

Annotation Lexicale

Une Mesure de Complexité Contextuelle

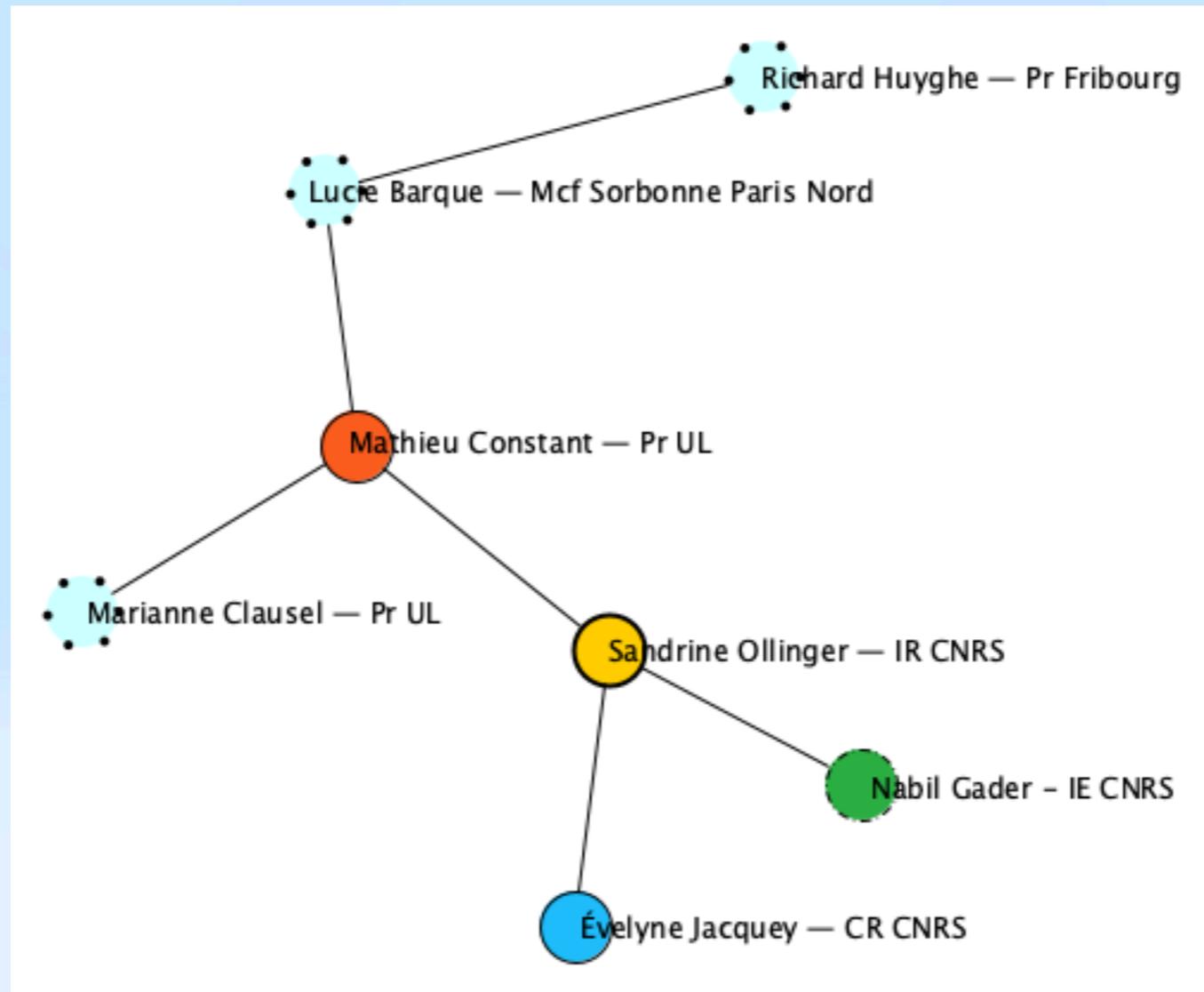


Sandrine Ollinger — 3 avril 2025

Rapide contextualisation

- Projet du GR PAL
- Volonté de travailler avec Lucie Barque (Université Sorbonne Paris Nord et LLF)
- Réponse Appel à Projet Pôle CLCS/ATILF
- Financement 2810€ (réunions + logiciel + petits matériel)
- Lancement mars 2024, prolongation → mai 2025

Participants



+ Jessika Cardinali + Benjamin Husson

Les objectifs initiaux du projet

Nous nous intéressons à la tâche d'annotation en sens lexicaux, traditionnellement désignée comme tâche de désambiguïsation lexicale en traitement automatique des langues. Dans le cadre du projet, deux annotatrices réalisent cette tâche manuellement et une annotation est réalisée automatiquement à l'aide du modèle développé par Aman Sinha.

Ces annotations sont réalisées sur le corpus brut. Il s'agit d'un corpus d'extraits (citations lexicographiques, paragraphes ou phrases; constitution en cours).

En parallèle, différentes analyses automatiques et manuelles sont réalisées pour enrichir le corpus de différentes couches d'annotation (classes sémantiques, voisins du RL-fr, syntaxe).

