



CORLI : diffuser, exploiter, et partager  
les corpus et les outils de linguistique  
de l'écrit et de l'oral

*C*ORpus  
*L*angues  
*I*nteractions



Carole Etienne, Juin 2018



## CORLI : CORpus, Langues et Interactions

- Fusion “Corpus Écrits” et “IRCOM : Corpus Oraux et Multimodaux”
- Porteur Fédération ILF-FR 2393 du CNRS (Institut de Linguistique Française)
- Comité de pilotage d'une vingtaine de personnes
- Groupes projet
  - Inter-Explo : Interopérabilité /Pratique et outils d'exploration de corpus
  - MULTICOM : Multimodalité et Nouvelles formes de communication
  - Corpus multilingues et plurilingues



## CORLI : CORpus, Langues et Interactions

### ■ Missions

- Valorisation, réutilisation, visibilité et accessibilité des ressources existantes
- Mise à disposition des ressources, mutualisation et interopérabilité afin d'intégrer les réseaux internationaux
- Partage des bonnes pratiques et diffusion des standards européens et internationaux
- Critères d'évaluation des corpus dans le cadre général de l'évaluation de la production scientifique des unités de recherche

### ■ Actions

- Organisation de journées d'études
- Organisation de formations
- Participation financière à des colloques
- Aide à la finalisations de projets (13 projets en 2017)
- Diffusion de nos réalisations dans les colloques : Floral, JLC, journées TEI, Clarin, ...
- Concertation avec l'équipex ORTOLANG : réalisations d'outils



## Cinq réalisations pour diffuser, exploiter, et partager corpus et outils

- Des métadonnées orientées recherche pour faciliter la réutilisation des corpus oraux et multimodaux
- Un format pivot pour les transcriptions de l'oral
- Un ouvrage Explorer un corpus textuel
- Un groupe de recherche CMC Corpora of Computer-Mediated
- Les réflexions sur les corpus multilingues et plurilingues



teimeta : des métadonnées pour  
faciliter la réutilisation des données



## teimeta : un jeu de métadonnées orienté recherche

- Table ronde sur la diffusion et la réutilisation des corpus
  - La diffusion
    - un jeu de métadonnées commun
    - la citation obligatoire de la ressource
    - une portée européenne, internationale
  - La réutilisation
    - anonymisation : où trouver l'information
    - signal : problèmes de qualité, temps de téléchargement, formats
    - transcriptions disponibles dans un format lié à un logiciel de transcription
    - pas d'indications claires sur la personne à contacter si les données ne sont pas accessibles



## teimeta : un jeu de métadonnées orienté recherche

- Un niveau minimal commun de métadonnées à l'oral pour
  - comprendre les données
  - sélectionner certaines de ces données
  - aider à l'analyse des résultats de recherche
  - enrichir par de nouvelles annotations
  
- Un format commun
  - évolutif
  - à granularité variable
  - diffusable dans un format standard
  - portée européenne et internationale
  - utilisable aussi pour les corpus écrits



## teimeta : un jeu de métadonnées orienté recherche

- Réutilisation des données : les verrous ?
  - Corpus déjà connus ... oui mais à un instant donné
    - volume de données : alimentation ?
    - audio vs vidéo dans les enregistrements plus récents
    - qualité hétérogène
    - nature des enregistrements
    - langue
    - annotations
    - politique d'accès
  - Corpus récents
    - peu connus





## teimeta : un jeu de métadonnées orienté recherche

- Constat : de plus en plus de projets de recherche concernent des corpus existants et peuvent impliquer plusieurs sources de données
- De plus en plus de projets impliquent des corpus oraux et des corpus écrits (écrits non planifiés)
- En début de projet, au moins un "Work Package" dédié à la mise en commun de données ... pourtant déjà décrites et annotées
- En fin de projet, de nouvelles annotations délivrées dans différents formats avec différents outils → comment l'indiquer dans les sources



