Recherche linguistique et production de ressources

Jeudi 15 mai 2015

Salle de documentation du CRISCO Campus 1,
Bât. N3, Salle SA S13 Université Caen Normandie

Atelier de l'Axe 2 du Centre de recherches inter-langues sur la signification en contexte (CRISCO)

Livret de résumés

Séance 1 : Lexicographie

Intervenante : **Laurette Chardon**

Titre: Les quatre projets en cours autour du Dictionnaire Électronique des Synonymes en 2025 (DÉS)

Résumé : Depuis sa création dans les années 1990, le DES (https://crisco4.unicaen.fr/des/) s'est continuellement enrichi non seulement au niveau des synonymes mais également avec les variantes, les antonymes et les catégories grammaticales. Actuellement, cinq grands projets sont en cours :

- la récupération des requêtes dans les logs permet après analyse de connaitre les mots recherchés (vedettes, variantes et mots inexistants) avec leurs fréquences.
  ** L'étude des mots les plus recherchés permet d'approfondir les usages : quels types de mots sont les plus demandés ? y a-t-il beaucoup de fluctuation d'un jour à l'autre, d'un mois à l'autre ?
  ** L'examen des requêtes sur des mots inexistants, après nettoyage du spam, met en évidence des manques qui peuvent être corrigés : des entrées principales et de variantes féminines, orthographiques ou correctives.
- une nouvelle interface réalisée avec le *framework* Django compatible avec les téléphones et les tablettes avec des outils pour mieux exploiter les cliques
- l'ajout progressif des contextes par paire de synonymes avec des méthodes automatiques de traitement des corpus (FRANTEXT, ChatGPT, Sketchengine, Réseau Lexical du Français, Wiktionnaire)
- enfin, un projet de visualisation 3D des différents sens d'un mot de l'espace sémantique est en cours avec le CIREVE. Il permettra d'affiner les méthodes de calculs pour obtenir un résultat satisfaisant en particulier pour faciliter l'apprentissage de la langue pour les apprenants.

CRISCO
Langues | Signification | Contexte

Intervenant : **Jacques FRANCOIS**

Membres du projet : Laurette CHARDON, Mathieu GOUX, Lina BLONTROCK (stagiaire orthophonie), Justine REYNAUD (jusqu'en 2024), Triss JACQUIOT (en 2022) Modélisation Graphique de la Polysémie Evolutive (MGPE)

Titre : Modélisation Graphique de la Polysémie Evolutive (MGPE)

Résumé : Ce projet a été préparé en 2020-21 par trois articles de Jacques François publiés respectivement dans les *Cahiers de lexicologie*, la *Zeitschrift für romanische Philologie* et *Travaux de linguistique*. A la suite de ces publications Jacques a travaillé sur cette question avec Justine Reynaud du GREYC en 2021-22, puis Triss Jacquiot a élaboré une maquette d'interface graphique et Laurette Chardon s'est associée au projet. En septembre 2023 nous avons rédigé un premier bilan d'étape envoyé à l'ATILF, qui suit attentivement nos progrès, et nous l'avons publié sous deux versions légèrement différentes comme Cahier 37 du CRISCO et comme contribution au n° 117 du Bulletin de la Société de Linguistique de Paris. En septembre 2024, Justine a souhaité se retirer du projet et Mathieu Goux a accepté de nous faire bénéficier de ses compétences en traitement de corpus historiques, en humanités et en gestion administrative du projet. En octobre 2024 nous avons rédigé un second bilan d'étape consacré en priorité à la présentation d'une version de démonstration de nos résultats sur 500 entrées historiques. Ces résultats (tableau de conversion graphique et interface graphique) sont accessibles sur le site du CRISCO (crisco3.unicaen.fr/notices.tlfi/) et ce second bilan est déposé sur la plateforme HAL avec toutes nos autres publications sur ce projet.

Notre objectif immédiat est d'augmenter progressivement le nombre des entrées historiques traitées, de corriger quelques bavures persistantes et de colorier les arêtes des interfaces graphiques historiques afin de clarifier le cheminement de chacune dans les interfaces les plus complexes (au-delà de 10 nœuds correspondant chacun à une des rubriques de l'entrée du TLFi et jusqu'à 150 nœuds pour le verbe TENIR). A moyen terme nous souhaitons poursuivre notre exposé des premiers acquis de différentes analyses transversales entre groupes d'entrées présentant des propriétés communes.