Le corpus ainsi enrichi nous servira à classer les extraits qu'il contient sur une échelle de complexité et de mettre au point une mesure de complexité contextuelle propre à la tâche d'annotation en sens lexicaux.

Les objectifs actualisés du projet

Nous nous intéressons à la tâche d'**annotation en sens lexicaux**, traditionnellement désignée comme tâche de désambiguïsation lexicale en traitement automatique des langues.

Nous nous intéressons à la réalisation **manuelle** et **automatique** de la même tâche : **assigner à une occurrence en contexte, un sens issu d'un inventaire de sens donné.**

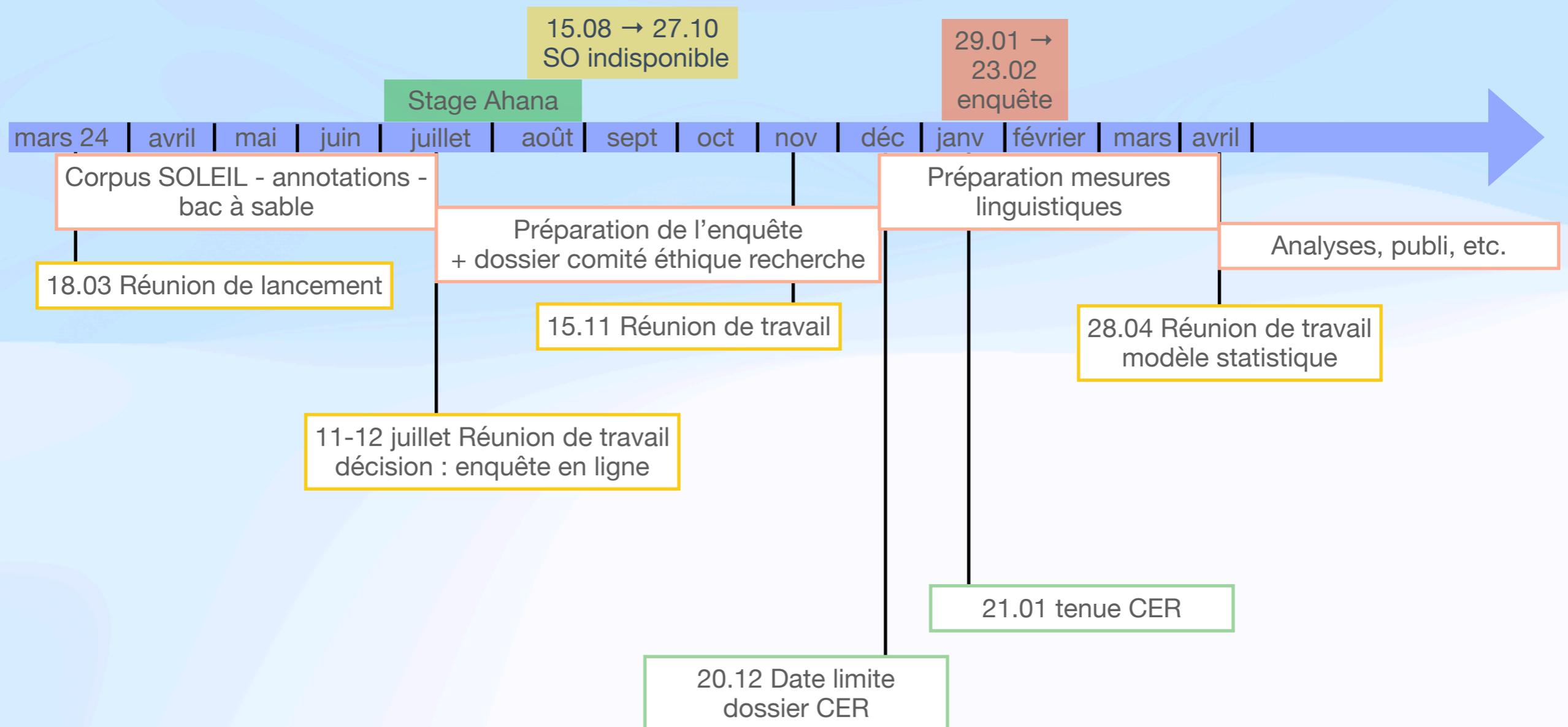
Nous nous concentrons sur la réalisation **manuelle** de la tâche et nous mettons en place une **enquête en ligne** pour sa réalisation par des annotateurs **non-experts.**

Nous cherchons à établir une **mesure de complexité contextuelle** pour cette tâche.

Ici, la complexité s'entend comme la **difficulté à accomplir la tâche.**

Nous recherchons des **indices linguistiques** qui permettent de mesurer cette difficulté, de la prédire.

Petite chronologie



À la carte

- Resana : environnement de travail collaboratif
- Préparation de l'enquête
 - ◉ Sélection des vocables (RL-fr, Lexique.org, SuperSens)
 - ◉ Sélection des contextes
 - ◉ Sélection des définitions
 - ◉ Limesurvey
- Com autour de l'enquête (Canva, LinkedIn, listes de diffusion)
- Premiers résultats statistiques à l'issue de l'enquête
- Mesures linguistiques (voisins lexicaux, entropie des sens)

Merci