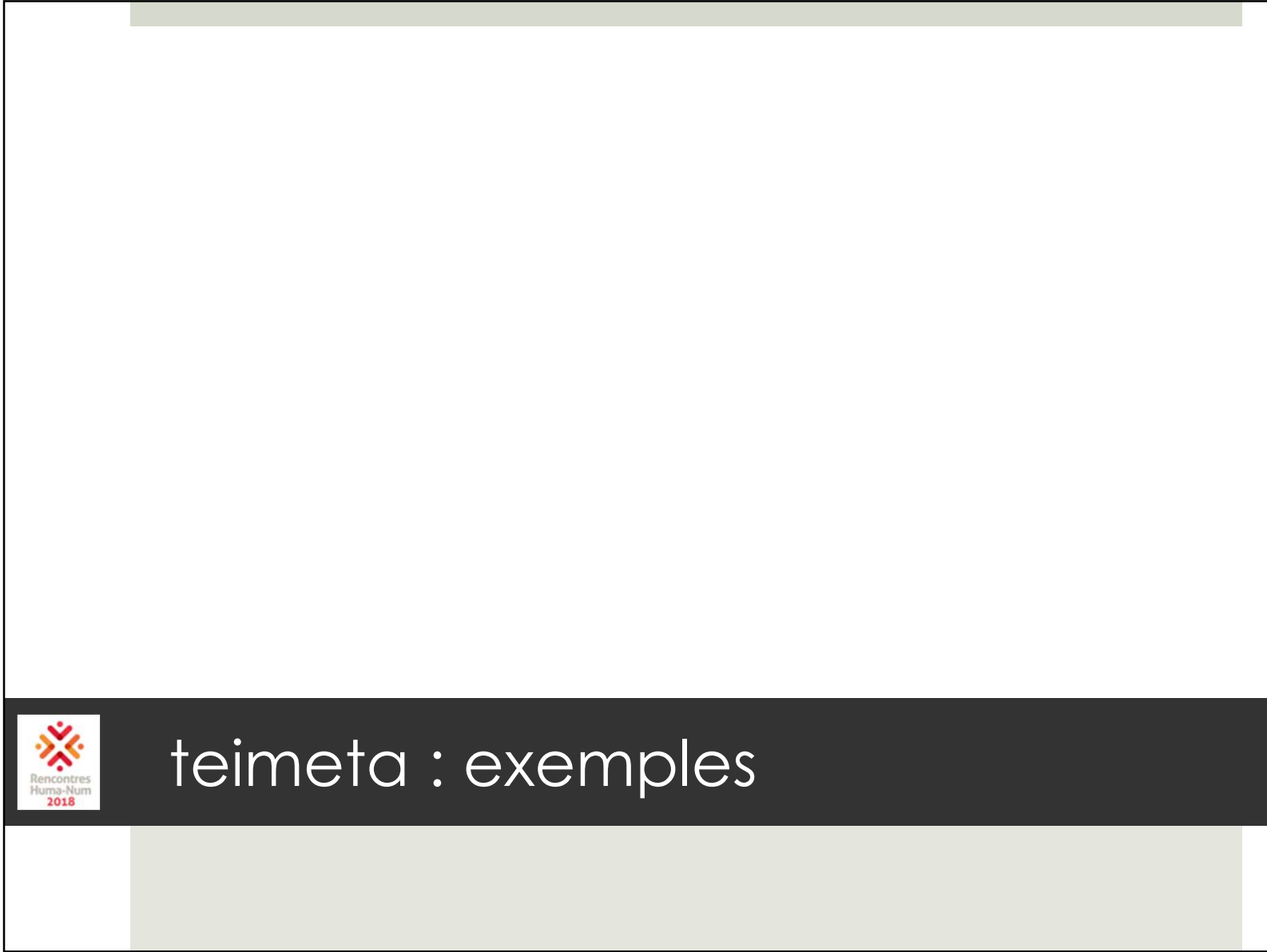
## teimeta : un jeu de métadonnées orienté recherche

- Analyse de l'existant notamment avec le projet ORFEO
  - Très hétérogènes tant au niveau du format
    - Fichier texte (pdf, word)
    - **Fichiers tabulaires (excel, csv)**
    - XML (OLAC, TEI Header, CMDI)
  - ... que du **contenu**
    - champs basiques : durée, âge, lieu , nom, ..
    - métadonnées absentes ?
    - critères subjectifs :
      - qualité, niveau de langue du locuteur : bon/moyen/mauvais
      - niveau de spontanéité
      - contenu : nature, catégorie, domaine




## teimeta : un jeu de métadonnées orienté recherche

- Le niveau commun "niveau 0"
  - Informations générales sur le corpus: **citation, diffusion**, version,...
  - Informations sur les données primaires: enregistrement, **anonymisation, contexte**
  - Informations sur les données secondaires: transcription, **anonymisation, annotations**
  - Informations sur les locuteurs : **natif/non natif, adulte/enfant, nombre**, âge, profil sociolinguistique...
- Vocabulaire contrôlé : sélection
- Multilingue
- Outil en ligne en concertation avec **ORTOLANG** : teimeta
- Format de diffusion : **TEI** → Personnalisation **ODD**
- Formation et diffusion Floral, JLC, PFC 2017, CMLF2018, TEI 2016 et 2018



teimeta : exemples





# Format pivot pour les transcriptions de l'oral

Edition de métadonnées TEI / CORLI ★☆☆ - 0.5.3

[Ouvrir](#) | [Sauver](#) | [Nouveau ↓](#) | [ODD prédéfinis ↓](#) | [Paramètres](#) | [Langues ↓](#) | [? Aide](#)

