CORPUSCOMERE
COMMUNICATION MÉDIÉE PAR LES RÉSEAUX
Salut s que NOM_4> c dcd à d pr sa cop main?
ETANDEM-NOUS BIEN !

http://comere.org
http://hdl.handle.net/11403/comere

The CoMeRe French CMC corpora and their modeling in TEI

Thierry Chanier, Céline Poudat, Ciara Wigham

ird-cmc-rennes :

International Research Days: Social Media and CMC Corpora for the eHumanities

23-24h October 2015

ORTOLANG
Open Resources and TOols for LANGuage

Consortium Corpus-écrits

Laboratoire de Recherche sur le Langage

TEI CMC
SIG TEI-CMC

CoMeRe (*Communication Médiée par les Réseaux*): a reference corpus of French CMC (2013-14)

http://comere.org
http://hdl.handle.net/11403/comere

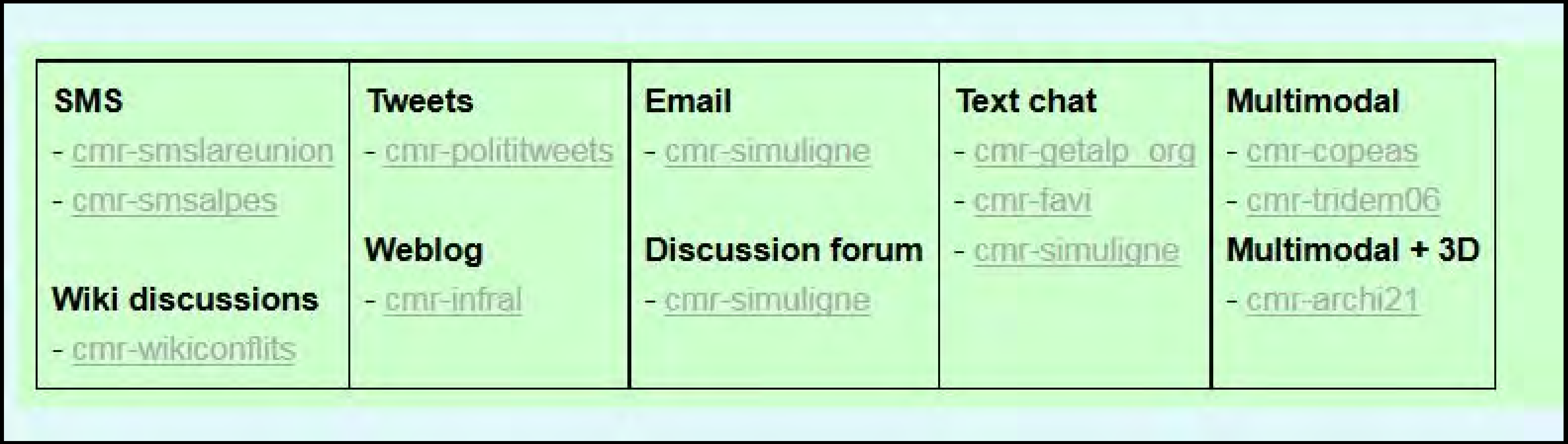
Project supported by the national consortium *Corpus-écrits*, sub-part of *Huma-Num*, and *Ortolang*

- **People:** 14 researc. from 8 research units. Coord: Chanier, T (Clermont), Poudat, C. & Sagot, B (Paris), Longhi, J. (Cergy), Antoniadis, G. (Grenoble)

Objective: Kernel corpus assembling existing corpora of different CMC genres and new corpora build on data extracted from the Internet. These heterogeneous corpora will be structured and processed in a uniform way, complemented with metadata. CoMeRe will be released as OpenData through the national infrastructure Ortolang, following constraints which will be reused for the forthcoming "*Corpus de Référence du Français*".

Variety + Standards + Open Access

Variety + Standards + Open Access



SMS	Tweets	Email	Text chat	Multimodal
- cmr-smslareunion	- cmr-polititweets	- cmr-simuligne	- cmr-getalp_org	- cmr-copeas
- cmr-smsalpes			- cmr-favi	- cmr-tridem06
	Weblog	**Discussion forum**	- cmr-simuligne	**Multimodal + 3D**
Wiki discussions	- cmr-infral	- cmr-simuligne		- cmr-archi21
- cmr-wikiconflits				

Variety + Standards + Open Access

- People often wonder: "what did you choose the *Text Encoding Initiative* to encode multimodal interactions?
- These interactions can be viewed as text
  - BALDRY & THIBAULT (2006) consider “texts to be meaning-making events whose functions are defined in particular social contexts,” following HALLIDAY (1989:10) “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.”
- Mainstream of oral corpora are encoded into TEI
- TEI offers a very rich way to describe the project corpus (on top of the interactions set)
- Opportunity to wrok at a European level

Variety + Standards + Open Access

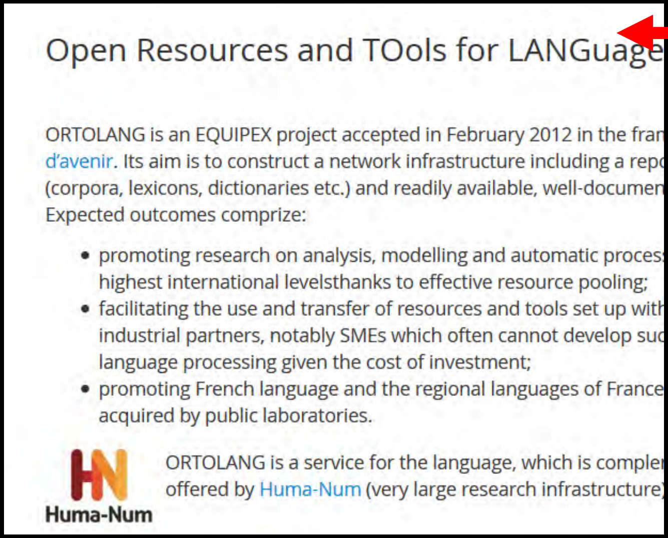
Opendata criteria

- “**Availability and Access**: the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Reuse and Redistribution**: the data must be provided under terms that permit reuse and redistribution including the intermixing with other datasets. The data must be machine-readable.
- **Universal Participation**: everyone must be able to use, reuse and redistribute – there should be no discrimination against fields of endeavor or against persons or groups. For example, ‘non-commercial’ restrictions that would prevent ‘commercial’ use, or restrictions of use for certain purposes (e.g. only in education), are not allowed. “OpenDefinition.org

OPEN KNOWLEDGE
OPEN DATA
OPEN CONTENT
OPEN SERVICE

Example of CoMeRe licences

- Falaise, A. (2014). *Corpus de français tchaté getalp_org*. [cmr-getalp_org]. cc BY NC SA
- Antoniadis, G (2014). *Corpus de SMS réels dans les Alpes, smsalpes* . [cmr-smsalpes]. cc BY NC SA
- Longhi, J., Marinica, C., Borzic, B., Alkhouli, A. Polititweets. (2014). *Corpus de tweets provenant de comptes politiques influents*. [cmr-polititweets] cc BY
- Ledegen, G. (2014). *Grand corpus de sms smslareunion* . [cmr-smslareunion] cc BY
- Yun, H. & Chanier, T. (2014). *Corpus d'apprentissage FAVI (Français académique virtuel international*). [cmr-favi]. cc 0 ZERO
- Abendroth-Timmer, D., Bechtel, M., Chanier T. & Ciekanski, M. (2014). *Corpus d'apprentissage INFRAL (Interculturel Franco-Allemand en Ligne)*. [cmr-infral] cc 0 ZERO
- Reffay, C. Chanier, T. Lamy, M.-N. & Betbeder, M.-L. (2014) *Corpus d'apprentissage Interactions Simuligne (Simulation en ligne en apprentissage des langues)*. [cmr-simuligne] cc 0 ZERO



