Corpus complexes et standards : un retour sur le projet CoMeRe

Ciara R. Wigham, Céline Poudat

Wigham, C.R. & Poudat, C. (2020). Corpus complexes et standards : un retour sur le projet CoMeRe. *Corpus*, 20 [En ligne]. http://journals.openedition.org/corpus/4736.

Corpus complexes et standards : un retour sur le projet CoMeRe

Ciara R. Wigham
Laboratoire de Recherche sur le Langage, Université Clermont Auvergne, France
Céline Poudat
Université Côte d'Azur, CNRS, BCL, France

Résumé

Le présent article se propose de revenir sur le projet national CoMeRe (Communication Médiée par les Réseaux) en insistant sur la complexité du corpus développé. Constitué de quatorze sous-corpus variés, le corpus CoMeRe est un corpus de référence de la communication médiée par les réseaux en français. Quatorze enseignants-chercheurs de huit laboratoires différents se sont impliqués dans le projet et ont été guidés par trois mots clés lors de leurs collaborations : *variété*, *standards* et *accès ouvert*.

Le corpus CoMeRe a ainsi été construit sur une **hypothèse de variété** et contient une gamme étendue d'interactions de la CMR dont nous restituons les différences et les caractéristiques principales (courriels, clavardage, SMS, forums de discussion Internet, blogs, tweets, discussions Wikipédia, interactions provenant de mondes synthétiques). Nous détaillons ensuite comme le corpus CoMeRe a été rendu **interopérable** : les quatorze sous-corpus ont été standardisés, suivant le modèle de l'espace d'interaction élaboré lors du projet (Chanier & Jin, 2013) d'une part et suivant les propositions de représentation standardisée des corpus de la CMR en TEI (Text Encoding Initiative, 2019) élaborées en lien avec les partenaires européens. Enfin, les collègues tenaient à diffuser le corpus en **accès ouvert** pour permettre son utilisation par la communauté scientifique. Nous revenons sur les retombées du projet liées à la diffusion du corpus dans la conclusion de notre article.

Abstract

The aim of this contribution is to review the national research project CoMeRe (*Communication Médiée par les Réseaux* - Networked-Mediated Communication) and, in particular, focus on the the complexity of the corpus it developed, structured, and disseminated. The CoMeRe corpus is a reference corpus for computer-mediated communication (CMC) in French comprising fourteen sub-corpora. Fourteen researchers from eight different laboratories were involved in the project and three key words guided their collaborations: variety, standards, and open access.

The CoMeRe corpus is composed of a wide range of heterogeneous CMC genres (emails, text chat, SMS, Internet discussion forums, blogs, tweets, Wikipedia discussions, interactions from synthetic worlds). In the first section of the article, we underline their main characteristics of the different CMC genres and highlight their similarities and differences. We then describe the choices made to support corpus interoperability: the fourteen sub-corpora were structured in a standardized manner in accordance with the Interaction Space model developed within the project (Chanier & Jin, 2013) and, in collaboration with European partners, following guidelines for standardizing CMC corpora in TEI (Text Encoding Initiative, 2019). The CoMeRe corpus was released in an open-access format so as to encourage future use by the scientific community. In the article's conclusion, we underline the different implications of the corpus' dissemination.

1. Introduction

Le présent article répond à un double objectif dans ce numéro : il s'agira d'une part de revenir sur le projet national CoMeRe (Communication Médiée par les Réseaux) en insistant sur la complexité du corpus de référence développé, constitué de quatorze sous-corpus variés, et sur les standards adoptés en lien avec cette complexité, et d'autre part de rendre hommage à Thierry Chanier qui a été l'initiateur de ce projet et l'artisan de sa réussite.

Le projet CoMeRe est né en 2014 dans le cadre du consortium de linguistique Corpus Écrits de la TGIR Huma-Num qui avait lancé un groupe de travail sur les 'Nouvelles formes de communication'. Piloté par Thierry Chanier et Céline Poudat, ce dernier rassemblait en grande partie des chercheurs qui avaient développé différents types de corpus de communication médiée par les réseaux (*Computer Mediated Communication*, CMC en anglais, CMR en français) au sein de leurs unités de recherche locales.

Souhaitant travailler en mode 'projet' au sein du groupe afin d'inciter une vraie collaboration à une échelle nationale, CoMeRe s'est inspiré de deux projets européens :

- Le premier, SONAR (STEVIN *Nederlandstalig Referentiecorpus*, Oostdijk et al., 2008), avait développé et structuré un corpus de référence du hollandais contemporain écrit de 500 millions de mots qui contenait non seulement des genres conventionnels (presse, rapports, etc.), mais également des données provenant des nouveaux médias, telles que le clavardage, les SMS, les forums Internet et les courriers électroniques.

- Le deuxième, DeRiK (Deutsches *Referenzkorpus zur internetbasierten Kommunikation*, Beißwenger, Ermakova, Geyken et al., 2013), avait développé un sous-corpus du corpus de référence de l'allemand oral DWDS (DWDS, 2013) représentatif de la communication médiée par les réseaux en allemand.

En se fixant un délai de deux ans, l'ambition du projet CoMeRe était donc de développer un corpus de référence de la communication médiée par les réseaux en français. Quatorze enseignants-chercheurs de huit laboratoires différents se sont impliqués dans le projet et ont été guidés par trois mots clés lors de leurs collaborations : *variété*, *standards* et *accès ouvert*.

Le corpus CoMeRe a été construit sur une **hypothèse de variété**. Partant de l'existant, plusieurs collègues ont contribué en apportant leurs corpus au projet. Structurés dans des formats de représentation différents et, pour la grande majorité, non standardisés, ces corpus comprenaient des interactions de courriels, de clavardage, de SMS, de forums de discussion Internet, des blogs ainsi que des interactions provenant de mondes synthétiques. Parallèlement, afin d'améliorer la représentativité du corpus en y adjoignant d'autres formes de communication qui n'y étaient pas déjà incluses, deux opérations de collectes ont été lancées, notamment pour recueillir des interactions provenant de Twitter et de Wikipedia. De ce fait, la variété de la CMR a été prise en compte, augmentant l'hétérogénéité des données et la complexité du corpus. La section 2 décrit la composition du corpus CoMeRe.

