(2.6)
<interpGrp type="POS">
<desc>etiquette provenant de l'analyseur MELt</desc>
 <interp xml:id="PRO">Pronom personnel</interp>
 <interp xml:id="V">Verbe</interp>
 <interp xml:id="P">préposition</interp>
 <interp xml:id="D">Déterminant</interp>
 <interp xml:id="P+D">preposition + déterminant</interp>
[…]
<bibl xml:id="MELt-pos"><!—ref à un document de MELt --></bibl>
</interpGrp>
```

3. Les graphies / token de l'Internet et de la communication médiée par les réseaux

3.1. Internet en général

Le tableau 4.1 de la section sur l'anonymisation, indique comment baliser les numéros de téléphone, adresse de courriels ou URL. Ainsi dans l'extrait de tweet (3.1), l'adresse apparaîtra comme dans (3.2)

```
(3.1)
un moment historique : Lyon est désormais relié à Oullins par le métro
http://t.co/PDBCSVh4vU #Lyon #TCL #MétroB #Sytral #Oullins
```

```
(3.2)
<rs type="url"><w ana="#??> http://t.co/PDBCSVh4vU </w></rs>
```

3.2. Les graphies propres à la CoMeRe

Nous avons pour obligations de suivre ce que nos collègues allemands du groupe TEI-CMC, auquel nous participons, ont déjà spécifié, notamment dans (Beißwenger et al, 2012). Dans cet article, ils indiquent que ces