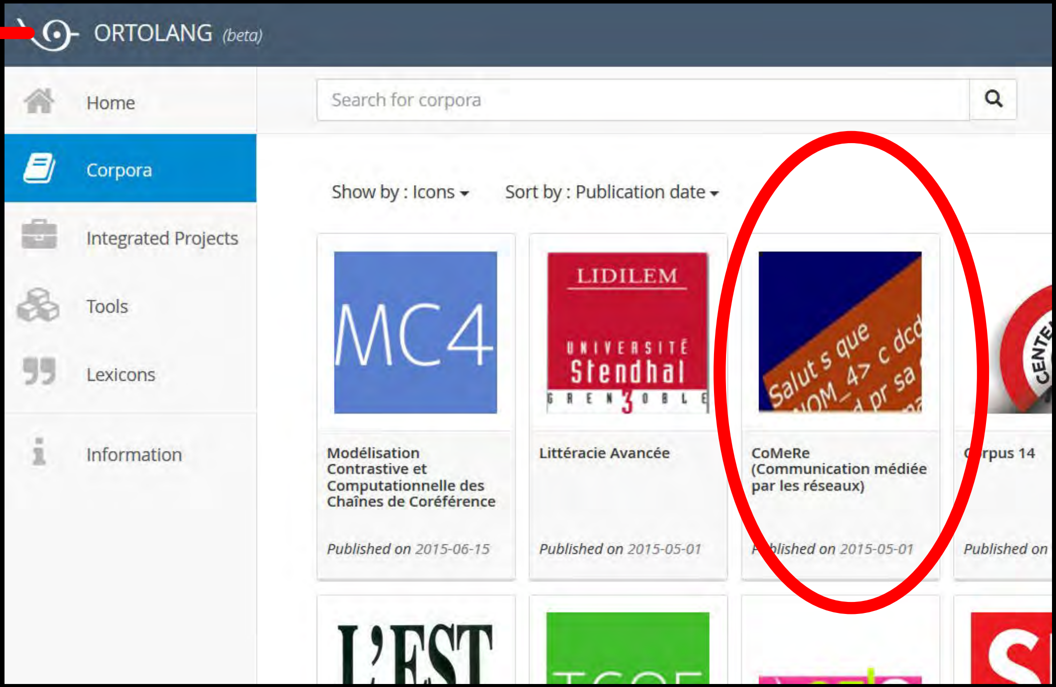
Open Resources and TOols for LANGuage

ORTOLANG is an EQUIPEX project accepted in February 2012 in the fran
d'avenir. Its aim is to construct a network infrastructure including a repo
(corpora, lexicons, dictionaries etc.) and readily available, well-documen
Expected outcomes comprize:

- promoting research on analysis, modelling and automatic proces
highest international levelsthanks to effective resource pooling;
- facilitating the use and transfer of resources and tools set up wit
industrial partners, notably SMEs which often cannot develop su
language processing given the cost of investment;
- promoting French language and the regional languages of France
acquired by public laboratories.

Huma-Num

ORTOLANG is a service for the language, which is comple
offered by Huma-Num (very large research infrastructure



Modélisation Contrastive et Computationnelle des Chaînes de Coréférence
Published on 2015-06-15

Littéracie Avancée
Published on 2015-05-01

CoMeRe (Communication médiée par les réseaux)
Published on 2015-05-01

rpus 14
Published on

Corpora repository in ORTOLANG

http://hdl.handle.net/11403/comere

Cuurent list of corpora

- 1) Antoniadis, G (2014). *Corpus de SMS réels dans les Alpes, smsalpes* [corpus]. In Chanier T. (ed.) Banque de corpus CoMeRe. Ortolang.fr : Nancy. [http://hdl.handle.net/11403/comere/cmr-smsalpes]
- 2) 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://hdl.handle.net/11403/comere/cmr-getalp_org]
- 3) Ledegen, G. (2014). *Grand corpus de sms SMS La Réunion* [corpus] ….
- 4) Reffay, C. Chanier, T. Lamy, M.-N. & Betbeder, M.-L. (2014). *Corpus Interactions Simuligne (Simulation en ligne en apprentissage des langues*) [corpus]…
- 5) Yun, H. & Chanier, T. (2014). *Corpus d'apprentissage FAVI (Français académique virtuel international)* [corpus…
- 6) Abendroth-Timmer, D., Bechtel, M., Chanier T. &Ciekanski, M. (2014). *Corpus d'apprentissage INFRAL (Interculturel Franco-Allemand en Ligne*). [corpus]…
- 7) Longhi, J., Marinica, C., Borzic, B. & Alkhouli, A. (2014) *Corpus de tweets provenant de comptes politiques influents*. [corpus]…
- 8) Chanier, T. & Audras, I. (2015). Tridem06 corpus: intercultural competence in online exolingual group exchanges [...]
- 9) Chanier, T. & Wigham, C.R. (2015). Archi21 corpus: collaborative language and architectural learning in Second Life [...]
- 10) Chanier, T., Reffay, C., Betbeder, M-L., Ciekanski, M. & Lamy, M-N. (2015). Copéas corpus: online language learning within an audiographic environment [...]
- 11) Poudat,C., Grabar , N. Kun, J. & Paloque-Berges, C. (2015). Corpus wikiconflits, conflits dans le Wikipédia francophone [...]

Corpora composed of verbal acts

Ref	Tokens	Partici.	Posts	Envir.	
(Antoniadis,2014)	449 313	359	22 052	SMS	→ Informal business
(Falaise, 2014)	35 M	25 000	3 M	textchat	→ Informal
(Ledegen, 2014)	357 000	850	22 000	SMS	→ Informal
(Reffay et al., 2014)	600 000	67 + 4 groups	- textchat: 6 790 - emails: 2 030 - forums: 2 686	LMS	→ education
(Yun, Chanier, 2014)	77 605	31 + 2 courses	7 750	textchat	→ education
(Abendroth-Timmer et al., 2014)	273 546	26 + 4 groups	1 200	Blog	→ education
(Longhi, Marinica, 2014)	567 851	205	34273	Tweet	→ politics
(Poudat et al., 2015)	489 000	3971	4456	Wiki discussions	→ science

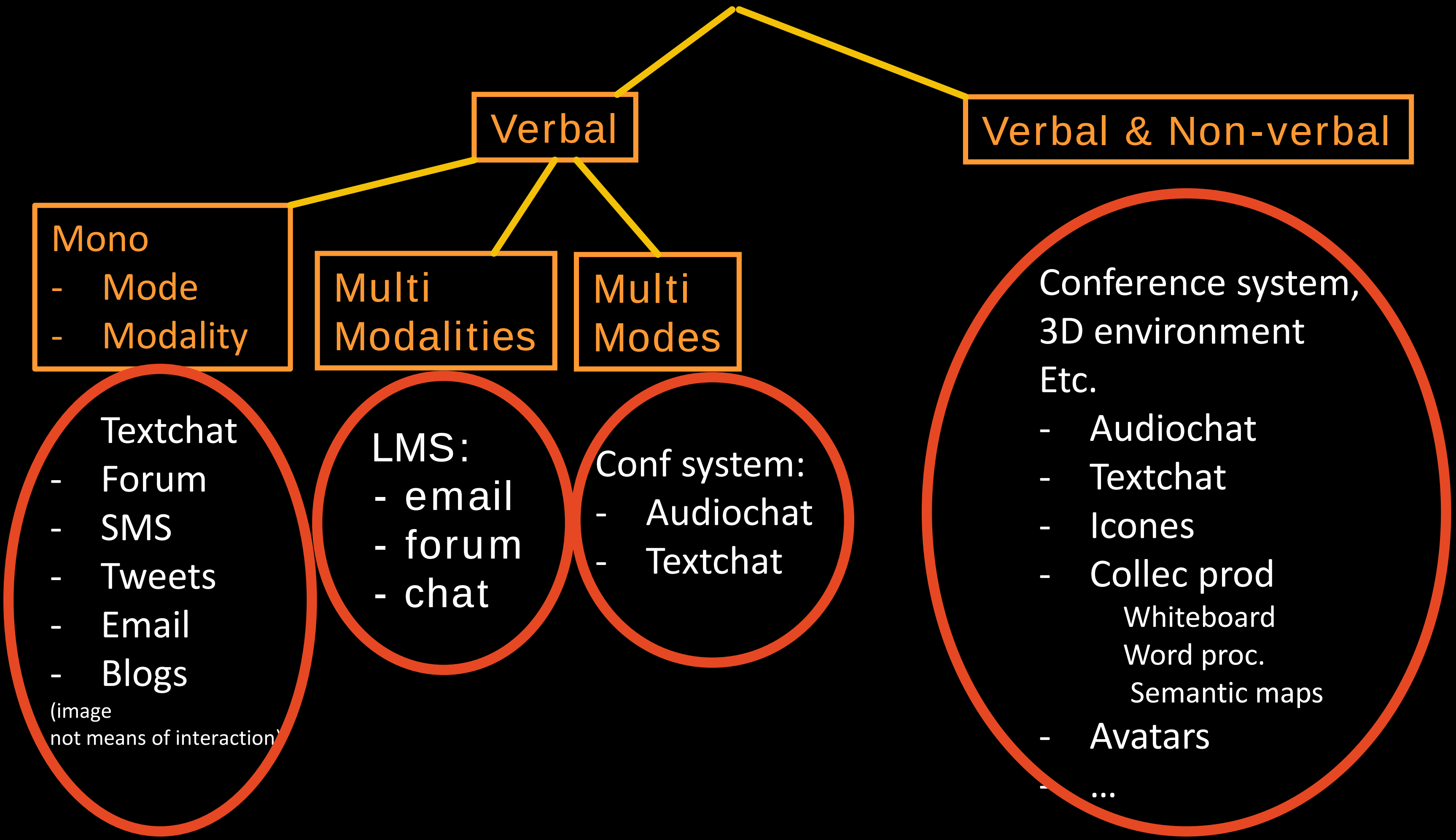
verbal & non-verbal acts (LETEC corpora)