D'entrée du jeu, le projet avait comme ambition que le corpus CoMeRe soit **interopérable**, à la fois pour permettre de futures analyses contrastives ou comparatives avec des données provenant d'autres corpus européens de CMR et pour poser les jalons de la construction d'un corpus de référence du français. Les genres de la CMR posant des problèmes de représentation et de traitement inédits et souvent ardus, nous avions pensé qu'il n'était pas déraisonnable de démarrer par le plus complexe. Dans cette optique, un modèle de l'espace d'interaction a été élaboré lors du projet (Chanier & Jin, 2013) ainsi que des propositions pour représenter la CMR suivant les standards de la TEI (Text Encoding Initiative, 2019) (voir Beisswenger et Lüngen dans le présent numéro qui détaille le schéma TEI CMC sous sa forme actuelle). Ces deux aspects seront détaillés dans la section 3.

Enfin, les collègues tenaient à diffuser le corpus en **accès ouvert** pour permettre son utilisation par la communauté scientifique. De ce fait, le projet a suivi les recommandations pour des données de recherche préconisées par l'initiative d'Open Data (2013) et a travaillé sur des métadonnées moissonnables pour permettre, en 2015, un dépôt sur la plateforme Ortolang (2019) dans le cadre des coopérations avec le consortium CORLI (Corpus, Langues, Interactions) de Huma-Num. Nous revenons sur les retombées du projet liées à la diffusion du corpus dans la conclusion de notre article.

2. Composition du corpus CoMeRe

La communication médiée par les réseaux est loin de former un ensemble homogène et sous cette bannière se trouvent réunis des objets et des dispositifs extrêmement différents. Cette section se propose de parcourir les différents dispositifs inclus dans le corpus CoMeRe en tentant de les organiser et de les spécifier à partir d'un ensemble de critères régulièrement convoqués dans la littérature pour les décrire.

Comme le montre le tableau 1, CoMeRe rassemble un florilège de huit dispositifs intégrant de la CMR écrite – certains dispositifs sont en effet multimodaux, proposant également de la CMR orale ou multimodale.

Tableau 1. Composition du corpus CoMeRe[1]

Dispositif	**Sous-corpus**	**Description**
Blogs	Sous-ensemble du corpus LETEC (LEarning and TEaching Corpus) Infral	1 200 messages produits par 26 apprenants francophones et germanophones sur un blog (273 546 tokens)
Wikipédia (discussions)	Wikiconflits	4 456 messages produits par 3 971 contributeurs autour de sept articles collaboratifs ayant donné lieu à des conflits d'édition (489 000 tokens)
Forums de discussion	Simuligne	2 686 messages entre une soixantaine d'apprenants de français de l'*Open University* en GB (600 348 tokens pour l'ensemble du corpus Simuligne)
Messagerie instantanée (dialogue en ligne, clavardage)	Corpus de français tchaté	5 millions de tours de clavardage produits par 53 000 participants au réseau IRC francophone Epiknet – domaines variés, de la conversation quotidienne à des interactions spécialisées, voire techniques (72 millions de tokens)
	Corpus d'apprentissage FAVI	7 780 tours de clavardage entre 31 apprenants de FLE (77 605 tokens)
	Simuligne	6 790 messages entre une soixantaine d'apprenants de français de l'*Open University* en GB (600 348 tokens pour l'ensemble du corpus Simuligne)

[1] https://hdl.handle.net/11403/comere

Messagerie Web (webmail, messagerie électronique)	Simuligne	2030 courriels entre une soixantaine d'apprenants de français de l'*Open University* en GB (600 348 tokens pour l'ensemble du corpus Simuligne)
Service de messagerie SMS (téléphonie mobile)	SMS La Réunion	12 622 SMS donnés pour la science par 884 contributeurs habitant à la Réunion (357 192 tokens)
	SMS Alpes	22 052 SMS donnés pour la science par 359 contributeurs habitant les Alpes (449 000 tokens)
	88milsms	88 522 SMS donnés pour la science par 422 contributeurs habitant majoritairement autour de Montpellier (1,2 million de tokens)
Second Life (monde synthétique)	Archi 21	669 messages 1 690 tours de parole audio 2 452 actes de communication non verbale (27 912 tokens)
Twitter	Polititweets	34 273 tweets provenant de 205 utilisateurs de Twitter, sélectionnés sur la base de critères de citation de personnalités politiques, d'affichage politique et d'activisme sur Twitter dans les *listes* politiques (567 851 tokens)
	Intermittent	11 307 tweets émanant de 215 comptes Twitter sélectionnés pour leur intérêt pour la question de l'intermittence du spectacle (180 790 tokens)
LMS (*Learning Management System*, plateforme d'apprentissage)	Copeas	5 506 tours de parole audio et 1 529 tours de clavardage entre 14 apprenants et 2 tuteurs de FLE (127 228 tokens)
	Tridem06	2 809 tours de parole audio, 248 tours de clavardage, 1058 actes de production et 779 billets de blog entre 62 apprenants d'un dispositif d'apprentissage interculturelle (184 594 tokens)

Chacun de ces huit dispositifs est un lieu de déploiement des interactions médiées par les réseaux (Internet, télécommunications). L'ensemble des interactions représentées dans CoMeRe relève d'une forme de communication à distance médiatisée par un dispositif de la CMR. Construit sur une hypothèse de variété, le corpus CoMeRe avait pour vocation de montrer la diversité de la CMR en veillant à une représentation équilibrée de ses différentes facettes (mono ou multimodalité, communications synchrones ou asynchrones, nature publique ou privée de la communication, variété des situations communicatives...). En ce sens, le corpus mis à disposition permet sans aucun doute à la communauté d'observer et d'explorer les particularités et les

variations linguistiques de cette forme de « communication écrite quasi immédiate » dont Anis et Lebrave (1986 : 126) saluaient le caractère inédit il y a plus de trente ans.

