*JOURNÉE D'ÉTUDE*

# REGARDS CROISÉS SUR LA CORÉFÉRENCE

## LINGUISTIQUE, TAL & SIC

**26 JUIN 2025**
**9H - 17H**

**SALLE 103**
**MAISON DE LA RECHERCHE SHS ANNIE ERNAUX**
**CY CERGY PARIS UNIVERSITÉ**

Organisation : Marine Delaborde & Hélène Manuélian

CY LT2D lexiques, textes, discours, dictionnaires

CY ADVANCED STUDIES

anr

# SOMMAIRE

# PRÉSENTATION

Les expressions référentielles désignant un même référent constituent les maillons d'une chaîne de référence et sont reliés par une relation de coréférence. Ce phénomène fait l'objet de travaux en linguistique d'un point de vue syntaxique, sémantique et discursif. Des travaux sur la coréférence s'inscrivent aussi en acquisition du langage, en didactique, en psycholinguistique, en traitement automatique des langues (TAL) ou encore dans le domaine des études littéraires. Si les cadres théoriques et les méthodes d'analyses diffèrent selon les disciplines, un dialogue interdisciplinaire pourrait permettre un enrichissement réciproque des approches.

En français, différentes initiatives de réalisation de corpus émergent depuis les années 2000 et s'entremêlent parfois pour traiter la coréférence ainsi que des phénomènes connexes. Le corpus ARCADE (2000) est l'un des premiers grands corpus annotés en relations anaphoriques. En 2006, le corpus DEDE a a proposé une annotation de descriptions définies et leurs liens de coréférence. Le projet ANNODIS (2011) a porté sur les chaînes topicales, qui concernent des segments possédant un même topique. Le projet ANCOR (2011-2013) a produit un corpus oral d'envergure annoté en relations anaphoriques (coréférentielles ou non). Pour l'écrit, le projet PEPS MC4 (2011-2012) et le projet ANR DEMOCRAT (2016-2020) ont donné lieu à des réalisations diverses : un corpus annoté, un outil d'annotation et de visualisation des chaînes de référence, des modèles de détection automatique de la coréférence et des publications qui ont contribué à la modélisation linguistique de ce phénomène. Une partie du corpus DEMOCRAT a été utilisée pour créer le corpus fr-litbank dans le projet du French BookNLP sur des textes littéraires. Le corpus DEMOCRAT (en 2022) et le corpus ANCOR (en 2025) ont été intégrés au projet Universal Anaphora pour représenter le français dans le corpus Coref-UD, un corpus annoté en coréférence constitué de 24 datasets pour 17 langues en 2025. Ce corpus est notamment utilisé dans les campagnes d'évaluation CRAC sur la détection automatique d'anaphore et de coréférence.

Cette journée d'étude entend proposer une réflexion structurée autour de trois sessions thématiques : (1) la caractérisation linguistique des chaînes et de leurs maillons, (2) la coréférence en corpus et la détection automatique en TAL et (3) l'étude des chaînes avec leur contexte. Au-delà de représenter une diversité d'approches et de points de vues, les travaux présentés permettront de nourrir une réflexion à propos de l'adaptation croisée d'outils et de méthodes pour l'étude de la coréférence en linguistique et en TAL, en envisageant leur réutilisation en sciences de l'information et de la communication.

**Cette journée d'étude est financée par CY Advanced Studies et par l'Agence Nationale pour la Recherche *via* la Chaire de Professeure Junior *Ressources numériques en sciences humaines et sociales* portée par Marine Delaborde au LT2D de CY Cergy Paris Université.**

# PROGRAMME

**9h** Accueil

**9h15** Introduction

## Session 1 : Caractérisation des chaînes

**9h30** **Hélène Manuélian & Marine Delaborde**
*Étendre la caractérisation des chaînes de référence*

**10h** **Catherine Schnedecker**
*Les pronoms indéfinis de première mention sont-ils réfractaires aux chaînes de référence ?*

**10h30** **Sylvia Federzoni**
*Caractériser les chaînes de référence par une approche linéaire*

**11h** Pause café

## Session 2 : Corpus et détection automatique de la coréférence

**11h30** **Frédérique Mélanie**
*De Democrat à fr-litbank : transformer les corpus pour de nouveaux usages*

**12h** **Antoine Bourgois**
*Extraction automatique des chaînes de coréférence de personnages dans les œuvres de fiction*