graphies ne s'inscrivent pas dans la syntaxe d'une phrase, mais ont une portée linguistique sur la/les phrases adjacentes (ou sur le tour de parole dans les clavardage, par exemple).

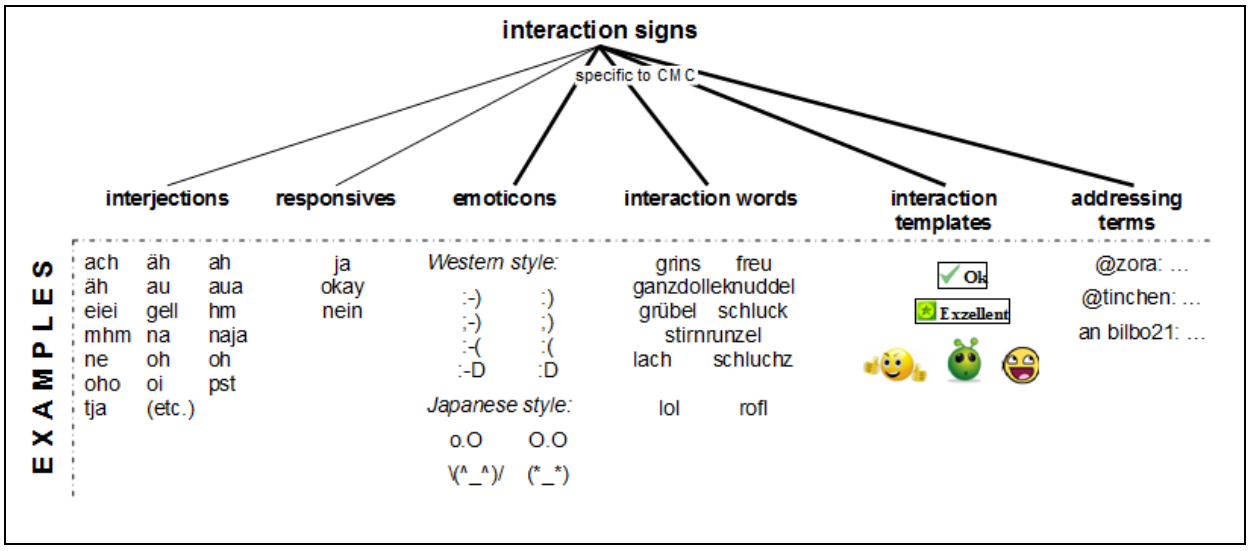

Figure 3.1 : signes d'interaction suivant (Beißwenger et al., 2012).

La figure 3.2 indique que les annotations, faites par ces collègues du groupe TEI-CMC, sont en partie manuelles.

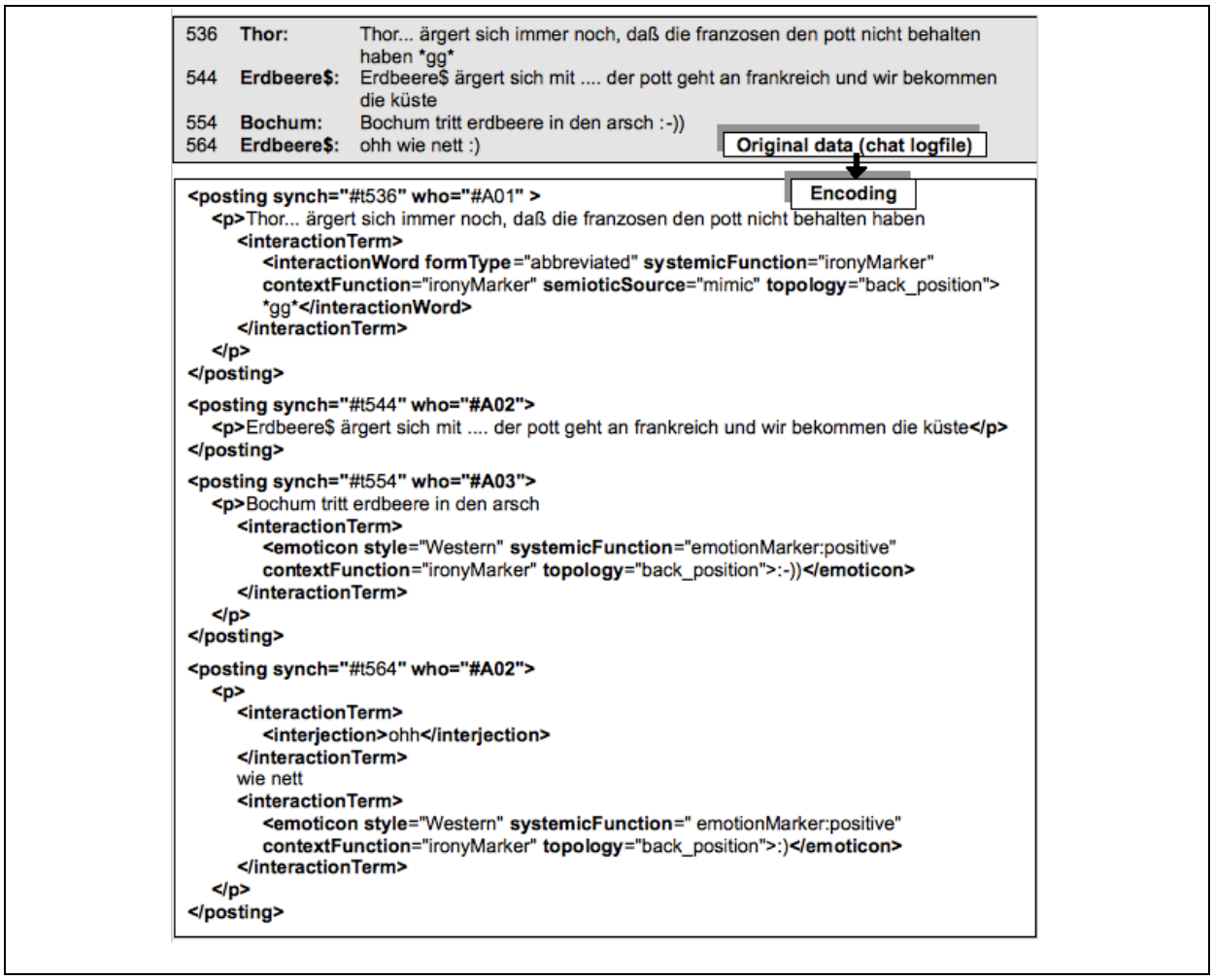

Figure 3.2 : encodage en TEI de signes d'interactions (Beißwenger et al., 2012).

Durant notre traitement automatique, nous simplifierons les choses (en permettant lors d'autres annotations post-CoMeRe d'ajouter les attributs nécessaires, voir les éléments de niveau supérieur).

Dans (3.3) extrait du corpus de clavardage informel `cmr-getalp_org-tei-v1`[2], on peut trouver un émoticône / binette `:)` et une graphie pour l'interaction (*interaction word*) `mdrrrrrrrr` , dans une version modifiée. Compte tenu des fortes variations /variantes graphiques de ces signes, il n'est pas sûr que l'on arrive à les repérer systématiquement automatiquement. Si tel est le cas, alors l'analyseur pourra les baliser come en (3.4).

