**École thématique « Annotation des corpus oraux et multimodaux : quels outils pour quels objets de recherche en linguistique et en SHS ? »**
**Saint Pierre d'Oléron, 9-13 juin 2025**

| Horaires | Intervenant | Modalité/Activité | Titre |
|---|---|---|---|
| **LUNDI** | | | |
| **Après-midi** | | Arrivée des participants | |
| ***19h30*** | | *Diner* | |
| **20h30** | Katja Ploog | Conférence introductive | **Défis pour le traitement de l'oral** |
| **MARDI** | | | |
| ***8h-9h*** | | *Récupération des vélos* | |
| **9h-9h30** | Céline Dugua | Temps d'inclusion | |
| **9h30-11h30** | Katja Ploog | Atelier de présentations des corpus/travaux des participants | |
| **11h30-12h30** | Loïc Liégeois | Conférence | **Structurer, analyser, diffuser et valoriser ses données : panorama des outils et des méthodes disponibles pour le traitement des corpus oraux** |
| ***12h30-14h*** | | *Déjeuner* | |
| **14h-17h** | Christophe Benzitoun | Conférence + Atelier | **Conférence : Annotation morphosyntaxique de corpus oraux : usages et limites ?**<br>**Atelier : Prise en main du Corpus d'Étude du Français Contemporain (CEFC)** |
| ***17h30-19h*** | | *Découverte de l'île*<br>*Possibilité d'une visite + dégustation* | |
| ***19h30*** | | *Diner* | |

| MERCREDI | | | |
|---|---|---|---|
| **8h30-9h30** | Philippe Boula de Mareüil | Conférence | **Transcription de langues peu dotées : expériences autour d'un atlas sonore de dialectes et langues minoritaires** |
| **9h30-12h30** | Céline Dugua et Flora Badin | Atelier | **Une chaîne de traitement pluri-outillée de la transcription à son enrichissement : l'exemple de la liaison en français** |
| ***12h30-14h*** | | *Déjeuner* | |
| **14h-15h** | James Trujillo | Conférence | **Bringing Together Qualitative and Quantitative Methods for Annotating Multimodal Communicative Behavior** |
| **15h15-17h15** | Isabel Colon de Carvajal | Atelier | **Annotations multimodales d'interaction en contexte de jeux de société dans ELAN à des fins d'analyses quantitatives** |
| ***17h15-19h30*** | | *Détente* | |
| ***19h30*** | | *Diner* | |

| JEUDI | | | |
|---|---|---|---|
| **8h30-9h30** | Lucas Ondel | Conférence | **Structurer, analyser, diffuser et valoriser ses données : panorama des outils et des méthodes disponibles pour le traitement des corpus oraux** |
| **9h30-12h30** | Benjamin Lecouteux | Atelier | **Transcription automatique de corpus oraux : méthodes, outils et bonnes pratiques avec Whisper et SpeechBrain** |
| ***12h30-14h*** | | *Déjeuner* | |
| **14h-15h** | Elisabeth Délais-Roussarie | Conférence | **Transcrire la prosodie : quels systèmes pour quels faits linguistiques ?** |
| **15h15-17h45** | Brigitte Bigi | Atelier | **Automatiser l'annotation prosodique, phonétique et syntaxique d'un corpus oral** |
| ***18h*** | | *Apéro des pratiques* | ***Rencontres informelles d'échanges autour des projets personnels*** |
| ***20h-22h*** | | *Diner* | ***Prolongation des échanges autour des projets personnels*** |

| VENDREDI | | | |
|---|---|---|---|
| ***8h-9h30*** | | *Restitution des vélos et des chambres* | |
| **9h30-11h30** | Conférenciers et animateurs d'ateliers<br>Benjamin Lecouteux, Katja Ploog | Table ronde | **Perspectives et défis pour l'annotation des corpus oraux et multimodaux** |
| **11h30-12h15** | Collectif | Perspectives | **Prolongements de l'école : enjeux stratégiques et projets à venir** |
| ***12h30-13h30*** | | *Repas* | |
| ***13h30*** | | *Départ des participants en bus* | |