Ref	Tokens	Partici.	Posts, U, Prod	Envir.
(Chanier & Audras, 2015)	184 594	62 + 12 groups	- audio: 2 809 - chat: 248 - non-verbal: 1 058 - blog: 779	Conference system
(Chanier & Wigham, 2015)	27 912	18 + 4 groups	- audio: 1 690 - chat: 669 - non-verbal: 2 452	3D environment
(Chanier , Reffay et al., 2015)	127 228	16 + 2 groups	- audio: 7 718 - chat: 1 566 - non-verbal: 5 790	Conference system

Interaction Space Model

Implementation in CoMeRe corpora

Environments



Verbal

Verbal & Non-verbal

Mono
- Mode
- Modality

Multi Modalities

Multi Modes

Textchat
- Forum
- SMS
- Tweets
- Email
- Blogs

(image
not means of interaction)

LMS:
- email
- forum
- chat

Conf system:
- Audiochat
- Textchat

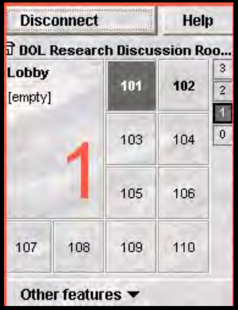
Conference system,
3D environment
Etc.
- Audiochat
- Textchat
- Icones
- Collec prod
  Whiteboard
  Word proc.
  Semantic maps
- Avatars
- …

Interaction Space Model

Interaction Space

Time(s)

Course
Session
Channel
Simultaneity

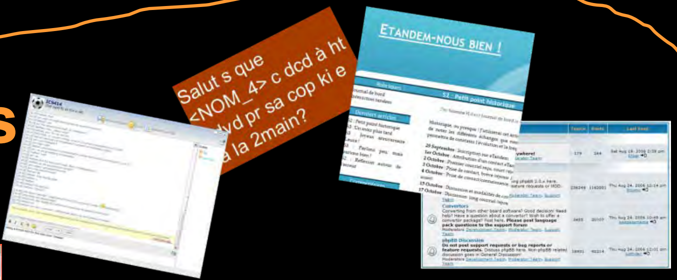


Locations

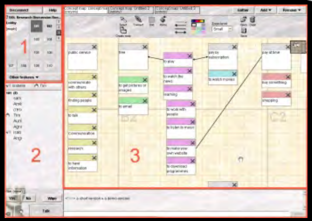
Participants

Author
Adresse(s)
Group
Network

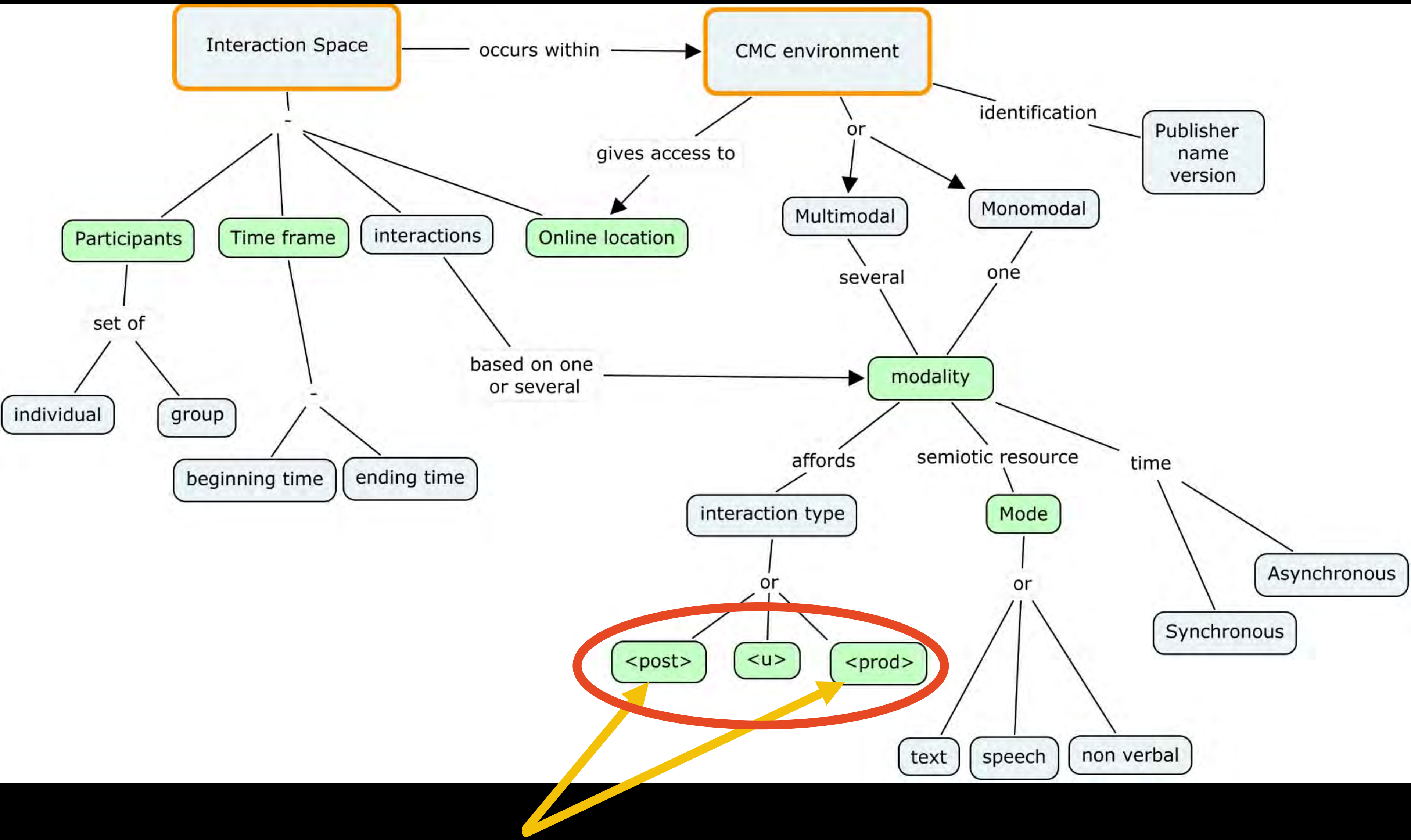
Environments



Salut s que <NOM_4> c dcd à ht
...vd pr sa cop ki e
...d la 2main?