Le corpus CoMeRe rassemble ainsi des genres et des sous-corpus très différents, dont nous restituons certaines des caractéristiques principales dans le tableau 2.

Tableau 2. Caractéristiques des dispositifs CMR et des genres représentés dans le corpus CoMeRe

Dispositif	Dénomination du message	Synchronicité	Sous-corpus	Visée communicative	Visée de l'analyste
Blogs	Post (message)	Asynchrone	Infral	Apprentissage du français	Didactique
Wikipédia (discussions)	Post (message)	Asynchrone	Wikiconflits	Rédaction d'un article encyclopédique	Étude de l'expression linguistique des conflits dans Wikipédia
Forums de discussion	Post (message)	Asynchrone	Simuligne	Apprentissage du français	Didactique
Messagerie instantanée (dialogue en ligne, clavardage)	Message instantané	Synchrone	Corpus de français tchaté	Variée	Étude de la langue du clavardage
			FAVI	Apprentissage du français	Didactique
			Simuligne	Apprentissage du français	Didactique
Messagerie Web (webmail, messagerie électronique)	E-(mail), courriel (courrier électronique)	Asynchrone	Simuligne	Apprentissage du français	Didactique
Service de messagerie SMS (téléphonie mobile)	SMS, texto	Asynchrone	SMS La Réunion	Variée	Étude du langage et de l'écriture SMS, perspective sociolinguistique (contact des langues)
			SMS Alpes	Variée	Étude du langage et de l'écriture SMS, développement de traitements automatiques
			88milsms	Variée	Étude du langage et de l'écriture SMS

Second Life (monde synthétique)	Messages instantanés (clavardage), tours de parole (audio), actes de communication non verbale	Synchrone	Archi 21	Apprentissage de l'architecture et du français / l'anglais langue étrangère	Didactique
Twitter	Tweets, micro-messages	Asynchrone	Polititweets	Politique	Étude du tweet politique
			Intermittent	Politique	Étude du tweet politique
LMS	Messages instantanés (clavardage), tours de parole (audio), actes de production, post (message)	Synchrone et asynchrone	Copéas Tridem06	Apprentissage du français / Apprentissage interculturelle	Didactique

Si les genres sont variés, on peut également observer que les messages eux-mêmes ont des dénominations distinctes et parfois même multiples et concurrentes d'un dispositif à l'autre – même si le terme de *message* a une valeur hyperonymique et commute facilement avec l'ensemble des appellations. On observe une spécialisation des messages les uns par rapport aux autres au sein de la CMR, selon des critères variés comme la taille (*e.g.* le *tweet* ou le SMS sont en moyenne plus courts que le courriel) ou la simultanéité – quatre sous-corpus contiennent de la communication CMR *synchrone*, c'est-à-dire des communications simultanées qui se tiennent en temps réel : il en va ainsi des trois sous-corpus de clavardage et du sous-corpus *Archi21*, qui se compose d'interactions dans *Second Life* ; un monde synthétique qui permet à ses utilisateurs d'incarner des personnages virtuels (avatars) dans un univers créé par les résidents eux-mêmes. Les cinq autres sous-corpus comprennent de la CMR *asynchrone*, c'est-à-dire que les interactions se tiennent en différé.

Les sous-corpus à vocation didactique sont particulièrement représentés : la didactique est en effet l'une des disciplines qui s'est le plus engagée dans l'étude de la CMR avec en arrière-plan une évaluation de son intérêt pour l'apprentissage (voir Lacroix, ce numéro). Thierry Chanier a largement contribué à ces développements, comme nous l'avons souligné dans le texte d'hommage qui accompagne ce numéro, et la majorité des sous-corpus intégrés dans CoMeRe dans ce domaine ont été construits dans le cadre de projets dont il était partie prenante – e.g. le projet ANR *Mulce* (Mulce documentation, 2015) qui avait pour objectif de structurer des interactions provenant de situations didactiques partiellement ou totalement en ligne en corpus d'apprentissage et de les diffuser en accès ouvert (Mulce repository, 2013). Les principaux champs scientifiques d'analyse de la CMR sont représentés : outre la didactique, l'analyse du discours, le TAL et la sociolinguistique sont certainement les domaines linguistiques qui ont le plus investi ces nouveaux objets.

On note également que les interactions se différencient par des visées communicatives bien distinctes ; outre la visée d'apprentissage (du français pour l'essentiel mais également de l'architecture), on relève un objectif collaboratif pour les interactions des Wikipédiens autour de la rédaction des articles, mais également une visée de communication politique pour les deux sous-corpus de Tweets. Le *Corpus de français tchaté* et les trois sous-corpus de SMS contiennent pour leur part une grande variété de messages dont il serait exclu d'isoler une visée communicative dominante – caractéristique qui s'accorde avec l'objectif descriptif des chercheurs à l'origine de ces corpus, qui cherchaient à décrire la *langue* du clavardage et du SMS.

Les sous-corpus intégrés sont donc variés, tant du point de vue de leurs objectifs et de leurs principes de constitution que de celui des situations communicatives dont ils relèvent. Construit sur une hypothèse de variété, le corpus CoMeRe est particulièrement hétérogène en genres et en dispositifs de communication. Cette hétérogénéité complexifie les possibilités de comparaison et d'analyse des sous-corpus qu'il fédère, requérant une organisation et une structuration standardisées des données, qui ne pourraient être échangées ou comparées sans cette étape.

3. Standards et complexité

Nous nous sommes par conséquent rapidement trouvés face à un constat un peu paradoxal : plus un corpus est complexe et intègre des données hétérogènes – *i.e.* plus son degré de variation est élevé – et plus il est indispensable de mettre en œuvre des standards. La standardisation permet d'abord de limiter l'hétérogénéité des données en partant d'une représentation commune à l'ensemble des sous-corpus considérés. Le choix que nous avons fait d'une structuration des sous-corpus suivant les recommandations de la TEI[2] rend d'autre part possible l'inscription de CoMeRe dans la communauté internationale, en adoptant les perspectives adoptées. Enfin, ce travail de standardisation s'accompagne d'une procédure d'évaluation de la bonne formation des sous-corpus avant leur diffusion, garantie de leur qualité.