**12h30** **Olga Seminck**
*Il est où ton LLM ? Sur les défis de l'utilisation des LLMs pour la résolution de la coréférence*

**13h** Pause déjeuner

## Session 3 : Les chaînes dans leur contexte

**14h** **Yoann Dupont & Marine Delaborde**
*Cooccurrence et coréférence : exploration du vocabulaire spécifique à un référent*

**14h30** **Frédéric Landragin**
*Saillance et saillance référentielle*

**15h** **Paola Pietrandrea / Caroline Bossant**
*Co-construction de la référence dans les interactions en ligne et hors-ligne en français*

**15h30** Pause café

**16h** Conclusion, discussions et perspectives

# RÉSUMÉS

**PAR ORDRE ALPHABÉTIQUE**

## BOURGOIS, ANTOINE

### Extraction automatique des chaînes de coréférence de personnages dans les œuvres de fiction

Les personnages constituent un objet central des œuvres de fiction. Dans le cadre des études littéraires computationnelles, être capable d'extraire automatiquement les chaînes de coréférence des personnages constitue une étape clé pour de nombreuses applications : réseaux de personnages, analyses de corpus diachroniques, études culturelles, narratologie. L'automatisation de la tâche d'extraction des chaînes de coréférence nécessite une formalisation rigoureuse permettant son exécution par un ordinateur. Cette présentation proposera une introduction à la résolution automatique de la coréférence, en s'appuyant notamment sur les approches par paires de mentions. Nous aborderons également les difficultés spécifiques aux œuvres de fictions : longueur des documents, diversité des formes de références, ambiguïtés propres à la narration.

## DUPONT, YOANN & MARINE DELABORDE

### Cooccurrence et coréférence : exploration du vocabulaire spécifique à un référent

Les chaînes de référence permettent de représenter le devenir discursif d'un référent : quand et comment est-il mentionné dans un discours ? Nous proposons une méthode permettant d'étudier le vocabulaire spécifique à un référent, en dehors de son paradigme désignationnel. Nous appliquons la mesure de spécificités de Lafon (1980), non pas à l'ensemble des occurrences d'une forme, mais à l'ensemble des occurrences des maillons d'une même chaîne. Ainsi, nous obtenons les cooccurrents d'un référent qui correspondent au vocabulaire statistiquement sur-représenté dans le contexte des maillons de la chaîne de référence.

# RÉSUMÉS

## *FEDERZONI, SILVIA*

### Caractériser les chaînes de référence par une approche linéaire

Ce travail propose une analyse des chaînes de référence en s'intéressant aux modalités d'enchaînement des maillons dans les textes. Nous faisons l'hypothèse que ces enchaînements obéissent à des régularités qui reflètent des stratégies discursives mises en place pour introduire, maintenir ou réactiver les référents. Notre objectif est de dégager des patrons récurrents et d'interroger leur contribution à l'organisation textuelle. Afin de fournir une description systématique des enchaînements des maillons, tout en prenant en compte un large volume de données, nous proposons une méthode outillée de corpus adoptant une approche linéaire — et non plus uniquement globale — des chaînes de référence. Cette méthode consiste dans la combinaison des techniques du Traitement Automatique des Langues et des techniques d'analyse des séquences, utilisées traditionnellement en Sciences Sociales. Nous présenterons les résultats obtenus en appliquant cette méthode aux chaînes de référence du corpus Democrat. Les résultats obtenus sont également croisés avec plusieurs facteurs de variation — genre textuel, type de texte, nature du référent — afin d'étudier leur impact sur les formes d'enchaînement observées.

## *LANDRAGIN, FRÉDÉRIC*

### Saillance et saillance référentielle

Beaucoup de travaux considèrent la saillance comme saillance référentielle, c'est-à-dire liée aux référents du discours et à leur réalisation textuelle. Les recherches s'intéressent alors aux anaphores, cataphores, coréférences et chaînes de référence, et convoquent des notions comme l'accessibilité, le centrage, la topicalité. Nous proposons un panorama des facteurs de saillance ainsi identifiés dans la littérature, et nous mettons en perspective ces facteurs avec d'autres, issus de la prosodie, du lexique ou de la syntaxe - cette fois en dehors de tout aspect référentiel. Cela nous amène à poser plusieurs questions de recherche, qui constituent une sorte de programme à discuter collectivement.