```
(3.3)
c vré c les kangourous ki saute banane de Wizou va mdrrrrrrrr bébé :)

(3.4)
[…] <interactionWord><w ana="#??">mdrrrrrrrr</w></interactionWord><w ana="#??"
lemma="??">bébé</w><emoticon><w ana="#??">:)</w></emoticon>
```

Le repérage automatique peut aussi être problématique sur les termes d'adresse (*addressing terms*). Ainsi en (3.5) l'auteur du tour de clavardage (non mentionné ici) s'adresse à Wizou,

[2] Surtout ne pas surgénéraliser à partir des exemples extraits du corpus de clavardage libre `cmr-getalp_org`. Les clavardages en contexte éducatif dans `cmr-favi` et ceux contenus dans `cmr-simuligne` n'ont rien à voir avec la syntaxe des textos.

qui est présent dans la salle de clavardage sans utiliser le signe `@`. Lorsque cette détection pourra avoir lieu, alors la balise qui devra être utilisée sera `<addressingTerm>`

```
 (3.5)
mé Wizou kestu fous ici toi ofait ?
```

3.3. Le cas spécial des termes d'adresse dans Twitter (et le hashtag).

Les termes d'adresse sont omniprésents dans les tweets, bien sûr, mais ne fonctionnent pas exactement de la même façon que dans les clavardages.

Dans la syntaxe Twitter, on ne s'adresse explicitement à une personne que si on cite le nom de son compte (son twitto) **en premier** dans le message. Ainsi dans les exemples (3.6) à (3.9), seul l'auteur (non mentionné ici) de (3.9) adresse son message à une personne explicite `@sonatatiyants`. Seule cette graphie devrait donc être balisée par `<addressingTerm>`. En (3.6) on mentionne un twitto sans s'adresser à son auteur explicitement. Donc `@claudebartolone` devrait être balisé tel que dans (3.10).

```
(3.6)
Déjeuner de travail avec @claudebartolone a l'Assemblee Nationale a l'ordre du
jour le rôle de la Caisse des dépôt a la SNCM

(3.7)
RT @AvecMennucci: L'équipe #Mennucci en marche! http://t.co/pBtgU1CrJN

(3.8)
RT @marcelvayre: Oh peuchere...L'illusion Gaudin via @Marianne2fr #Marseille
#Marseille2014 #Mennucci http://t.co/jYtSTW1gPn

(3.9)
@sonatatiyants Thanks so much for your kind words! Hope that you're having a
fantastic weekend! :)
```

En (3.7) et (3.8), apparaissent des réplications de tweet (Retweet) qu'il faudrait noter spécifiquement (voir (3.11) pour (3.7) où l'on englobe 3 graphies). En (3.8), en plus du retweet apparaît le phénomène de référence. Le terme via suivi d'un twitto indique que ce qui a été dit précédemment provient / réfère à l'auteur du compte Twitter (le twitto), ce que l'on traduira en (3.12). Enfin dans ce même message (3.8) apparaissent plusieurs mots-clés renvoyant à des fils de discussion dans Twitter et symbolisés par des *hashtags* (voir traduction en (3.12)).

```
(3.10)
<rs type="twitter-account"><w ana="#??>@claudebartolone</w></rs>

(3.11)
<rs type="twitter-retweet"><w ana="#??>RT @AvecMennucci:</w></rs> <w ana="#DET
lemma="la">l</w> <w ana"#N" lemma="équipe">équipe</w> […]

(3.12)
[…]  <w ana="#PN">Gaudin</w> <rs type="twitter-ref-account"><w ana="#??>via
@Marianne2fr</w></rs> <rs type="twitter-hashtag"><w
ana="#??">#Marseille</w></rs>
```

❖ On remarquera que dans ces exemples (3.10) à (3.12) on fait l'hypothèse que l'analyseur ne cherche pas à traiter le détail des graphies. MELt pourrait par contre décider de les traiter et trouver ainsi que `#Marseille` renvoie à une ville. A décider ? En tous cas, il est important que la balise `<rs>` englobe l'ensemble (où pourrait se trouver plusieurs `<w>` en cas d'analyse plus fine).

4. Cas particuliers

4.1. Anonymisation

Comme expliqué dans le rapport comere-is-tei-v2 (Chanier & Jin, 2013), le travail accompli pour le passage en TEI a servi à harmoniser les différentes approches sur l'anonymisation adoptées initialement par chaque dépositeur de corpus. Le format choisi (une structure de trait en TEI, élément `<f>`) a permis de conserver toutes les informations au prix d'une certaine lourdeur.

Nous reproduisons l'exemple (4.1) provenant d'un texto.