3.1 Modélisation générale de la CMR : paramétrage de la situation communicative

Première étape de cet effort de standardisation, une modélisation de la situation de communication particulière qui sous-tend l'ensemble des interactions médiées par les réseaux a été développée. Il s'agissait ainsi de proposer une représentation simplifiée mais pertinente, qui se concentrait sur les propriétés les plus décisives et les plus pertinentes pour décrire et contraster chacun des environnements en ligne : la Figure 1 illustre ce modèle de l'**espace d'interaction** (Chanier et Jin, 2013).

[2] *Text Encoding Initiative*, https://tei-c.org/

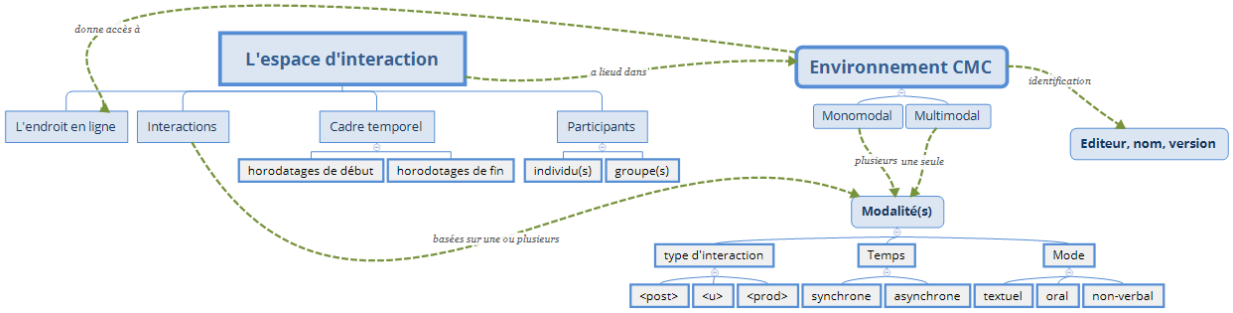

Figure 1. Modèle de l'espace d'interaction (Traduit de Chanier et al., 2014 :9)

Chanier et Jin (*ibid*) définissent l'espace d'interaction (IS) comme un concept abstrait, situé dans un cadre temporel dans lequel les interactions entre les participants (individus ou groupes) se produisent en ligne. La temporalité est saisie à partir des informations dont dispose le chercheur concernant les horodatages de début et de fin des interactions. Si les mêmes interactants communiquent sur des échelles temporelles plus longues (plusieurs semaines par exemple), on considère que plusieurs sessions d'interaction ont eu lieu. La désignation 'en ligne' est définie par les propriétés de l'ensemble des environnements utilisé dans lesquels les interactions sont transmises via des réseaux (Internet, intranet, téléphone …). Ces propriétés concernent le descriptif et les affordances de l'environnement d'interaction en ligne :

- Est-ce que l'environnement est **synchrone** ou **asynchrone** ?
- Quels sont les **modes de communication** (oral, textuel, visuel, non-verbal) à disposition des interactants ?
- Dans un mode particulier, quelles sont les **modalités de communication** disponibles ? À chaque mode de communication correspond en effet une modalité, c'est-à-dire une forme concrète particulière de communication. Un mode communicationnel peut correspondre à une variété de modalités. Par exemple, dans le cas du mode textuel il peut exister la modalité *clavardage*, la modalité *traitement de texte*, la modalité *carte conceptuelle* ou encore la modalité *tableau blanc*. À certains modes peut ne correspondre qu'une seule modalité, par exemple, le mode oral ne correspond dans les pratiques qu'à la modalité audio.
- Est-ce que l'environnement est **monomodal** ou **multimodal** ? Offre-t-il par exemple aux locuteurs la possibilité de combiner des modes ou des modalités de communication, que ce soit en associant deux modes de communication (par exemple l'emploi d'une image avec un commentaire textuel ou une production orale accompagnée d'une combinaison de gestes) ou en combinant deux modalités relevant du même mode de communication (par exemple, dans le mode *texte*, l'environnement peut offrir la possibilité de saisir du texte dans un système de traitement de texte mais également dans un outil de clavardage) ? Dans CoMeRe, nous parlerons de corpus multimodal quand les données

font intervenir plusieurs modes, ou quand plusieurs modalités sont associées à un seul mode.

- L'environnement propose-t-il un seul **espace d'interaction** ou plusieurs ? Dans certains sous-corpus, les interactants peuvent par exemple naviguer dans plusieurs salles de *chat* ; dans le cas des sous-corpus provenant des dispositifs d'apprentissage et relevant des plateformes audio-synchrones, il existe ainsi une salle de discussion « classe » et la possibilité pour l'enseignant de fractionner l'espace d'interaction pendant une période prédéterminée en différentes sous-salles « groupe ».

3.2 De la représentation générale aux particularités des dispositifs : illustration de la mise en œuvre de la standardisation TEI à travers deux exemples

L'identification de l'espace d'interaction et de ses propriétés a permis au projet CoMeRe de proposer des pistes d'extension du schéma de la TEI sur le plan des éléments et des attributs, afin de développer un nouveau standard qui permette de modéliser les interactions médiées par les réseaux (voir Beisswenger et Lüngen, ce numéro, pour une description du *CMC-core*).

Ces propositions se situent à deux niveaux : (i) au niveau des métadonnées d'abord, dans l'en-tête TEI (`<teiHeader>`) et plus précisément dans la section `<profileDesc>` pour une description des environnements utilisés ainsi que des caractéristiques des interactions et des participants ; (ii) au niveau des interactions elles-mêmes ensuite, dans le corps du texte (`<body>`).