http://wiki.tei-c.org/index.php/SIG:CMC/Draft:_A_basic_schema_for_representing_CMC_in_TEI



New macro-level elements

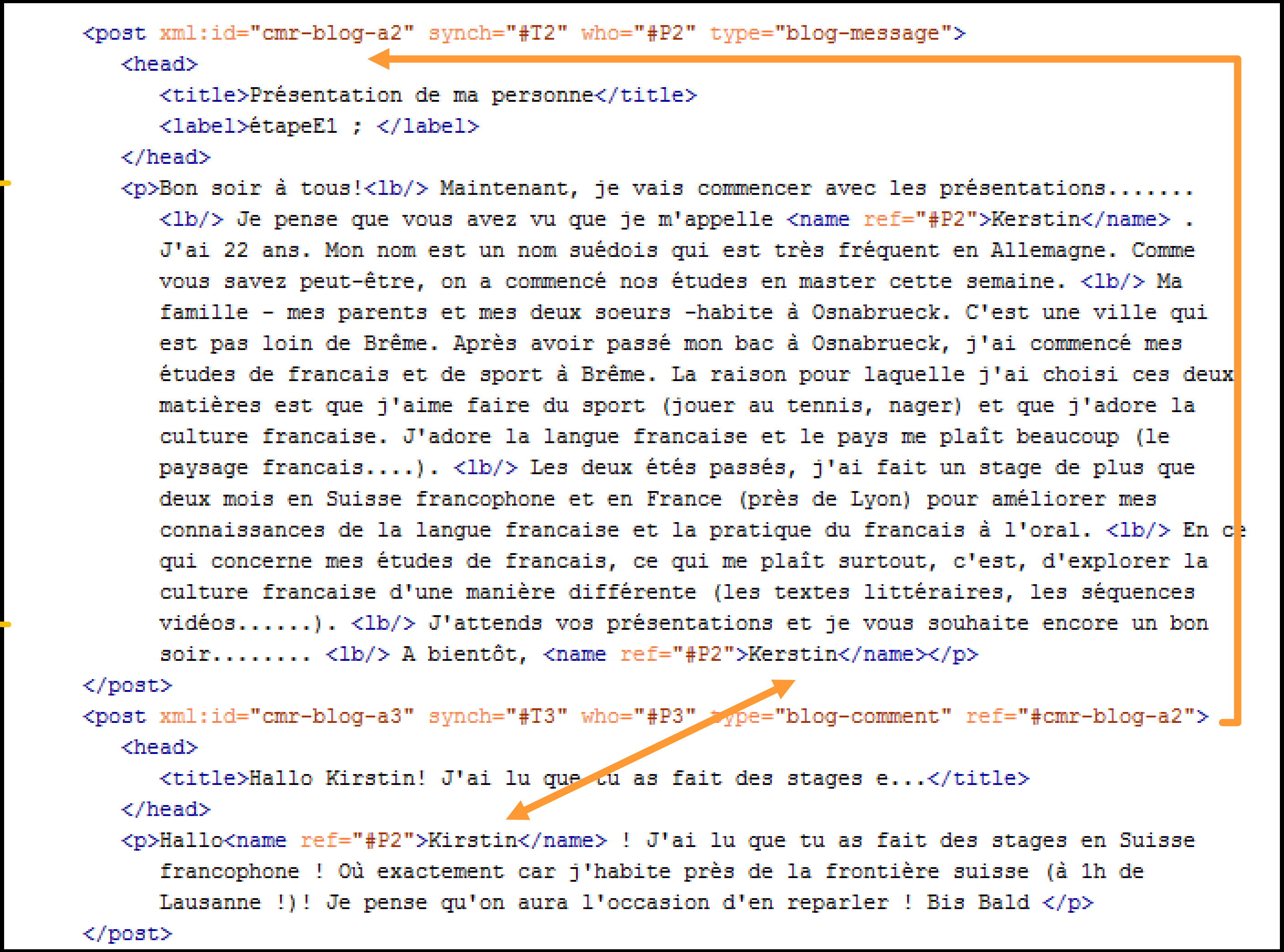
Blog: message and comment

message

Title
label

Contents
/ body

comment



```
<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
      est pas loin de Brême. Après avoir passé mon bac à Osnabrueck, j'ai commencé mes
      études de francais et de sport à Brême. La raison pour laquelle j'ai choisi ces deux
      matières est que j'aime faire du sport (jouer au tennis, nager) et que j'adore la
      culture francaise. J'adore la langue francaise et le pays me plaît beaucoup (le
      paysage francais....). <lb/> Les deux étés passés, j'ai fait un stage de plus que
      deux mois en Suisse francophone et en France (près de Lyon) pour améliorer mes
      connaissances de la langue francaise et la pratique du francais à l'oral. <lb/> En ce
      qui concerne mes études de francais, ce qui me plaît surtout, c'est, d'explorer la
      culture francaise d'une manière différente (les textes littéraires, les séquences
      vidéos......). <lb/> J'attends vos présentations et je vous souhaite encore un bon
      soir........ <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>
```

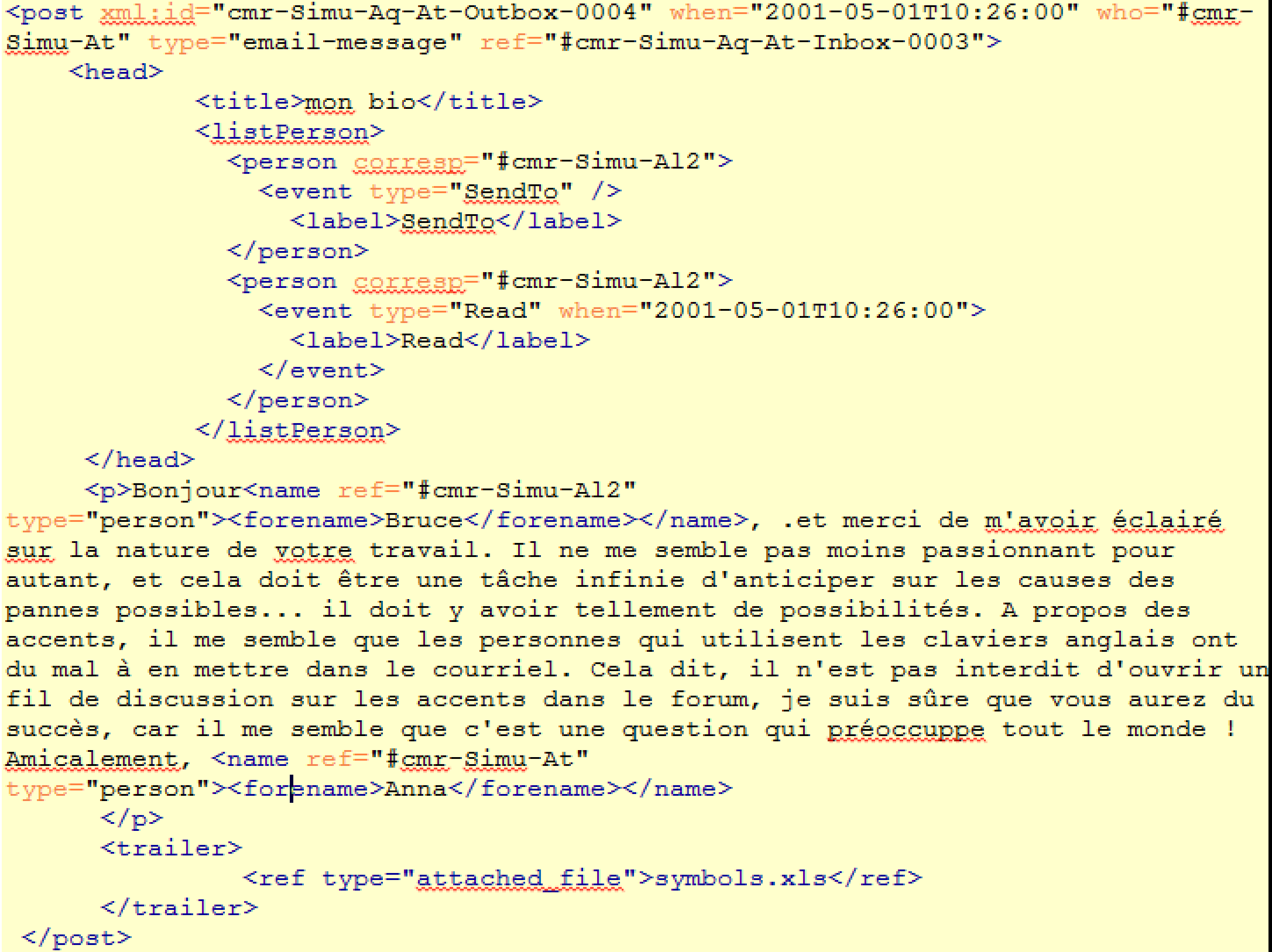
More complex (cumbersome?): email (here), forum