```
(4.1)
  <post xml:id="cmr-smsalpes-c001-a8" when-iso="2010-10-01"
          who="#cmr-smsalpes-c001-p288726353160825" type="sms">
          <p>Bon, d'accord ! Elle est ou miss <fs type="anonymisation">
              <f name="numOrder"><numeric value="522"/></f>
              <f name="numCharacter"><numeric value="5"/></f>
              <f name="origfrom"><string>Depositor</string></f>
              <f name="anonyString"><string>[_forename_]</string></f>
            </fs> ? Usa ? On se retrouve ds 1 coin sympa ? Qu'est-ce qui
t'arrange ?</p>
        </post>
```

Le message contenu dans le <p>peut se lire plus simplement comme en (4.2) en sachant qu'à la place de `[_forename_]` figurait initialement un prénom.

```
(4.2)
  Bon, d'accord ! Elle est ou miss [_forename_] ? Usa ? On se retrouve ds 1 coin
sympa ? Qu'est-ce qui t'arrange ?
```

En (4.3) sont listés tous les segments de remplacement.

```
(4.3)
                [_forename_]
                [_surname_]
                [_addName_]
                [_tel_]
                [_email_]
                [_url_]
                [_code_]
                [_address_]
```

Pour cette version `tei-V2`, nous proposons de **ne pas conserver en sortie** la structure de traits. Celle-ci donne des indications, notamment, sur qui a accompli ce processus d'anonymisation. En cas de besoin, on pourra toujours se reporter à la version `tei-v1` des corpus.

Cependant, lors de ce traitement, il faudrait conserver les balises TEI identifiant la nature du constituant, comme indiqué dans le tableau 4.1. .

Token provenant de tei-v1	éléments de TEI correspondant	POS
[_forename_]	<forename>[_forename_]</forename>	Celui des noms propres (PN ?)
[_surname_]	<surname>[_surname_]</surname>	Celui des noms propres
[_addName_]	<addName>[_addName_]</addName>	Celui des noms propres
[_tel_]	<rs type="telephone">[_tel_]</rs>	??
[_email_]	<email>[_email_]</email>	??
[_url_]	<rs type="url">[_url_]</rs>	??
[_code_]	<rs type="code">[_code_]</rs>	??
[_address_]	<address>[_address_]</address>	??

Tableau 4.1.: éléments de TEI en colonne droite

Prenons un autre exemple extrait de (2.3). Dans le tableau 4.2, les informations contenues dans le message initial (le fait qu'il s'agisse du nom d'une personne, l'identifiant de la personne, etc.) doivent être conservées pour analyses ultérieures, notamment sur les interactions.

entrée	Sortie
<hi style="#000088 ; +1">La synthèse proposée par <name ref="#P2" type="person"><surname><fs type="anonymisation"><f name="origfrom"><string>Depositor</string></f><f name="anonyString"><string>[_surname_]</string></f></fs></surname></name> est :</hi>	<hi style="#000088 ; +1">La synthèse proposée par <name ref="#P2" type="person"><surname><w ana="PN">[_surname_]</w></surname></name> […]

Tableau 4.2

(4.1) prendra donc la forme (4.4).

```
(4.4)
  <post xml:id="cmr-smsalpes-c001-a8" when-iso="2010-10-01"
          who="#cmr-smsalpes-c001-p288726353160825" type="sms">
          <p><w ana="#XX" lemma="XX">Bon</w> <w ana="#PONC">,</w> […] <w
ana="#XX" lemma="XX">miss</w> <forename><w ana="#PN">[_forename_]</w></forename>
[…]</p>
   </post>
```

4.2. Partie du discours dans une autre langue

Nous reproduisons en (4.5) une partie du message d'un blogue de `cmr-infral-tei-v1` affiché complètement en (2.1).