Pour illustrer ce passage du modèle standardisé général aux dispositifs particuliers, nous prendrons deux exemples concrets provenant d'environnements CMR différents, sous-tendus par des objectifs interactionnels distincts et construits suivant des visées différentes de recherches. Le premier a trait à l'exploitation du monde synthétique *Second Life* dans le cadre d'un dispositif d'enseignement-apprentissage de l'architecture d'une part et des langues étrangères d'autre part (sous-corpus *ARCHI21*, Chanier & Wigham, 2015). Le deuxième exemple s'appuie sur un corpus de tweets provenant de comptes politiques influents (*Polititweets*, Longhi et al., 2014), développé dans le cadre d'un projet qui cherchait à mieux comprendre les tweets politiques en tant que genre de discours (voir Longhi, ce numéro). Le tableau 3 détaille les propriétés de l'espace d'interaction de chaque sous-corpus.

Tableau 3. Propriétés des espaces d'interactions pour les sous-corpus *Archi21* et *Polititweets.*

Propriétés de l'environnement	ARCHI 21	Polititweets
Synchronicité	synchrone	asynchrone

Modes de communication	verbal (oral, textuel), non verbal (proxémique, kinésique, apparence), visuel	verbal (textuel), visuel (vidéos et images), non-verbal (emojis, vidéos et images mimant le gestuel)
Type d'interaction	en bloc (textuel) en continu (oral) productions (actions, constructions)	en bloc (textuel, vidéos, images et émojis)
Mono/Multimodal	multimodal	multimodal
Espaces de communication	plusieurs (publics et privés)	publics
Cadre temporel	07/02/2011 - 11/02/2011 avec des séances journalières	2013-2014 avec la majeure partie des tweets envoyés en mars 2014
Participants	17 étudiants et 4 tuteurs divisés en 4 groupes de travail Utilisation des alias	comptes politiques. Il peut s'agir d'individus, de partis, d'influenceurs, ou de comptes satiriques.
Type de discours	interactions pédagogiques en L2, parole privée	discours politique, parole publique

Configuration de l'en-tête TEI

L'en-tête du document TEI permet de regrouper des métadonnées concernant les environnements utilisés, les caractéristiques des interactions et les participants.

Les environnements de CMR sont décrits dans la section `<taxonomy>` avec un identifiant ainsi qu'un descriptif (`<catDesc>`) de l'environnement. Ensuite sont détaillés les modes de communication possibles de l'environnement ainsi que leurs spécificités.

Si nous comparons les deux sous-corpus (Tableau 4), leurs en-têtes décrivent que pour le sous-corpus *ARCHI21* la balise `<post>` fait référence au tour de parole dans le clavardage tandis que pour le sous-corpus *Polititweets*, elle renvoie au tweet. Dans le corpus provenant du monde synthétique, nous avons également des modes oraux représentés dans le corpus par `<u>` (*utterance*), correspondant à des tours de parole audio, ainsi que des modes qui relèvent du non verbal encodés par la balise `<prod>`. Cette balise décrit les actions, les gestes ou l'apparence d'un participant à travers son avatar. Les spécificités de la syntaxe des messages sur Twitter entraînent pour leur part une utilisation particulière des attributs de la TEI, qui est également documentée, *ex.* les balises et les attributs utilisés pour annoter les termes d'adresse, les informations concernant le nombre de fois qu'un message a été retweeté ou l'utilisation des hashtags.

Le `<TEI Header>` permet ainsi de décrire comment chaque balise, élément et attribut de la TEI a été employé.

Tableau 4. Description des environnements dans le TEI Header

ARCHI21

```
<namespace name="TEI-CMC">
   <tagUsage gi="post">one post corresponds to one texchat turn within Second Life<list>
         <item><att>xml:id</att>ID of the post.</item>
         <item>
            <att>synch</att>date of the message when created, given by the system see
               <gi>timeline</gi></item>
         <item>
            <att>who</att>id of the author of the message. Every participant have been
            described in the <gi>particDesc</gi>.</item>
         <item>
            <att>type</att>type of the post cf. taxononomy. </item>
      </list></tagUsage>
   <tagUsage gi="u">One audio turn within Second Life (for transcription code see <ref
         target="#slmanual"/>). This coding scheme has been converted to TEI speech (see
         <ref target="#umanual"/>).</tagUsage>
   <tagUsage gi="prod">participants multimodal action <list>
         <item><att>type</att> of action</item>
         <item><att>subtype</att></item>
      </list>cf. taxononomy.</tagUsage>
   <tagUsage gi="code">Code created by Mulce when transcribing nonverbal action. see
         <ref target="#slmanual"/></tagUsage>
</namespace>
```

Polititweets

```
<namespace
    name="http://wiki.tei-c.org/index.php/SIG:CMC/Draft:_A_basic_schema_for_representing_CMC_in_TEI">
    <tagUsage gi="text">each text correspond to the set of tweets coming from the same Twitter
        account</tagUsage>
    <tagUsage gi="post">one post corresponds to one tweet.<list>
        <item><att>xml:id</att>ID of the posting.</item>
        <item>
            <att>when</att>is date of message on Twitter.</item>
        <item>
            <att>who</att>ID of the twitter account, see <gi>listPerson</gi> .</item>
       </list></tagUsage>
    <tagUsage gi="p">This element appears inside the <gi>post</gi>. Encode the message
        contents. </tagUsage>
    <tagUsage gi="distinct">This element appears inside <gi>p</gi>. It describes Twitter
        syntax in the following way. <att>type</att>
        <list>
            <item><val>twitter-hashtag</val>. Then the element contains <gi>ident</gi> with
                <val>#</val>, and <gi>ref</gi> with the URL of discussion topic</item>
        </list></tagUsage>
    <tagUsage gi="addressingTerm">Addressing terms address an utterance to a particular
        interlocutor / twitto or refers to a twitto. It includes : <list>
            <item><gi>addressMarker</gi> with <val>@</val></item>
            <item><gi>addressee</gi> refers to a Twitter account</item>
        </list></tagUsage>
</namespace>
```

La section <textDesc> est pour sa part utilisée pour documenter les interactions dans chaque environnement avec un jeu de métadonnées adaptées. Sont par exemple renseignés dans cette rubrique le niveau de spontanéité des interactions (<preparedness type>, leur objectif communicatif (<purpose>) ou encore l'orientation de la communication vers un seul ou vers plusieurs destinataires (<interaction type> <active>). Le tableau 5 en donne un exemple.