Response to what?

Sent to whom? Read by whom?

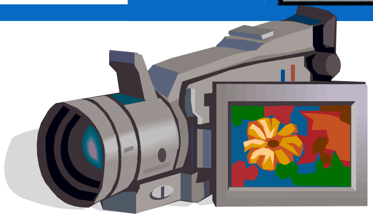
May contain HTML, Table,etc.

Attached doc



```
<post xml:id="cmr-Simu-Aq-At-Outbox-0004" when="2001-05-01T10:26:00" who="#cmr-
Simu-At" type="email-message" ref="#cmr-Simu-Aq-At-Inbox-0003">
    <head>
            <title>mon bio</title>
            <listPerson>
              <person corresp="#cmr-Simu-Al2">
                <event type="SendTo" />
                  <label>SendTo</label>
              </person>
              <person corresp="#cmr-Simu-Al2">
                <event type="Read" when="2001-05-01T10:26:00">
                  <label>Read</label>
                </event>
              </person>
            </listPerson>
     </head>
     <p>Bonjour<name ref="#cmr-Simu-Al2"
type="person"><forename>Bruce</forename></name>, .et merci de m'avoir éclairé
sur la nature de votre travail. Il ne me semble pas moins passionnant pour
autant, et cela doit être une tâche infinie d'anticiper sur les causes des
pannes possibles... il doit y avoir tellement de possibilités. A propos des
accents, il me semble que les personnes qui utilisent les claviers anglais ont
du mal à en mettre dans le courriel. Cela dit, il n'est pas interdit d'ouvrir un
fil de discussion sur les accents dans le forum, je suis sûre que vous aurez du
succès, car il me semble que c'est une question qui préoccuppe tout le monde !
Amicalement, <name ref="#cmr-Simu-At"
type="person"><forename>Anna</forename></name>
      </p>
      <trailer>
               <ref type="attached_file">symbols.xls</ref>
      </trailer>
 </post>
```

Modality interplay

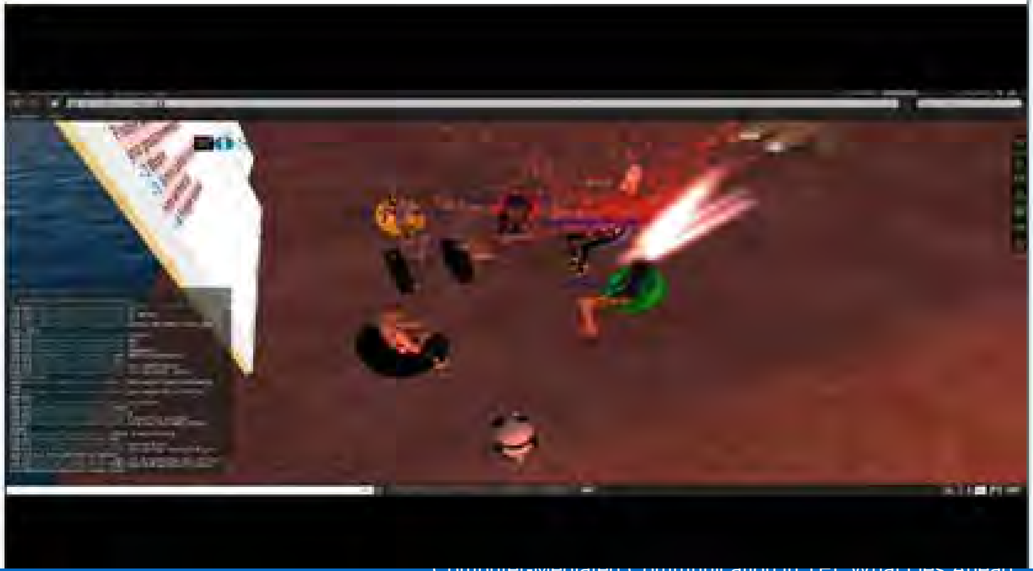


1.5 mn video

tingrabu	tfrez2	romeorez	quentinrez
ok for me this presentaion was become too fast because it's always the same in our architectural school we have not time and too quickly sorry and we can't do good images because it's less time euh I don't know [...] and it's a big matter because we always talk about teleportation [...] an everyday lack of time ok thank you quentinrez and this is very difficult [...]	it went too quickly or it was too early in the week? ok too quickly means you didn't have enough time	i think it was to early too	yes, it's an everyday lack of time