```
(4.5)
<p>Hallo<name ref="#P2">Kirstin</name> ! J'ai lu que tu as fait des stages en
Suisse francophone ! Où exactement car j'habite près de la frontière suisse (à
1h de Lausanne !)! Je pense qu'on aura l'occasion d'en reparler ! Bis Bald </p>
```

On remarquera la présence d'un certain nombre de mots ou expressions en allemand. Ce corpus assemble des échanges entre deux groupes d'étudiants franco-allemands. L'allemand y est très minoritaire, mais présent assez régulièrement.

Si l'analyseur détecte des mots d'une autre langue, nous proposons de l'indiquer dans l'attribut `@xml:lang`, comme en (4.6)

```
(4.6)
<p><w xml:lang="deu">Hallo</w><name ref="#P2"><w ana="#PN">Kirstin</w></name> <w
ana="#PONC">!</w> […] <w xml:lang="deu">Bis</w> <w xml:lang="deu">Bald</w> </p>
```

Lorsque tout le message est dans une autre langue que le français (ce qui est possible de façon minoritaire dans le corpus `cmr-simuligne-tei-v1`), alors le message pourrait ne pas être traité, mais la langue du paragraphe indiquée. Dans ce cas (4.7) deviendrait (4.8) (attention au titre en français).

```
(4.7)
        <post xml:id="cmr-forum-a2" synch="#T2" who="#P2"
            type="forum-message" ref="#cmr-forum-a1">
            <head>
               <title>Petit coucou de Besancon</title>
               <listPerson>
                  <person corresp="#P1">
                     <event type="Read" when="2001-05-25T09:18:00">
                        <label>Read</label>
                     </event>
                  </person>
                  <person corresp="#P2">
                     <event type="Read" when="2001-05-25T09:18:00">
                        <label>Read</label>
                     </event>
                  </person>
                  <person corresp="#P3">
                     <event type="Read" when="2001-05-25T09:18:00">
                        <label>Read</label>
                     </event>
                  </person>
               </listPerson>
            </head>
            <p>It was interesting to see that all the reactions to the word
               'school' were fairly neutral (including my own) - and also
               I'm surprised that nobody has contributed anything here yet.
               I LOVED school, had a great time, learned a lot, made many
               lasting friendships... I was sorry when the time came to
               leave the place. <name ref="#P3" type="person"
                  ><forename>Marja</forename></name>
            </p>
         </post>.
```

```
(4.8)
        <post xml:id="cmr-forum-a2" synch="#T2" who="#P2"
            type="forum-message" ref="#cmr-forum-a1">
            <head>
               <title><w ana="#??" lemma="??>Petit</w> <w ana="#??"
lemma="??>coucou</w> <w ana="#??" lemma="??>de</w> <w ana="#??"
lemma="??>Besancon</w></title>
               <listPerson>
[…] recopier comme tel            </head>
            <p xml:lang="eng">It was interesting to see that all the reactions
to the word 'school' were fairly neutral (including my own) - and also I'm
surprised that nobody has contributed anything here yet. I LOVED school, had a
great time, learned a lot, made many lasting friendships... I was sorry when the
time came to leave the place. <name ref="#P3"
type="person"><forename>Marja</forename></name></p>
         </post>.
```

4.3. Types de messages de clavardage dans le corpus cmr-getalp_org

Le corpus `cmr-getalp_org` contient différent types de messages de clavardage: `chat-message`, `chat-event`, `chat-command`, etc.) Seuls les messages de type `chat-message` (4.9) devront être analysés. Les autres (tels (4.10) et (4.11) seront recopiés en l'état.