Tableau 5. Description des environnements dans le TEI Header

ARCHI21

```
<textDesc xml:lang="en-GB">
   <channel mode="m" xml:lang="en-GB">
      <term ref="#SecondLife">Second Life, Synthetic environment, 3D world with
         avatars</term>
   </channel>
   <constitution>This corpus is made of interactions between participants (learners,
      tutors) during the online Archi21 experiment (2011) focusing on second language
      learning through a content and language integrated approach (CLIL) with architectural
      and design education. All these interactions happened within the Second Life
      environment. In Second Life, they are made of verbal actions (textchat turns, audio
      turn), and non-verbal ones (actions on objects, or related to avatars). Participants
      were organized in small groups. All details about groups are in <gi>particDesc</gi>. </constitution>
   <derivation type="original"/>
   <domain>education</domain>
   <factuality type="fact"/>
   <interaction type="complete" active="many">
      <note>Interactions happened acordingly to the guidelines of the learning activities
         (see <gi>projectDesc</gi> for access to guidelines) </note>
   </interaction>
   <preparedness type="spontaneous"/>
   <purpose degree="high">English and French for architectural design (CLIL)</purpose>
</textDesc>
```

Polititweets

```
<textDesc xml:lang="en-GB">
  <channel mode="w" xml:lang="en-GB"><term ref="#tweet">Message sent through a Twitter
      account</term></channel>
  <constitution>Selected through automatic processing. See <gi>projectDesc</gi> for more
    information</constitution>
  <derivation type="original"/>
  <domain type="public"><note>domain of a message: politics</note></domain>
  <factuality type="fact"/>
  <interaction type="complete" active="many"/>
  <preparedness type="spontaneous"/>
  <purpose>political local elections</purpose>
</textDesc>
```

Les métadonnées concernant les participants sont décrites dans la rubrique <person> (voir Tableau 6). Pour les deux sous-corpus, un <id> permet de faire figurer le nom ou le code attribué à l'interactant par les chercheurs lors du recueil des données. Pour le sous-corpus *Polititweets*, le pseudonyme ou le nom de l'interactant figure également sous <name>, les données provenant de comptes Twitter politiques. Pour *ARCHI21*, afin de préserver l'anonymat des participants suivant le protocole de recherche, la balise <addName type="alias"> est employée pour faire figurer les autres noms utilisés dans les interactions qui faisaient référence

au participant ou à son avatar, tandis que pour *Polititweets* le *screenname*, e.g. le nom du compte Twitter de l'interactant est renseigné `<addName type="ScreenName">`.

Tableau 6. Description des participants dans le TEI Header.

ARCHI21	`<person role="learner" xml:id="antoinobri">`
	`  <sex>male</sex>`
	`  <age value="25"/>`
	`  <residence>fra</residence>`
	`  <affiliation>`
	`  <orgName>École Nationale Supérieure d'Architecture Paris-Malaquais</orgName>`
	`  </affiliation>`
	`  <persName>`
	`  <addName type="alias">Antoine</addName>`
	`  </persName>`
	`  <langKnowledge>`
	`  <langKnown tag="fra">`
	`  <ident type="L1">First language or language used every day</ident>`
	`  </langKnown>`
	`  </langKnowledge>`
	`  <langKnowledge>`
	`  <langKnown tag="eng">`
	`  <ident type="L2">Second language (first foreign)</ident>`
	`  </langKnown>`
	`  </langKnowledge>`
	`  <langKnowledge>`
	`  <langKnown tag="">`
	`  <ident type="L3">Third language (second foreign)</ident>`
	`  </langKnown>`
	`  </langKnowledge>`
	`</person>`

Polititweets	`<person xml:id="cmr-politweets-p18814998">`
	`  <persName>`
	`  <name>François Hollande</name>`
	`  <addName type="ScreenName" ref="https://twitter.com/fhollande">fhollande</addName`
	`  </persName>`
	`</person>`

Le sous-corpus *ARCHI21* provenant du domaine de la didactique, plusieurs métadonnées d'ordre sociolinguistique décrivent les participants : leur sexe et leur âge, leur pays d'origine (`<residence>`) pour permettre aux chercheurs de distinguer les étudiants français des étudiants ERASMUS inscrits dans le dispositif, et de disposer d'une information concernant leur langue(s) maternelle(s) ainsi que les langues étrangères qu'ils pratiquent (`<langKnowledge>`) : celles-ci pourront en effet avoir une influence sur l'apprentissage de la langue cible lors de la formation en langue étrangère. Le `<role>` du participant en tant qu'apprenant ou tuteur est également stipulé ainsi que l'institution dans laquelle la personne est inscrite (`<affiliation>`).

Structuration des interactions dans le corps du texte

Le `<TEI body>` permet de structurer les interactions elles-mêmes. Pour les sous-corpus CoMeRe, la balise `<div>` délimite une interaction. Selon l'environnement, nous avons distingué plusieurs types de divisions :

- `div type="thread"` pour structurer des interactions provenant des forums ou des blogs, chaque interaction étant décrite avec l'élément enfant `<post>` ;
- `div type="logfile"` pour les interactions de clavardage et de SMS avec l'élément enfant `<post>` ;
- `div type="oral-discourse"` pour des interactions audio avec l'élément enfant `<u>` ;
- `div type="multi-modalities"` pour les interactions multimodales, qui contiennent des `<post>` pour les interactions écrites, `<u>` pour les interactions orales et `<prod>` pour les interactions dans le mode non verbal.