ODD: TEI Oral

CSS:

Fichier: new-file.xml

*Titre, citation, diffusion et données primaires – signal audio/vidéo*

Titre, description, citation ★ Titre de la ressource, nom d'usage  ★ Description courte   
 + Citation(s) accompagnant la ressource: projet, équipe de recherche, référence bibliographique  ←

*La ressource: titre, description, citation, responsable, contributeurs*

+ Responsable de la ressource: organisme, laboratoire, projet, personne  ★ Nom  projet

+ Contributeurs

+ Rôle, fonction  annotateur ★ Nom  projet

*Diffusion : identifiant unique, diffusion, sites web, licence*

+ ★ Projet, archive diffusant la ressource  ←

+ ★ Diffusion dans d'autres sites

+ ★ URL  Lien vers la ressource (URL)

+ ★ Identifiant unique de la ressource  handle

+ Conditions de diffusion

+ ★ License de diffusion  Creative Common CC\_BY\_NC\_SA : Attribution sans usage commercial et partage suivant les mêmes modalités ↓



# teimeta : un jeu de métadonnées orienté recherche

Données primaires (signal)

+ Session, Enregistrement : une même session peut correspondre plusieurs fichiers (qualité et/ou format différents)

+ Description courte

★ Média : chaque média peut avoir un type et une durée différente audio/vidéo signal vidéo format format mp4 durée du média Format: 00:00 ou 00:00:00 00:00 url du média

★ Qualité moins de 5% de bruit Anonymisation anonymisation partielle

Catégorie, participants, contexte, langue

Type de données orales

★ Canal de l'interaction : radio/tv/téléphone/présence vidéo tous les locuteurs sont présents Ressource ressource autonome

★ original ou adaptation d'une autre ressource (résumé, traduction...) ressource d'origine

★ Domaine privé ou professionnel professionnel - au moins un locuteur est en situation professionnelle Genre interactionnel réunion

★ Nombre de locuteurs : au total, actifs et passifs

★ Consignes, instructions spontané Contexte commercial

Session enregistrée : lieu, langue, date

Lieu

★ Ville

★ Région

★ Description courte

★ Lieu Pays

★ Session enregistrée : description, date, environnement, langue

★ Intervalle (depuis/jusqu'à) ou date exacte Depuis jj / mm / aaaa Jusqu'à jj / mm / aaaa Date exacte jj / mm / aaaa

★ Description de la session enregistrée

Langue(s) parlée dans la session (si plusieurs pourcentage d'utilisation)

★ Langue(s) de la situation, si plusieurs pourcentage d'utilisation



# teimeta : un jeu de métadonnées orienté recherche

**Locuteurs**

+ **Locuteur** Identifiant unique du locuteur  Sexe masculin  URL Locuteur  Rôle père

+ **Langue des locuteurs** Code Iso  maternelle

+ **Information supplémentaire**  Type

+ **Nom du locuteur**  **Tranche d'âge** de 21 à 60 ans  **Pseudonyme (dans la transcription)** Type

+ **Situation** en activite  **Indice socio-économique**  **Niveau de scolarité**

**Logiciel, projet, annotations**

+ **Annotations : type d'annotation, logiciel d'annotation, convention, anonymisation**

+ **Annotations : type d'annotation, logiciel d'annotation, convention, anonymisation** anonymisation de la transcription  transcription anonymisée

+ **Annotations : type d'annotation, logiciel d'annotation, convention, anonymisation** nature(s) des annotations  prosodique  Description étendue du projet

+ **Annotations : type d'annotation, logiciel d'annotation, convention, anonymisation** nature(s) des annotations  orthographique  Description courte

+ **Annotations : type d'annotation, logiciel d'annotation, convention, anonymisation** nature(s) des annotations  syntaxique

**Convertisseur TEI : TEI\_CORPO ou un autre convertisseur**

+ **Convertisseur TEI**  convertisseur teicorpo d'Ortolang  Description courte

15



# teicorpo : Format pivot pour les transcriptions de l'oral





## Format pivot pour les transcriptions de l'oral

- Table ronde sur la diffusion et la réutilisation des corpus
  - **verrou : transcriptions disponibles dans un format lié à un logiciel de transcription**
  - besoin : un format commun pour les données, diffusable dans un standard, portée européenne/internationale
  