```
(4.9)
      <post xml:id="cmr-chat-a1-1" when="2013-01-01T20:00:00" who="#P1"
alias="ufofiles"  type="chat-message">
        <p>salut KeViN2A non pas plus que sa</p>
      </post>
(4.10)
      <post xml:id="cmr-chat-a3" when-iso="2013-02-03T02:26" who="#P2"
alias="KeViN2A" type="chat-event" subtype="disconnexion">
        <p>KeViN2A(31@3d35082caf1b1a8b) s'est déconnecté: Quit:KeViN2A</p>
      </post>
(4.11)
      <post xml:id="cmr-chat-a4" when-iso="2013-02-04T00:45" who="#p2"
alias="KeViN2A" type="chat-command">
        <p>!dep 55</p>
      </post>
```

5. Vérification de qualité sur les traitements

Le projet CoMeRe ne dispose pas des ressources financières permettant de valider les résultats des traitements. Cependant, comme indiqué, les traitements MELT peuvent être très satisfaisants pour certains corpus et pas du tout pour d'autres. Il faudrait donc faire un contrôle qualité post-traitement. Si le contrôle qualité se révèle positif, alors la version `tei-v2` du corpus correspondant sera diffusé" dans ORTOLANG. Dans le cas contraire, elle ne le sera pas.

Je propose à George de s'en charger. Il faudrait alors qu'il décide d'un taux acceptable d'erreurs d'annotation, prélever des échantillons dans le corpus et décider si on peut ou non le conserver. Pour ce faire, on pourrait introduire dans AJAX un répertoire supplémentaire, par exemple `traitements-comere > tei-v2_1` dans lequel ne seraient déposés que les corpus ayant passé ce mini-test. Des commentaires seraient ajoutés (commentaire XML) en début de fichiers.

6. Citer et référencer un corpus

Rappelons nos principes de référencement / citations, qui seront mis en valeur sur le site ORTOLANG dans chaque page (de métadonnées) associée à un corpus et qui serviront à chacun de référencer son travail dans sa liste de publications.

Pour un corpus en version tei-v1

```
(6.1)
AUTEURS/DEPOSITEUR (2014). <nom du corpus> [corpus]. EDITEURS (editors).
Ortolang : Nancy [HANDLE Ortolang du corpus]
```

Pour mémoire, la chaîne `[corpus]` provient des spécifications APA qui oblige à indiquer la nature de l'objet référencé, quand ce n'est pas une publication "standard" (conférences, article de revue, livre).

Parmi les auteurs figurant en début de référence, figure le nom du dépositeur. Mais c'est le dépositeur qui décide avec ses collègues quels auteurs doivent figurer et dans quel ordre. Voici quelques exemples :

```
(6.2)
Antoniadis, G (2014).  Corpus de SMS réels dans les Alpes, smsalpes [corpus]. In
Chanier T. (ed.) Banque de corpus CoMeRe.  Ortolang.fr : Nancy.
[http:handle.net/XXX/cmr-smsalpes-tei-v1]

(6.3)
Falaise, A.  (2014). Corpus de français tchaté getalp_org [corpus]. In Chanier
T. (ed) Banque de corpus CoMeRe  Banque de corpus CoMeRe.  Ortolang.fr : Nancy.
[http:handle.net/XXX/cmr-getalp_org-tei-v1]

(6.4)
Abendroth-Timmer, D., Bechtel, M., Chanier T. & Ciekanski, M. (2014). Corpus
d'apprentissage INFRAL (Interculturel Franco-Allemand en Ligne) [corpus]. In
Chanier T. (ed.) Banque de corpus CoMeRe.  Ortolang.fr : Nancy.
[http:handle.net/XXX/cmr-infral-tei-v1]
```

Pour la version tei-v2, nous proposons

```
(6.5)
Sagot,B. (2014).  Etiquetage morpho-syntaxique du corpus smsalpes [corpus].
D'après Antoniadis, G (2014), Corpus de SMS réels dans les Alpes, smsalpes [cmr-
smsalpes-tei-v1]. Banque de corpus CoMeRe.  Ortolang.fr : Nancy.
[http:handle.net/XXX/cmr-smsalpes-tei-v2]
```

Autrement dit :