Figure 2. Capture d'écran du moment de l'interaction dans le monde synthétique

La Figure 3 offre un exemple de structuration dans le `<body>` des interactions multimodales synchrones du sous-corpus *ARCHI21*. Trois participants sont en train d'interagir dans le monde synthétique : le premier intervient *via* l'audio, le second à travers les actes de son avatar et le troisième dans l'espace de clavardage. Nous comprenons qu'il s'agit d'une interaction entre trois participants et non pas d'un seul participant interagissant sous des alias différents, car les identifiants uniques sont distincts. Le participant `cmr-archi21-slrefl-es-j3-1-a191`, sous son nom alias (`@who`) `tingrabu` prend la parole par le mode audio. Son tour de parole est décrit dans la balise `<u>` et les informations concernant le temps de début et de fin de son tour sont encodées. Lors de son tour de parole, l'avatar d'un autre participant, `id="cmr-archi21-slrefl-es-j3-1-a192"`, sous son alias `romeorez`, fait l'acte de manger du popcorn décrit avec l'élément enfant `<prod>` (mode non verbal). Nous pouvons voir, grâce aux balises temporelles que cette interaction non verbale se chevauche avec le tour de parole de `tingrabu`. Juste après, toujours en même temps que `tingrabu` parle dans le mode audio, le participant

`cmr-archi21-slrefl-es-j3-1-a195` sous son alias `tfrez2` envoie un message dans le clavardage, décrit dans l'élément enfant `<post>`.

```
<u xml:id="cmr-archi21-slrefl-es-j3-1-a191" who="#tingrabu"
    start="#cmr-archi21-slrefl-es-j3-1-ts373"
    end="#cmr-archi21-slrefl-es-j3-1-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"/> ...
</u>
<prod xml:id="cmr-archi21-slrefl-es-j3-1-a192" who="#romeorez"
    start="#cmr-archi21-slrefl-es-j3-1-ts376"
    end="#cmr-archi21-slrefl-es-j3-1-ts377" type="body" subtype="kinesics">
    <code>eat(popcorn)</code>
</prod>
<prod xml:id="cmr-archi21-slrefl-es-j3-1-a193" who="#tfrez2"
    start="#cmr-archi21-slrefl-es-j3-1-ts378"
    end="#cmr-archi21-slrefl-es-j3-1-ts386" type="body" subtype="kinesics">
    <code>type</code>
</prod>
<post xml:id="cmr-archi21-slrefl-es-j3-1-a195" who="#tfrez2"
    start="#cmr-archi21-slrefl-es-j3-1-ts380"
    end="#cmr-archi21-slrefl-es-j3-1-ts381" type="chat-message">
    <p>it went too quickly?</p>
</post>
```

Figure 3. Structuration d'une interaction multimodale en TEI.

Le sous-corpus *Polititweets* offre d'autres exemples d'interactions asynchrones. L'élément enfant `<post>` est employé pour les interactions écrites. Comme pour le sous-corpus *ARCHI21*, l'interactant est identifié au moyen d'un identifiant unique. Vu qu'un tweet consiste en un message envoyé en un bloc, il n'y a pas d'empan temporel : la date et l'heure d'envoi sont décrites dans l'attribut `when`. Les attributs TEI sont employés pour décrire l'utilisation des termes d'adresse `<addressTerm>`, le nombre de fois que le *tweet* a été republié, ou *retweeté* (`<f name="retweetcount">, <numeric value="60">`) ou encore l'emploi de *hashtags* (`<distinct type="twitter-hashtag">`).

```
<post xml:id="cmr-politweets-a382908783045128192" who="#cmr-politweets-p17211968"
  when="2013-09-25T18:45:42" xml:lang="fra">
  <p>Je suis aujourd'hui convaincu que François <distinct type="twitter-hashtag"
        ><ident>#</ident><rs ref="https://twitter.com/search?q=%23Hollande&amp;src=hash"
        >Hollande</rs></distinct> a renoncé à penser les réformes nécessaires. <ref
      target="http://t.co/6mE5vIThxD http://mouvementdemocrate.fr/article/francois-
      hollande-a-renonce-a-penser-les-reformes-necessaires"
      >http://t.co/6mE5vIThxD</ref></p>
  <trailer>
    <fs>
      <f name="medium">
        <string>TweetDeck</string>
      </f>
      <f name="favoritecount">
        <numeric value="8"/>
      </f>
      <f name="retweetcount">
        <numeric value="60"/>
      </f>
    </fs>
  </trailer>
</post>
```

Figure 4. Exemple de tweet du sous-corpus *Polititweets*

Ces exemples d'interactions de deux types de sous-corpus de la CMR nous ont permis d'illustrer la mise en œuvre d'un standard malgré la variation des données. Les deux exemples considérés relèvent en effet de différentes temporalités (synchrone/asychrone), de différentes configurations multimodales (texte/ image et vidéo ; audio/clavardage/non verbal) selon les affordances des plateformes, et d'interactions CMR issues de différents domaines (l'éducation/la politique). Un fichier décrivant la structure TEI complète des sous-corpus CoMeRe est d'ailleurs disponible[3].

Nous avons ainsi tenté de restituer les efforts consentis par les membres du projet CoMeRe pour développer une modélisation générale et standardisée de la communication médiée par ordinateur, applicable à l'ensemble des sous-corpus et des genres représentés. La section qui suit propose un bilan de l'exploitation de CoMeRe, et de la complexité des traitements de ce type de données.

4. CoMeRe en perspective

Diffusé sur Ortolang (CoMeRe Repository, 2014) selon les termes de la licence Creative Commons Attribution 4.0 International, le corpus CoMeRe regroupe actuellement quatorze sous-corpus structurés en CMC-TEI et totalise 75M tokens. CoMeRe présente sans aucun doute un intérêt remarquable pour observer la CMR.