# RÉSUMÉS

## MANUÉLIAN, HÉLÈNE & DELABORDE, MARINE

### Étendre la caractérisation des chaînes de référence

La notion de chaîne de référence a 50 ans cette année (Chastain, 1975). De nombreuses études linguistiques ont depuis cherché à les décrire et les caractériser. Dans quel but ? Elles permettent d'analyser le devenir discursif d'un référent, de faire toutes sortes d'analyses contrastives (langues, genres de textes…). Pour autant, la caractérisation des chaînes est orientée par les ressources utilisées, les outils disponibles mais surtout par le sujet et le domaine scientifique de l'analyse. Sur la base de l'état des lieux proposé par Schnedecker (2021), nous nous demanderons comment les outils, mesures et critères utilisés en linguistique outillée peuvent être adaptés à d'autres domaines et sur des concepts proches.

## MÉLANIE, FRÉDÉRIQUE

### De Democrat à fr-litbank : transformer les corpus pour de nouveaux usages

La création d'un modèle d'annotation suppose des données préalablement annotées. Nous comparons ici deux projets – DEMOCRAT et Propp – tous deux reposent sur un corpus annoté en « mentions » et « chaînes de référence », c'est-à-dire en unités textuelles identifiées et reliées parce que désignant un même référent. DEMOCRAT a produit un corpus de référence inédit en français ; Propp, en s'appuyant sur ces données, a développé un modèle d'annotation de personnages.
Les deux projets poursuivent des objectifs distincts, le corpus de référence de chacun d'eux en témoigne. Celui de DEMOCRAT, issu de sources variées, couvre des textes allant du XI$^{e}$ au XX$^{e}$ siècle ; fr-litbank (corpus de Propp) ne conserve de DEMOCRAT que les textes littéraires du XIX$^{e}$ siècle et les adapte, afin de permettre l'annotation et la détection de chaînes de personnages.
Notre analyse visera à qualifier et quantifier ces adaptations, à illustrer le travail fastidieux et consciencieux requis pour constituer un corpus exploitable.

# RÉSUMÉS

## *PIETRANDREA, PAOLA & BOSSANT, CAROLINE*

### Co-construction de la référence dans les interactions en ligne et hors-ligne en français

Nous définissons, avec Charolles (2002), la référence comme un acte propositionnel d'introduction d'entités dans le discours. Dans la lignée de l'analyse conversationnelle et des approches psycholinguistiques (Clark & Wilkes-Gibbs 1986), nous étudions sa dimension collaborative : listes, deixis, reformulations, marqueurs de négociation (Voghera 2012). Nous faisons l'hypothèse que la variation socio-sémiotique influence le choix des stratégies référentielles. Dans ce cadre, la communication digitale peut transformer les stratégies de co-construction de la référence, les rendant parfois plus complexes (Pietrandrea 2021). Pour tester cette hypothèse, nous avons constitué un corpus pilote (échanges oraux, Facebook, LinkedIn, Twitch) et conçu un schéma d'annotation des chaînes de co-référence. Les premières analyses tendent à confirmer notre hypothèse : la communication digitale peut avoir un impact sur la fluidité des processus de co-construction référentielle (Pietrandrea & Bossant 2024).

## *SCHNEDECKER, CATHERINE*

### Les pronoms indéfinis de première mention sont-ils réfractaires aux chaînes de référence ?

Cette étude vise à déterminer dans quelle mesure un pronom indéfini (désormais PI) peut servir de premier maillon à une chaîne de référence (désormais CR), ou suite des expressions coréférentielles d'un texte, s'inscrivant ainsi à la croisée de deux domaines d'études : celui des chaînes de référence (cf. Landradin 2021 ; Landragin & Schnedecker 2014 ; Schnedecker, Glikman & Landragin 2017 ; Schnedecker 2021) et celui des pronoms indéfinis, peu abordé en termes d'anaphores ou de coréférence, ou circonscrit à la forme on dont le traitement coréférentiel a été abordé par Delaborde 2021. La question des premiers maillons vs singletons n'est pas anodine compte tenu du bruit abondant auquel est confronté le traitement des CR, puisque pratiquement trois quarts des mentions sont des singletons.
Après un état de la question sur les position et fonctions des SN indéfinis dans les CRet en nous appuyant sur une mini étude de corpus, nous aborderons les paramètres susceptibles d'influer sur la (non(-) reprise suite à des PI : leurs catégorie et fonctions grammaticale, référentielle et textuelle, la catégorie ontologique du référent, le genre discursif d'occurrence ainsi que leur forme-même.