* Paper: (Wigham & Chanier, 2013) CALL journal

* Data: (Wigham, 2013) LETEC corpus



Multimodalité : Verbal et non verbal

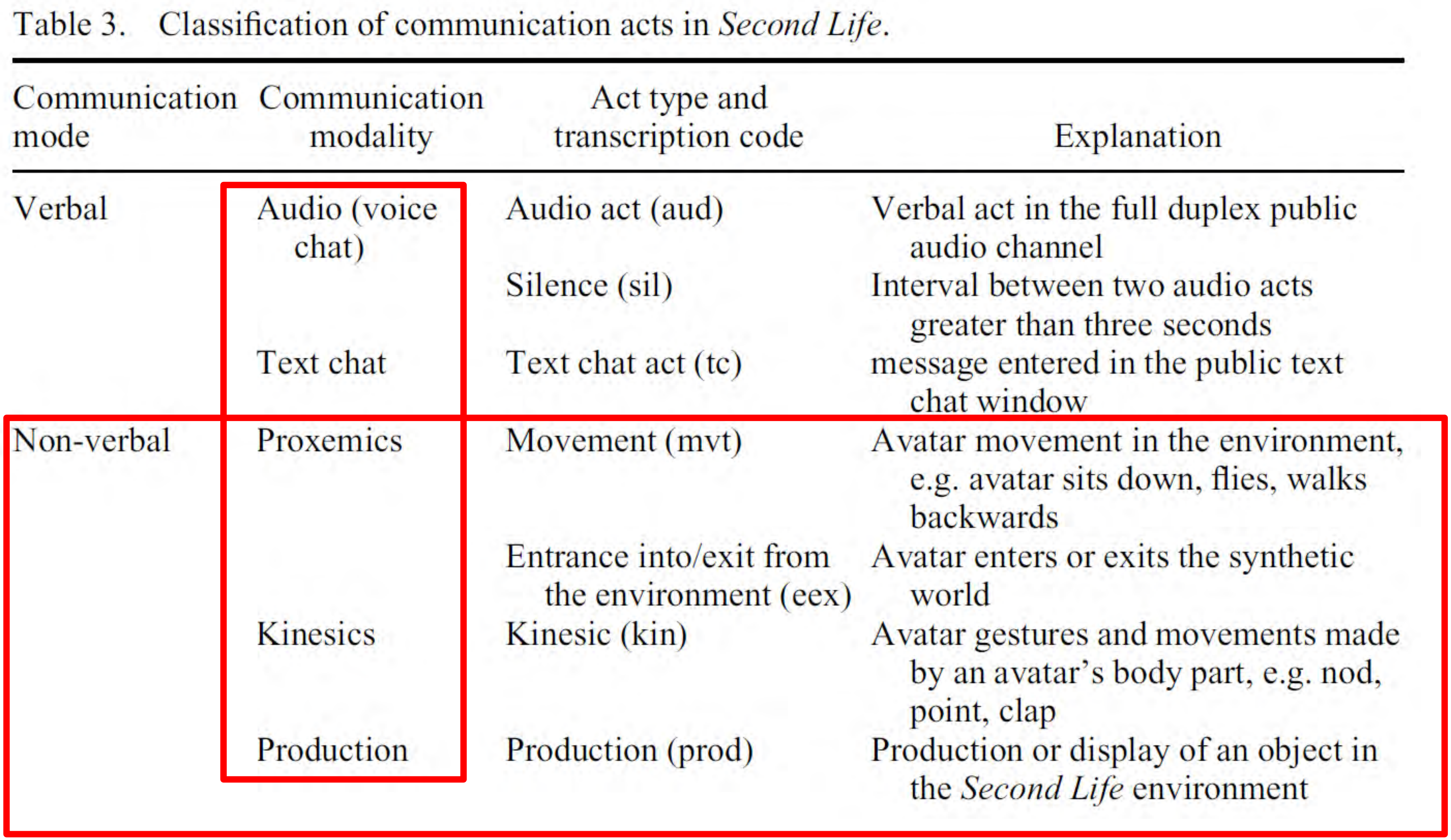


Table 3. Classification of communication acts in *Second Life*.

Communication mode	Communication modality	Act type and transcription code	Explanation
Verbal	Audio (voice chat)	Audio act (aud)	Verbal act in the full duplex public audio channel
		Silence (sil)	Interval between two audio acts greater than three seconds
	Text chat	Text chat act (tc)	message entered in the public text chat window
Non-verbal	Proxemics	Movement (mvt)	Avatar movement in the environment, e.g. avatar sits down, flies, walks backwards
		Entrance into/exit from the environment (eex)	Avatar enters or exits the synthetic world
	Kinesics	Kinesic (kin)	Avatar gestures and movements made by an avatar's body part, e.g. nod, point, clap
	Production	Production (prod)	Production or display of an object in the *Second Life* environment

(Wigham & Chanier, 2013)