- Méthode
  - analyse des pratiques
  - choix du format de diffusion : TEI
  - contribution au groupe européen ISO-TEI
  - ouverture vers les logiciels : Lexico, Trameur, Hyperbase, Txm
  - intégration des annotations PoS de Treetagger dans Praat, Elan



# Format pivot pour les transcriptions de l'oral



IRCOM



ORTOLANG



TEI

## Conversions au format TEI pour l'Oral et le Multimodal

### 1) Choisir le Format Destination

- TEI (xml / tei\_corpo.xml / teiml / trjs)
- TRS (transcriber)
- CHA (chat - childes)
- TXT (texte - utf8)
- DOCX (microsoft word)
- XLSX (microsoft excel)
- CSV (tableurs)
- TEXTGRID (praat)
- EAF (elan)
- TXM (xml/w)
- Lexico/Le Trameur (.txt)

- Conserver ces locuteurs/champs dans la sortie
  - Supprimer ces locuteurs/champs de la sortie
- Valeur du locuteur ou du champ (caractères génériques acceptés)
- Supprimer les marqueurs spécifiques de l'oral

### 2) Choisir le Fichier source (extension: TRS/CHA/TEXTGRID/EAF/TXT/DOCX/XLSX)

Faire glisser ici un (ou plusieurs) fichier(s)

Ou cliquer ici pour sélectionner un fichier =>  Aucun fichier sélectionné.

Demander les paramètres pour les fichiers praat.

Résultats (Effacer)

Le format TEI\_CORPO suit les propositions du GT2 IRCOM et du groupe TEI Oral ISO. Il est conforme au standard TEI.



# Ouvrage Explorer un corpus textuel



## Ouvrage Explorer un corpus textuel

Auteurs : Céline Poudat, Frédéric Landragin

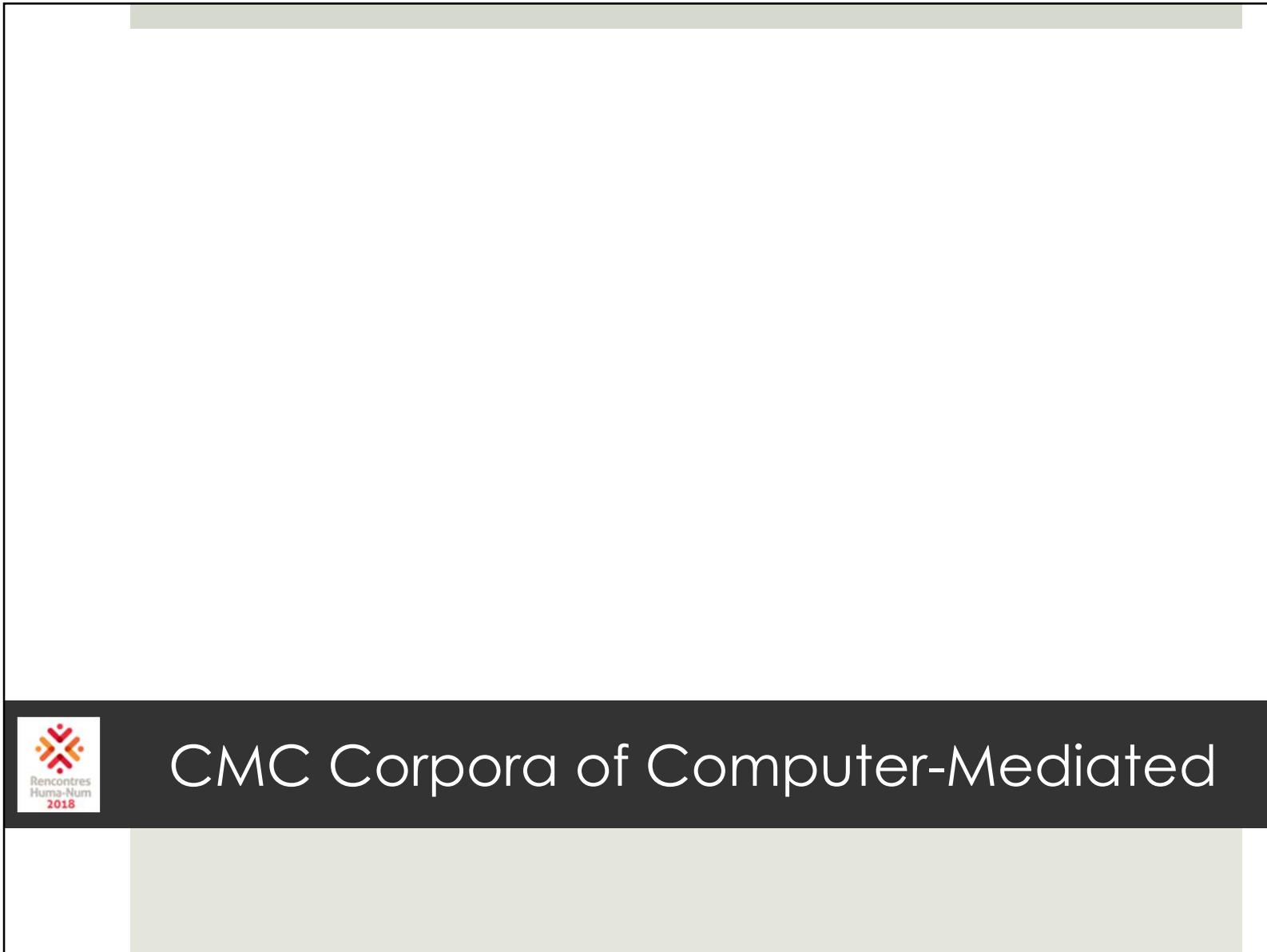
- Synthèse des méthodes et des pratiques d'exploration de corpus
  - Ouvrages existants datés et souvent trop techniques
  - Réunir **annotation et ADT** dans un même ouvrage
  