# RÉSUMÉS

## *SEMINCK, OLGA*

### Il est où ton LLM ? Sur les défis de l'utilisation des LLMs pour la résolution de la coréférence

La résolution automatique des anaphores et de la coréférence a une longue histoire, allant des systèmes à règles aux approches statistiques, puis à l'utilisation des réseaux neuronaux. Depuis l'arrivée des grands modèles de langue génératifs (LLMs, pour Large Language Models, ou GenAI, pour Generative Artificial Intelligence), pilotables à l'aide d'instructions en langue naturelle, de nombreuses tâches « classiques » du traitement automatique des langues (TAL) se retrouvent remises en question. Offrant des solutions de bout en bout prêtes à l'emploi, ces modèles semblent rendre superflues certaines analyses linguistiques intermédiaires à la coréférence, comme l'analyse syntaxique ou l'identification des entités nommées. Avec quelques essais simples et des instructions élémentaires, un LLM semble capable d'identifier assez fiablement les chaînes de coréférence, sans passer par les étapes complexes typiques des systèmes traditionnels de TAL. On peut donc se demander s'il ne serait pas temps de remplacer les systèmes de résolution de la coréférence par des LLMs.

Dans cette présentation, je montrerai en quoi cette idée est naïve. En plus des problèmes bien connus liés à l'utilisation des LLMs — manque de transparence quant à la constitution du modèle, aux données de pré-entraînement, et à son comportement ; concentration des ressources computationnelles nécessaires entre les mains d'acteurs très puissants, rendant leur entraînement inaccessible à la majorité —, la tâche de la coréférence pose des défis spécifiques qui rendent l'usage des LLMs loin d'évident.

# BIBLIOGRAPHIE


- **Chastain,** C. (1975). Reference and Context: Language, Mind, and Knowledge (Minnesota Studies in the Philosophy of Science 7).
- **Charolles**, M. (2002). La référence et les expressions référentielles en français. Paris : Ophrys.
- **Clark**, H. H., **Wilkes-Gibbs**, D. (1986). “Referring as a collaborative process”. Cognition, 22, 1-30.
- **Delaborde**, M. (2021). La coréférence floue dans les chaînes du corpus DEMOCRAT. Langages, 224(4), 47-65.
- **Lafon**, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. Mots. Les langages du politique, 1(1), 127-165.
- **Landragin**, F. (2021). Le corpus Democrat et son exploitation. Présentation. Langages, 224(4), 11-24.
- **Pietrandrea**, P. (2021). Comunicazione, dibattito pubblico, social media. Come orientarsi con la Linguistica. Carocci: Roma.
- **Pietrandrea**, P., **Bossant**, C. (2024). « Discours numérique et co-construction de la référence ». Séminaires du Lattice – 17 décembre 2024.
- **Schnedecker**, C., **Glikman**, J., & **Landragin**, F. (2017). Les chaînes de référence: annotation, application et questions théoriques. Langue française, 195(3), 5-16.
- **Schnedecker**, C., & **Landragin**, F. (2014). Les chaînes de référence: présentation. Langages, 195(3), 3-22.
- **Schnedecker**, C. (2021). Les chaînes de référence en français. Paris : Ophrys.
- **Voghera**, M. (2012). “Chitarre, violino, banjo e cose del genere”. In M. Voghera & A. Thornton (a cura di) Per Tullio De Mauro. Studi offerti dalle allieve in occasione del suo 80° compleanno, 341-364. Aracne.

# INFOS UTILES

## MAISON SHS

Salle 103 (1er étage)
33, Boulevard du Port
95011 Cergy-Pontoise-Cedex
Accès

## SITE DE LA JOURNÉE

https://marine-delaborde.github.io/coreference/je25.html