Différents développements se sont poursuivis depuis la mise en ligne de CoMeRe : l'un des sous-corpus, *FAVI* (Yun & Chanier, 2014), qui est un corpus d'apprentissage du français langue

[3] https://hdl.handle.net/11403/comere/tei_cmr.rng

étrangère structurant des échanges par clavardage, a été étiqueté en morphosyntaxe (Riou & Sagot, 2016), l'annotation de surface de l'ensemble du corpus restant encore à réaliser. Sur un autre plan, le schéma TEI proposé par le projet CoMeRe, que nous venons de décrire dans la section précédente, a été retravaillé par des chercheurs impliqués dans le groupe d'intérêt TEI autour de la CMR dans l'objectif de déposer une demande d'enrichissement de la TEI (*TEI feature request*) pour l'annotation standardisée de la CMR. Il s'agit du résultat d'une collaboration franco-allemande actuellement en cours (voir Beisswenger et Lüngen, ce numéro).[4]

Élaboré dans un cadre européen grâce à l'action de Thierry Chanier qui a été le moteur de cette mise en réseau, CoMeRe reste un projet régulièrement cité comme modèle par les consortiums de linguistique de la TGIR Huma-Num (IRCE et IRCOM jusqu'à 2015, et CORLI depuis 2016). Huma-Num, et donc CORLI, s'engagent dans la voie du *FAIR data* (*Findable*, *Accessible*, *Interoperable*, *Reusable*) que CoMeRe (et Chanier) avait déjà empruntée avant l'heure.

Différentes collaborations initiées dans le cadre du projet CoMeRe sont en outre toujours actives, comme par exemple les travaux sur les discussions Wikipédia qui vont donner lieu à la parution d'un volume dans la collection *Studies in Corpus Linguistics* chez John Benjamins en 2020, dans le cadre de la coopération toujours productive entre BCL (Nice), CLLE-ERSS (Toulouse), l'institut de la langue allemande (IDS, Mannheim) et l'Université de Mannheim.

Le dialogue avec l'infrastructure CLARIN (Common Language Resources and Technology Infrastructure, 2019) reste actif. Son objectif est de créer et maintenir une infrastructure pour soutenir le partage, l'utilisation et la conservation pérenne des données linguistiques et des outils de recherche en sciences humaines et sociales. CLARIN travaille également à augmenter la visibilité des corpus CMR, par exemple, en proposant des "familles de ressources" – des sites web qui regroupent des informations concernant des corpus structurés et en accès ouvert provenant d'un domaine spécifique (https://www.clarin.eu/resource-families/cmc-corpora) et, par le biais de la formation, en organisant des ateliers pratiques sur l'annotation des données CMR en TEI.

Références

Anis, J. & Lebrave, J-L. (1986). Des textes interactifs ?, *Linx*, 14(19) : 107-131, [https://doi.org/10.3406/linx.1986.1041].

Beißwenger, M., Ermakova, M., Geyken, A;, Lemnitzer, L. & Storrer, A. (2013). « A German reference corpus of computer-mediated communication », *Literary and Linguistic Computing*, 28(4) : 531–537, [https://doi.org/10.1093/llc/fqt038].

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

[4]Voir également https://wiki.tei-c.org/index.php?title=SIG:CMC/CMC-core_schema_for_representing_CMC_in_TEI_(2019)

genres », in Beißwenger, M., Oostdijk, N., Storrer, A & van den Heuvel, H. Building and Annotating Corpora of Computer-Mediated Discourse : Issues and Challenges at the Interface of Corpus and Computational Linguistics, *Journal of Language Technology and Computational Linguistics* (special issue), 29(2) : 1-30. [http://halshs.archives-ouvertes.fr/halshs-00953507].

Chanier, T. & Jin, K. (2013) *Defining the online interaction space and the TEI structure for CoMeRe corpora.* Projet CoMeRe (Communication Médiée par les Réseaux), IR Corpus-écrits. [http://corpuscomere.org/, comere-is-tei-v2].

Chanier, T. & Wigham, C.R. (2015). « Archi21 corpus: collaborative language and architectural learning in Second Life ». In Chanier T. (ed) *Banque de corpus CoMeRe*. Ortolang.fr : Nancy. [https://hdl.handle.net/11403/comere/cmr-archi21].

CLARIN (2019). « Common Language Resources and Technology Infrastructure ». [website] [https://www.clarin.eu/].

CoMeRe Repository (2014). « CoMeRe corpora ». [website] [https://hdl.handle.net/11403/comere].

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

Longhi, J., Marinica, C., Borzic, B. & Alkhouli, A. (2014). « Polititweets, corpus de tweets provenant de comptes politiques influents ». In Chanier T. (ed) *Banque de corpus CoMeRe*. Ortolang.fr : Nancy. [https://hdl.handle.net/11403/comere/cmr-polititweets].

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

Mulce documentation (2015). « Documentation on Mulce repository and Mulce methodology ». [website]. [http://mulce.org].

Oostdik, N., Reynaert, M., Hoste, V. & Schuurmann, I. (2013). « The Construction of a 500-Million Word Reference Corpus of Contemporary Written Dutch ». In Spyns, P., & Odijk, J. (ed.), *Essential Speech and Language Technology for Dutch*. Berlin: Springer : 219-247.

OpenData (2013) « Principles for "openness" in relation to data and content » .[Document]. *Open Knowledge Foundation*. [http://opendefinition.org/od/].

ORTOLANG (2019). « Open Resources and TOols for LANGuage ». [website]. ATIFL / CNRS - Université de Lorraine : Nancy. [http://ortolang.fr]

Riou, S. & Sagot, B. (2016). « Etiquetage morpho-syntaxique du corpus FAVI ». D'après Yun, H. & Chanier, T. (2014). Corpus d'apprentissage FAVI (Français académique virtuel international) [cmr-favi-tei-v1]. In Chanier T. (ed) *Banque de corpus CoMeRe.* Ortolang.fr : Nancy. [https://hdl.handle.net/11403/comere/cmr-favi/cmr-favi-tei-v2]

Text Encoding Initiative (2019) « Text Encoding Initiative Consortium c [website] [https://tei-c.org/].

Yun, H. & Chanier, T. (2014). « Corpus d'apprentissage FAVI (Français académique virtuel international) ». In Chanier T. (ed) *Banque de corpus CoMeRe*. Ortolang.fr : Nancy. [https://hdl.handle.net/11403/comere/cmr-favi/cmr-favi-tei-v1].