# Résumés des conférences et ateliers

## MARDI

**Loïc Liégeois - Structurer, analyser, diffuser et valoriser ses données : panorama des outils et des méthodes disponibles pour le traitement des corpus oraux**

La chaîne des traitements à appliquer à un corpus oral est souvent longue, complexe et diversifiée. Afin de couvrir les différentes étapes de cette chaîne, le chercheur ou la chercheuse est amenée à prendre en main une série de méthodes et d'outils différents dans le but de répondre à des problématiques hétérogènes liées, par exemple, à la transcription, à la structuration, à l'analyse ou à la diffusion de son corpus oral.

Cette conférence a l'ambition de proposer un panorama, loin d'être exhaustif, des solutions outillées existantes permettant de nous assister au cours des différentes étapes de la chaîne de traitement d'un corpus oral. L'accent sera mis sur l'utilisation d'outils libres, faciles à prendre en main et interopérables.

**Christophe Benzitoun**

**Conférence : "Annotation morphosyntaxique de corpus oraux : usages et limites ?"**

Depuis plus de 10 ans, rares sont les linguistes qui travaillent encore à partir des corpus oraux bruts (à partir de la seule transcription et de l'enregistrement correspondant). Le recours à la lemmatisation, aux annotations automatiques en parties du discours voire en relations de dépendances est de plus en plus systématique. Bien que cette pratique puisse rendre de grands services, elle fait aussi courir un certain nombre de risques qu'il est fondamental d'appréhender. Je présenterai un bref historique des annotations morphosyntaxiques dans le champ des corpus oraux en mettant l'accent sur les questions pratiques et théoriques qu'elles continuent de soulever.

**Atelier : "Prise en main du Corpus d'Étude du Français Contemporain (CEFC)"**

Dans cet atelier, je présenterai, dans un premier temps, le CEFC élaboré dans le cadre du projet Orféo sous la direction de Jeanne-Marie Debaisieux en me focalisant uniquement sur la partie orale. Je montrerai également les trois interfaces en ligne qui permettent actuellement de l'interroger. En plus de la plateforme du corpus sur laquelle il est possible de faire une recherche simple, on peut exploiter les annotations syntaxiques et les métadonnées grâce à Grew-match et le Lexicoscope. Ces trois outils seront présentés brièvement avant de passer à l'étude de cas à l'aide du logiciel TXM. Ce dernier point constituera le cœur de l'atelier.

## MERCREDI

**Philippe Boula de Mareüil - Transcription de langues peu dotées : expériences autour d'un atlas sonore de dialectes et langues minoritaires**

Cette communication s'appuie sur un corpus comparable constitué, pour un atlas sonore, d'une même histoire traduite dans plus d'un millier de versions, en dialectes et langues minoritaires de France et d'autres pays du monde. Nous présentons les solutions proposées pour la transcription orthographique voire phonétique, les problèmes d'orthographe ou parfois de translitération que cela pose, a fortiori quand il s'agit d'annoter les enregistrements sous forme de gloses. Différents cas d'école sont détaillés.

**Céline Dugua et Flora Badin - Une chaîne de traitement pluri-outillée de la transcription à son enrichissement : l'exemple de la liaison en français**

A travers cet atelier, nous montrerons un ensemble d'outils qui, à partir d'enregistrements oraux, permettent d'ajouter des niveaux d'annotations pour aboutir à des transcriptions enrichies. Les outils que nous verrons sont à la fois des outils qui permettent des transcription automatique (whisper), des transformations de fichiers (TeiCorpo), un enrichissement des transcriptions (Jtrans, TEICORPO), des analyses (Elan, Python). L'annotation sous TXM pourra également être abordée. Pour illustrer cette chaine, nous prendrons l'exemple du phénomène de liaison dans un sous-corpus d'ESLO.