```
(6.6)
AUTEURS du POS (2014). Etiquetage morpho-syntaxique du corpus <nom cours du
corpus> [corpus]. D'après DEBUT REF DU CORPUS EN tei-v1 [URN corpus en tei-v1].
Banque de corpus CoMeRe.  Ortolang.fr : Nancy. [http:handle.net/XXX/cmr-<nom
court du corpus>-tei-v2]
```

Bien que le travail éditorial existe aussi en dans cette deuxième version (construction de `<teiheader>` notamment, il ne nous parait pas opportun de lister le nom des éditeurs afin de ne pas surcharger la référence, mais de privilégier la traçabilité de la nouvelle version. Bien sûr, comme d'habitude, les noms des contributeurs à cette nouvelle œuvre seront cités dans le `<teiheader>`

- ❖ Merci de donner votre avis car nous devons conclure et ne pourrons revenir en arrière vu le flux de traitement de CoMeRe.

7. Annexes

7.1. Organisation des répertoires dans le serveur AJAX

❖ **Notion de "corpus"**: pour mémoire un "corpus" souvent ne se réduit pas à un seul fichier TEI/XML. Un corpus peut avoir été fractionné en de multiples sous-corpus (`cmr-getalp_org-tei-v1` contient 80 fichiers/TEI). Il peut contenir un nombre plus ou moins grand de documents associés. Par exemple le corpus `cmr-smslareunion-tei-v1` contient un fichier TEI dans sa version tei-V1, un manuel PDF, un questionnaire, le tableur du questionnaire dépouillé, la fiche OLAC (fichier XML ne format DC/OLAC).

Dans le `serveur http://msh-handle.univ-bpclermont.fr/comere > Mes Fichiers`

7.1.1. Répertoires généraux

- `echantillons_comere` : répertoire où sont stockés des échantillons de corpus destinés à être lus par les partenaires extérieurs de CoMeRe : ORTOLANG, responsables formations corpus-écrits, collègues du SIG TEI-CMC. Les utilisateurs ayant accès à ce répertoire ne peuvent voir le reste du site
- `traitements-comere > tei-v1` : répertoire destiné au dépôt par le groupe Comere-LRL des corpus à traiter, par exemple `sms-lareunion-tei-v1.xml`. Des sous-répertoires peuvent apparaitre (par exemple pour `cmr-getalp_org` avec 80 fichiers XML)
- `traitements-comere > tei-v2` : répertoire destiné au groupe CoMeRe-traitements pour y déposer les fichiers résultats du traitement.

7.1.2. Par corpus

Pour chaque corpus, par exemple ici pour `cmr-smslareunion` :

- `smslareunion-v1` : répertoire où seront déposées **les versions définitives** du corpus `smslareunion` pour transfert sur le site ORTOLANG fin 2014. Il contiendra donc les versions tei-V1 (identifiant `cmr-smslareunion-tei-v1)` et tei-v2 (identifiant `cmr-smslareunion-tei-v2)` du corpus. Seuls les membres de CoMeRe-LRL peuventt y écrire. A l'heure actuelle ce répertoire est vide.
- `smslareunion-v0 > depots` : répertoire **réservé aux dépositeurs de la version originale** du corpus `smslareunion` (manuel compris). CoMeRe-LRL laisse les fichiers en l'état (tableur, XML maison, etc.) et n'y modifient rien. Le groupe vient seulement récupérer la version d'origine pour travailler dessus. Souvent plusieurs interactions avec les dépositeurs sont nécessaires afin d'affiner la qualité du dépôt, d'y ajouter des informations.
- `smslareunion-v0 > tei-v1` : répertoire où le groupe CoMeRE-LRL (et CoMeRe-TEI avec Linda Hriba) dépose la nouvelle version du corpus `smslareunion` après passage des corpus en format TEI-CMC, version tei-V1. Linda y ajoute les fiches OLAC. Seuls les membres de ce groupe peuvent y écrire

- `smslareunion-v0.1` : répertoire destiné à **l'interface entre le groupe CoMeRe-LRL et CoMeRe-qualité**. Dans le sous-répertoire `smslareunion-v0.1 >tei-v1` , CoMeRe-LRL dépose le corpus sur lequel le groupe qualité devra travailler (intrant). Dans le répertoire `smslareunion-v0.1 >tei-v1_1`, le groupe qualité dépose le résultat de son travail (extrant).

Cette organisation des répertoires est reproduite pour chaque corpus `cmr-getalp_org`, `cmr-mulce` et ses sous-corpus, `cmr-sms-alpes`, et d'autres corpus à venir dans quelques semaines ou mois (`cmr-polititweets`, `cmr-wikiconflits`, d'autres `cmr-mulce`).

7.2. Récapitulatif des éléments et attributs TEI

Nom élément	explication	ref texte	Intrant	extrant