- Visée : **diffusion des bonnes pratiques**
  
- Propos **méthodologique** plutôt que technique
  - Explorer la structure d'un corpus avec l'analyse factorielle : AFC, ACP
  - Explorer la structure d'un corpus avec une classification : ascendante CAH, descendante CDH ou une analyse arborée




## Ouvrage Explorer un corpus textuel

Auteurs : Céline Poudat, Frédéric Landragin

- Souci **pédagogique**
  
- Cible:
  - Étudiants en sciences du langage, de niveau Master
  - **Tout** étudiant, **tout** chercheur intéressé par la manipulation et l'exploitation de données textuelles
  
- 8 logiciels: AntConc, Dtm-Vic, Hyperbase, IRaMuTeQ, Le Trameur, Lexico, TXM, Unitex
  
- 18 encadrés restituant des **exemples de recherches** de nos collègues mobilisant les méthodes présentées



  
Rencontres  
Huma-Num  
2018

CMC Corpora of Computer-Mediated



## CMC Corpora of Computer-Mediated communication médiée par les réseaux

- Groupe de recherche européen CMC-corpora
  - CoMeRe : communication médiée par les réseaux
  - partenaires allemands → élargissement du standard TEI à la CMC
  
- Actions
  - animation d'un site , d'un groupe Facebook , d'une liste de discussion
  - cycle annuel de conférences 2015-2018
  - groupe de travail TEI-CMC créé en 2014
  - table ronde Franco-Allemande 'Standards for CMC corpora'
  - French-German colloquium WikiCorp 2018 'Fostering linguistic studies on Wikipedia discussions'



## CMC Corpora of Computer-Mediated communication médiée par les réseaux

- CLARIN
  - Inclusion de CoMeRe dans CLARIN Resources Families
  - UIE : User Involvement Event 'How to use TEI for the annotation of CMC and social media resources: a practical introduction'





# Corpus multilingues et plurilingues



## Corpus multilingues et plurilingues des données dans plusieurs langues / plusieurs langues dans la même donnée

- ❑ Constitution de corpus écrits et oraux pour **des langues de grande diffusion** vs constitution de corpus oraux pour **des langues peu décrites** : quels outils, quels annotateurs, quelles priorités de recherche ?
- ❑ Exploitation quantitative de **corpus massifs** vs exploitation quantitative de **corpus réduits de langues peu étudiées** : quels modèles statistiques, quelles questions théoriques et quelles méthodes ?
- ❑ Difficulté à constituer des corpus de volume suffisant pour l'étude du **code-switching** : un étiqueteur par langue ou un seul étiqueteur qui accepte des passages dans une autre langue, Universal Dependency Tagset ou plusieurs tagsets, adapter les annotateurs automatiques aux corpus plurilingues
- ❑ Journée d'étude 2017 "Annotations et traitements automatiques"



Un grand merci pour votre attention !



Rencontres  
Huma-Num  
2018