Modality interplay

Audio

kinesics

chat

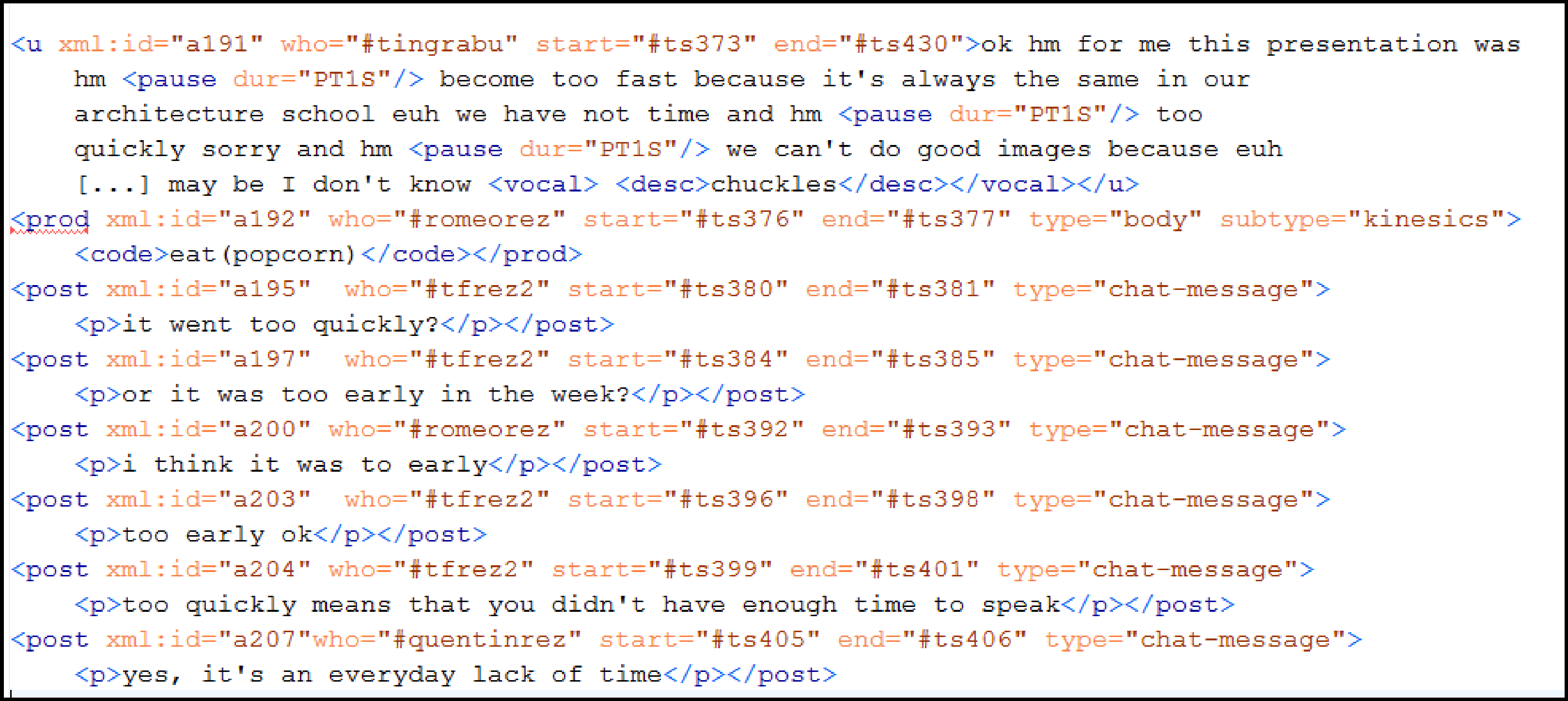
chat

chat

chat

chat

chat

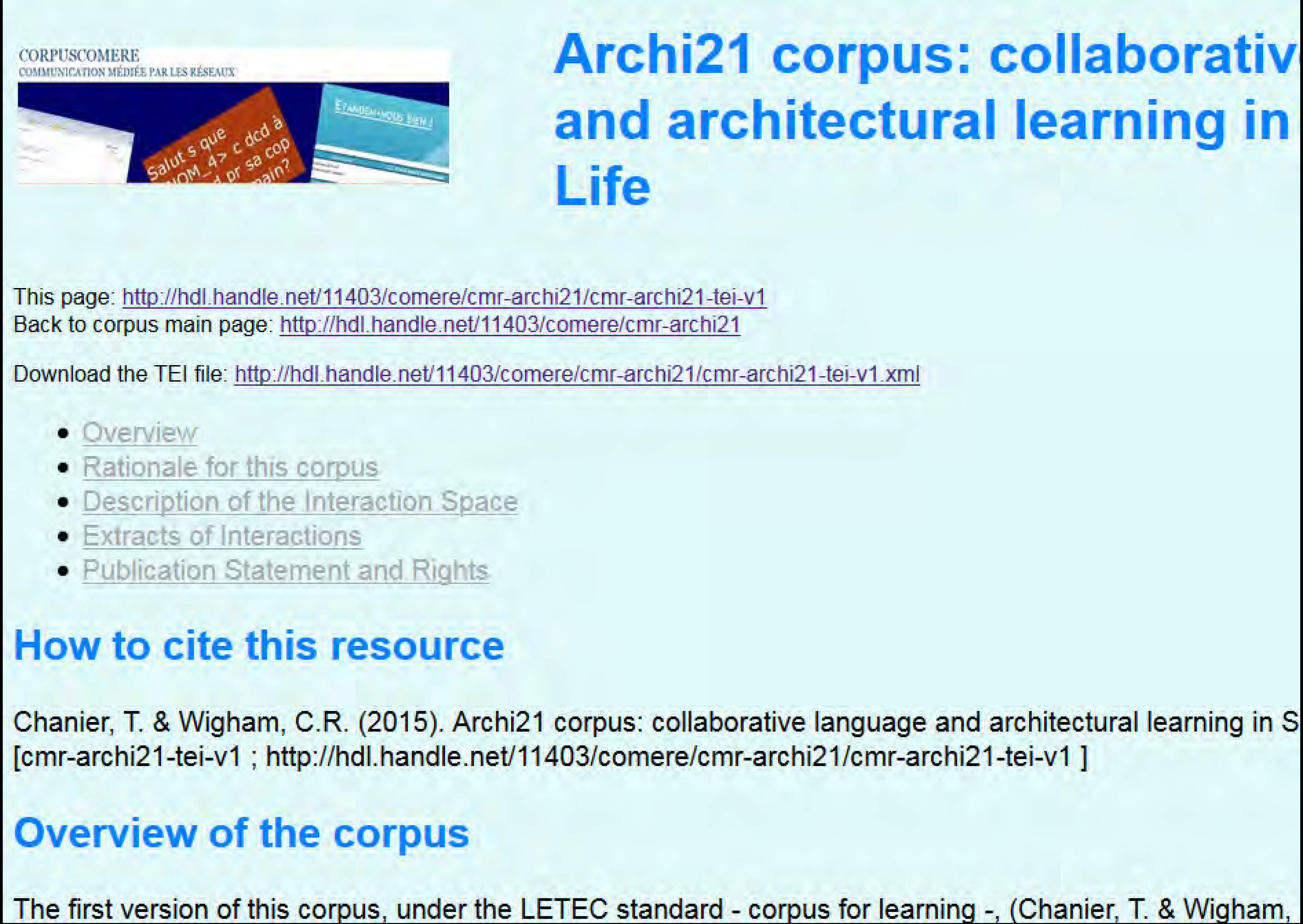


```
<u xml:id="a191" who="#tingrabu" start="#ts373" end="#ts430">ok hm for me this presentation was
    hm <pause dur="PT1S"/> become too fast because it's always the same in our
    architecture school euh we have not time and hm <pause dur="PT1S"/> too
    quickly sorry and hm <pause dur="PT1S"/> we can't do good images because euh
    [...] may be I don't know <vocal> <desc>chuckles</desc></vocal></u>
<prod xml:id="a192" who="#romeorez" start="#ts376" end="#ts377" type="body" subtype="kinesics">
    <code>eat(popcorn)</code></prod>
<post xml:id="a195"  who="#tfrez2" start="#ts380" end="#ts381" type="chat-message">
    <p>it went too quickly?</p></post>
<post xml:id="a197"  who="#tfrez2" start="#ts384" end="#ts385" type="chat-message">
    <p>or it was too early in the week?</p></post>
<post xml:id="a200" who="#romeorez" start="#ts392" end="#ts393" type="chat-message">
    <p>i think it was to early</p></post>
<post xml:id="a203"  who="#tfrez2" start="#ts396" end="#ts398" type="chat-message">
    <p>too early ok</p></post>
<post xml:id="a204" who="#tfrez2" start="#ts399" end="#ts401" type="chat-message">
    <p>too quickly means that you didn't have enough time to speak</p></post>
<post xml:id="a207"who="#quentinrez" start="#ts405" end="#ts406" type="chat-message">
    <p>yes, it's an everyday lack of time</p></post>
```

Detailing the corpus project in TEI

To support resuse by other researchers

Using TEI header



Archi21 corpus: collaborativ
and architectural learning in
Life

This page: http://hdl.handle.net/11403/comere/cmr-archi21/cmr-archi21-tei-v1
Back to corpus main page: http://hdl.handle.net/11403/comere/cmr-archi21

Download the TEI file: http://hdl.handle.net/11403/comere/cmr-archi21/cmr-archi21-tei-v1.xml

- Overview
- Rationale for this corpus
- Description of the Interaction Space
- Extracts of Interactions
- Publication Statement and Rights

How to cite this resource

Chanier, T. & Wigham, C.R. (2015). Archi21 corpus: collaborative language and architectural learning in S
[cmr-archi21-tei-v1 ; http://hdl.handle.net/11403/comere/cmr-archi21/cmr-archi21-tei-v1]

Overview of the corpus

The first version of this corpus, under the LETEC standard - corpus for learning -, (Chanier, T. & Wigham,

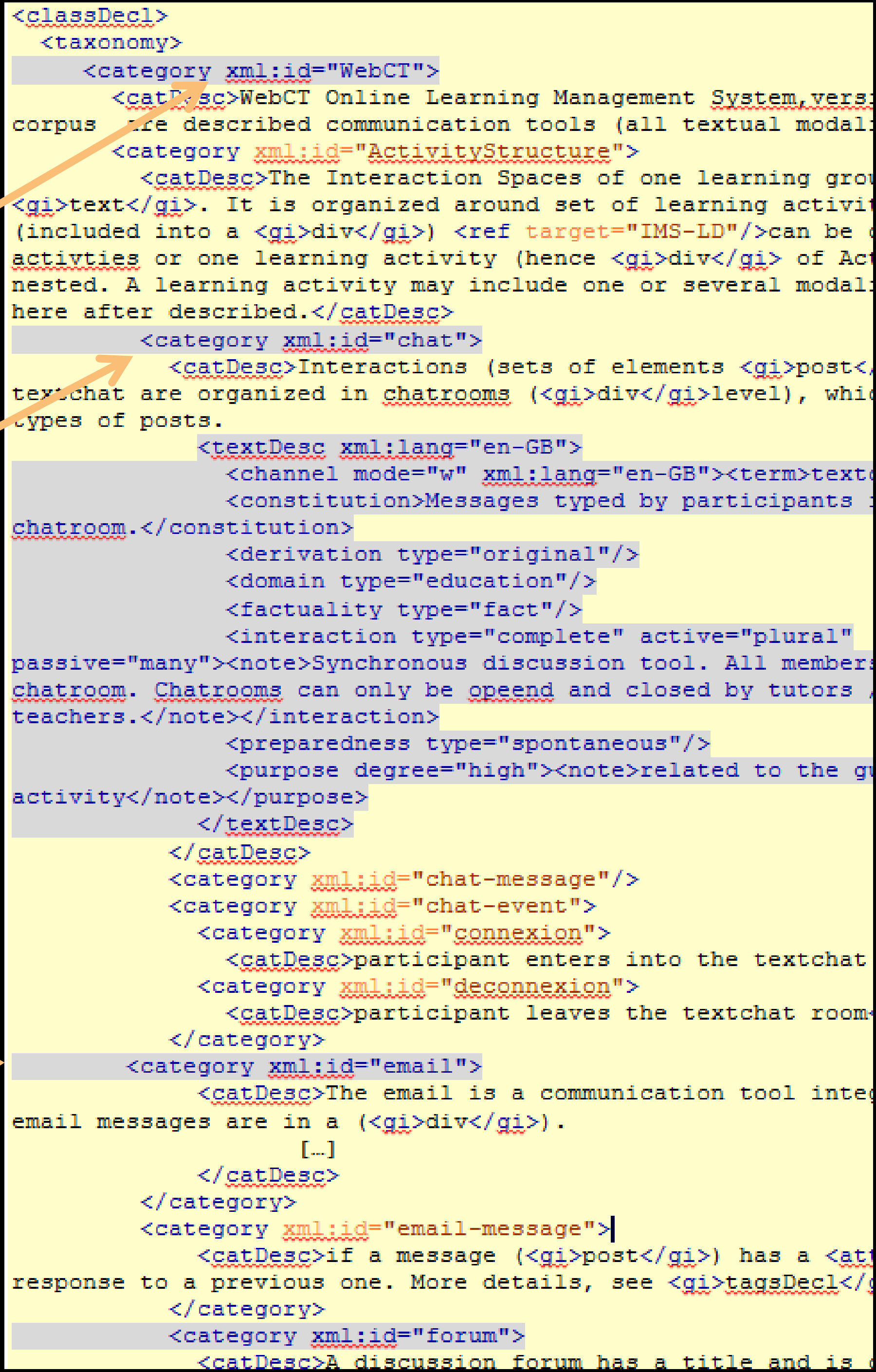
Multimodality environment: general features and affordances (LMS- Learning Management System)