<addName>[_addName_]</addName>	Elément balisant les surnoms, alias, uniquement lorsque la graphie d'anonymisation est rencontrée	§ 4.1	N	O
<address>[_address_]</address>	Elément balisant les adresses, uniquement lorsque la graphie d'anonymisation est rencontrée	§ 4.1	N	O
<addressingTerm>	Elément balisant des termes d'adresse, souvent, mais pas toujours préfixés par le symbole @	(3.5)	N	O
<email>	Elément balisant les adresses de courriel, qu'elles soient anonymisées ou non	§ 4.1	N	O
<f>	Elément balisant une procédure d'anonymisation	§ 4.1	O	N
<forename>	Elément balisant les prénoms, en particulier quand ils sont anomymisés (mais pas seulement ?)	§ 4.1	N	O
<head>	Sous-élément d'un <post> pouvant contenir des éléments à analyser (titre, étiquette): <title> <label>	§ 2.1.2	O	O
<interactionWord>	Elément balisant des graphies propres à l'interaction en ligne (émoticones, raccourcis de clavardage, etc.)	(3.3)	N	O
<interGrp> et <interp>	Elément à inclure dans la partie <teiheader>/<encodingDesc>/<editorialDecl>/<interpretation>/<p> du fichier TEI. Là seront ldécrits chaque balise POS de MELt avec la balise <interp>.	(2.6)	N	O
<label>	Dans le <head> d'un <post> correspond à l'étiquette d'un message. A anlyser	§ 2.1.2	O	O
<lb/>.	Intrant dans fichier TEI : élément indiquant que l'auteur du message à introduit plusieurs retours à la ligne	§ 2.1.1	O	O
<p>	Elément contenus dans le <post> où l'analyse devra se faire.	§ 2.1.1	O	O
<phr>	Elément extrant entourant tout groupe de mots à un niveau inférieur à la phrase ou l'énoncé complet (*phrase*)	§ 2.1.2	N	O
<post>	Elément supérieur contenant les messages de types courriels, forums, clavardage, blogues, textos, tweets, etc. L'analyse se fera dans le ou les sous-éléments <p>	§ 2.1	O	O
<rs type="code">	Elément et attribut entourant tout ce qui peut être repéré comme étant un code, soit code provenant d'un langage formel, soit digicode, en particulier quand la graphie	§ 4.1	N	O

	d'anonymisation est présente dans le texte			
<rs type="telephone">	élément et attribut entourant tout numéro de téléphone, anonymisé ou non	§ 4.1	N	O
<rs type="twitter-account">	Elément et attribut balisant un compte Twitter	(3.10)	N	O
<rs type="twitter-hashtag">	Elément et attribut balisant un hashtag	(3.12)	N	O
<rs type="twitter-ref-account">	Elément et attribut indiquant qu'on cite un compte Twitter	(3.12)	N	O
<rs type="twitter-retweet">	Elément et attribut balisant un retweet	(3.11)	N	O
<rs type="url">	élément et attribut entourant tout adresse URL, anonymisée ou non	§ 4.1	N	O
<s>	Elément extrant entourant tout groupe de mots correspondant à une phrase ou énoncé complet (*sentence*)	§ 2.1.1	N	O
<surname>[_surname_]</surname>	Elément balisant les patronymes, en particulier quand ils sont anomymisés (mais pas seulement ?)	§ 4.1	N	O
<title>	Dans le <head> d'un <post> correspond au tire d'un message. A nalayser	§ 2.1.2	O	O
<u>	élément contenant les tours de paroles audio. Les contenus seront à analyser, mais pour l'instant pas encore de corpus CoMeRe disponible (en préparation)	NA	O	O
<w ana="#refpos" lemma="XX">	Balise entourant une graphie. @ana référence son étiquette qui est préfixé par #. L'étiquette "refpos" doit être déclarée dans <interp>. La valeur de @lemma contient le lemme. La graphie de départ est contenu entre les balises <w>	§ 2.2.1	N	O

8. References

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L &, and Storrer, A (2012). "A TEI Schema for the Representation of Computer-mediated Communication", *Journal of the Text Encoding Initiative*, 3. [http://jtei.revues.org/476 ; DOI : 10.4000/jtei.476]

Chanier, T. & Jin, K. (2013) *Defining the online interaction space and the TEI structure for CoMeRe corpora*. [Rapport de travail]. Projet CoMeRe (Communication Médiée par les Réseaux), IR Corpus-écrits : comere.org . [comere-is-tei-v2 ; http://corpuscomere.files.wordpress.com/2014/01/tei-cmc-comere-interactionspace_131231.pdf]

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C.R., Hriba, L., Longhi, J. & Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres [Preprint] [http://hal.archives-ouvertes.fr/halshs-00953507]

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