**James TRUJILLO - Bringing Together Qualitative and Quantitative Methods for Annotating Multimodal Communicative Behavior [Conférence en anglais]**

Face to face language use is multimodal. It involves (in the case of spoken language) not only speech, but also visual signals such as facial signals and manual gestures. These signals need to be annotated in order to understand phenomena of multimodal language use, particularly in naturalistic or unconstrained settings. This process can be quite time-consuming and potentially challenging to maintain consistency when performed entirely by hand. At the same time, automated and quantitative methods are not advanced enough to replace human coding. In this talk, I will provide an introductory framework to annotating multimodal communicative behaviors using qualitative methods that are informed and augmented by automated and quantitative tools.

**Isabel Colon de Carvajal - Annotations multimodales d'interaction en contexte de jeux de société dans ELAN à des fins d'analyses quantitatives**

Cet atelier pratique sur ELAN a pour objectif de créer un schéma d'annotations multimodales à partir d'un corpus qui réunit 5 joueurs en train de jouer à un jeu de société diffusé simultanément en ligne sur Twitch.Tv. Cet atelier permet de voir la chaine de traitement à mettre en place pour produire des analyses quantitatives sur son jeu de données : création du vocabulaire contrôlé, des types de piste, des acteurs (parents/enfants), annotations des données, exportation des données dans un tableur, tableau croisé dynamique pour visualisation quantitative des données codées.

## JEUDI

**Lucas Ondel - Structurer, analyser, diffuser et valoriser ses données : panorama des outils et des méthodes disponibles pour le traitement des corpus oraux**

La chaîne des traitements à appliquer à un corpus oral est souvent longue, complexe et diversifiée. Afin de couvrir les différentes étapes de cette chaîne, le chercheur ou la chercheuse est amenée à prendre en main une série de méthodes et d'outils différents dans le but de répondre à des problématiques hétérogènes liées, par exemple, à la transcription, à la structuration, à l'analyse ou à la diffusion de son corpus oral.

Cette conférence a l'ambition de proposer un panorama, loin d'être exhaustif, des solutions outillées existantes permettant de nous assister au cours des différentes étapes de la chaîne de traitement d'un corpus oral. L'accent sera mis sur l'utilisation d'outils libres, faciles à prendre en main et interopérables.

**Benjamin Lecouteux - Transcription automatique de corpus oraux : méthodes, outils et bonnes pratiques avec Whisper et SpeechBrain**

La transcription automatique offre aux linguistes des possibilités pour explorer des corpus oraux à grande échelle. Cet atelier propose une initiation pratique à l'utilisation des outils Whisper (OpenAI) et SpeechBrain (open source), adaptés à la transcription automatique de données vocales pour des analyses linguistiques. Nous aborderons les enjeux méthodologiques liés à la préparation des corpus oraux, au choix et à

l'ajustement des modèles selon les spécificités des langues étudiées, et à l'évaluation critique des résultats. Les participants découvriront comment exploiter ces outils afin d'obtenir des transcriptions exploitables pour la recherche en linguistique.

**Elisabeth Delais-Roussarie - Transcrire la prosodie : quels systèmes pour quels faits linguistiques ?**

Cette communication se fixe un triple objectif : (i) rappeler à partir d'exemples précis quels sont les domaines de la prosodie (accentuation, phrasing, intonation), (ii) voir comme ils sont traités et représentés dans différents systèmes de transcription de la prosodie (manuels ou semi-automatiques), (iii) analyser quels systèmes sont les plus pertinents pour étudier des faits linguistiques mettant en jeu le phrasing et ses liens avec la syntaxe, l'intonation et l'interprétation des énoncés en contexte et l'accentuation.

**Brigitte Bigi - Automatiser l'annotation prosodique, phonétique et syntaxique d'un corpus oral**

L'un des moyens les plus sûrs qui permet de mettre en relation les phénomènes de différents domaines consiste à aligner ses annotations sur le signal sonore. Les données temporelles sont la plupart du temps des points ou des durées de segments linguistiques de natures et de tailles variées. Cet ancrage temporel, s'il est fiable, n'en reste pas moins difficile à obtenir : l'annotation est chronophage. Dans cette formation, nous verrons comment il est possible d'obtenir automatiquement, et de mettre en relation l'annotation prosodique Momel, avec les segmentations phonétiques en mots/phonèmes/syllabes, et avec l'annotation morpho-syntaxique.