LMS

textchat

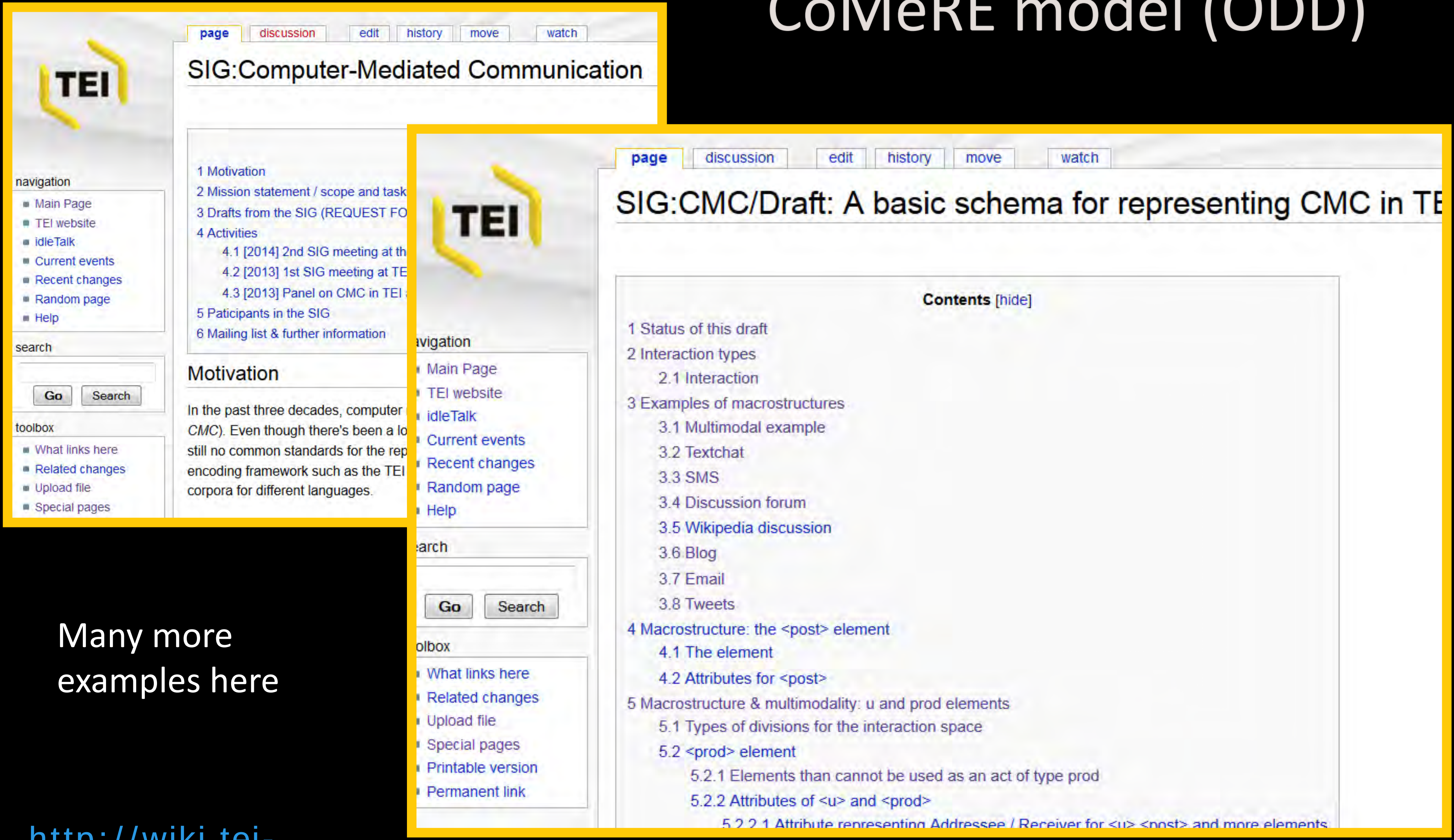
email

forum



```
<classDecl>
  <taxonomy>
     <category xml:id="WebCT">
       <catDesc>WebCT Online Learning Management System,vers
corpus  are described communication tools (all textual modal
       <category xml:id="ActivityStructure">
         <catDesc>The Interaction Spaces of one learning grou
<gi>text</gi>. It is organized around set of learning activit
(included into a <gi>div</gi>) <ref target="IMS-LD"/>can be 
activties or one learning activity (hence <gi>div</gi> of Act
nested. A learning activity may include one or several modal
here after described.</catDesc>
         <category xml:id="chat">
           <catDesc>Interactions (sets of elements <gi>post<
textchat are organized in chatrooms (<gi>div</gi>level), whic
types of posts.
             <textDesc xml:lang="en-GB">
               <channel mode="w" xml:lang="en-GB"><term>text
               <constitution>Messages typed by participants 
chatroom.</constitution>
               <derivation type="original"/>
               <domain type="education"/>
               <factuality type="fact"/>
               <interaction type="complete" active="plural"
passive="many"><note>Synchronous discussion tool. All member
chatroom. Chatrooms can only be opeend and closed by tutors 
teachers.</note></interaction>
               <preparedness type="spontaneous"/>
               <purpose degree="high"><note>related to the g
activity</note></purpose>
             </textDesc>
           </catDesc>
           <category xml:id="chat-message"/>
           <category xml:id="chat-event">
             <category xml:id="connexion">
               <catDesc>participant enters into the textchat
             <category xml:id="deconnexion">
               <catDesc>participant leaves the textchat room
           </category>
        <category xml:id="email">
             <catDesc>The email is a communication tool inte
email messages are in a (<gi>div</gi>).
                    […]
             </catDesc>
         </category>
         <category xml:id="email-message">
             <catDesc>if a message (<gi>post</gi>) has a <at
response to a previous one. More details, see <gi>tagsDecl</
           </category>
           <category xml:id="forum">
             <catDesc>A discussion forum has a title and is 
```

CoMeRE model (ODD)



page discussion edit history move watch

SIG:Computer-Mediated Communication

1 Motivation
2 Mission statement / scope and task
3 Drafts from the SIG (REQUEST FO
4 Activities
4.1 [2014] 2nd SIG meeting at th
4.2 [2013] 1st SIG meeting at TE
4.3 [2013] Panel on CMC in TEI
5 Paticipants in the SIG
6 Mailing list & further information

Motivation

In the past three decades, computer CMC). Even though there's been a lo still no common standards for the rep encoding framework such as the TEI corpora for different languages.

page discussion edit history move watch

SIG:CMC/Draft: A basic schema for representing CMC in TE

Contents [hide]

1 Status of this draft
2 Interaction types
2.1 Interaction
3 Examples of macrostructures
3.1 Multimodal example
3.2 Textchat
3.3 SMS
3.4 Discussion forum
3.5 Wikipedia discussion
3.6 Blog
3.7 Email
3.8 Tweets
4 Macrostructure: the <post> element
4.1 The element
4.2 Attributes for <post>
5 Macrostructure & multimodality: u and prod elements
5.1 Types of divisions for the interaction space
5.2 <prod> element
5.2.1 Elements than cannot be used as an act of type prod
5.2.2 Attributes of <u> and <prod>
5.2.2.1 Attribute representing Addressee / Receiver for <u> <post> and more elements

Many more examples here

http://wiki.tei-c.org/index.php/SIG:CMC/Draft:_A_metadata_schema_for_CMC

http://wiki.tei-c.org/index.php/SIG:CMC/CoMeRe_schema_draft_for_representing_CMC_in_TEI_%282014%29

CoMeRe team

Documentation and events : http://comere.org

Repository: http://hdl.handle.net/11403/comere