Séance 2 : Syntaxe

Intervenants : **Stéphane FERRARI, Richard RENAULT**

Titre : Intégration et évaluation des outils d'analyse syntaxique du corpus Malherbe

Résumé : Une analyse syntaxique du corpus (lemmes, catégories, flexions et fonctions), combinée à une analyse métrique des vers et strophes permet de dégager des observations pertinentes dans le cadre de l'étude de la concordance entre la structure métrique et la structure syntaxique. Notre présentation portera

1) sur l'intégration des données syntaxiques dans les fichiers XML du corpus analysé,

2) sur l'étude d'un cas particulier lié à la problématique du statut des auxiliaires dans le cadre des grammaires de dépendance, et

3) sur la mise en place d'une procédure d'évaluation des outils d'analyse syntaxique et notamment des mesures relatives aux performances des différents outils et à l'accord entre les annotateurs.

Intervenants : **Adeline PATARD, Kylian ROUSSEL**

Titre : Un premier treebank du normand

Résumé : Cette intervention portera sur une première expérience d'analyse syntaxique d'un texte littéraire en dialecte normand en utilisant des outils de parsing syntaxique. Issue du corpus Paroles de Normands (https://mrsh.unicaen.fr/parolesdenormands/ Lajoye, Laîné et Manœuvrier 2024), la « Légende du roi Arthur racontée par les instituteurs de Saint-Georges-de-Rouelley (canton de Barenton, Manches) en 1913 », publiée en 1969 par Fernand Lechanteur (*Parlers et traditions populaires de Normandie*, 5, p. 4-6) a été tokenisée, annotée en catégories et fonctions syntaxiques et lemmatisée en utilisant l'interface ArboratorGrew (Guibon et al 2020). Nous présenterons les défis rencontrés lors de cette première expérience de constitution de *treebank* du normand et les solutions élaborées ainsi que les premières évaluations de l'adaptation des outils du parsing automatique sur cette variété de la langue d'oïl.

Intervenant : **Pierre LARRIVÉE**

Titre : A quoi peut bien servir un protocole optimisé d'annotation syntaxique ?

Résumé : Cette intervention évoquera les protocoles d'annotation syntaxique utilisé dans le cadre de projets récents au CRISCO (ConDÉ, MICLE, High-Tech, AUTOMATED, corpus CorAG) afin de s'interroger sur les applications de la démarche de l'outillage des corpus en diachronie. Premièrement, la visée théorique est de définir le changement de la forme des phrases à travers le temp pour établir des comparaisons historiques (Samo 2023, Samo et Isolani 2024) ou typologiques (Chen *et al* 2023). On peut également envisager l'utilisation des données syntactiquement annotées dans le domaine de l'acquisition de la langue écrite, et en particulier dans les langues diglossiques (Biber and Gray 2016). Une autre application possible est dans le domaine de la santé, et en particulier la santé mentale. L'analyse des productions peut en effet soutenir le diagnostic, y compris en principe le diagnostic précoce (Smirnova et al 2018 pour la dépression ; Fraser et al 2016 pour la maladie d'Alzheimer ; Ehlen *et al* 2023 pour la schizophrénie). Enfin, un ensemble de données annotées syntaxiquement favorise l'extraction d'information (Osenova & Simov 2011).

Séance 3 : Multimodalités

Intervenante : **Catharine MASON**

Titre : Projet VOVA (Vocal and Verbal Arts)

Résumé : Cette présentation du projet VOVA insistera sur le double processus de l'entextualisation, définie comme la configuration stylistique d'un discours qui lui donne un cadre cohésif et cohérent de signification intersubjective en contexte. Terme issu des études en anthropologies linguistique américaine, 'entextualisation' se réfère, à la fois, à la formulation spontanée multifactorielle et multifonctionnelle d'un discours, et à sa capacité d'être décontextualisé et recontextualisé pour une nouvelle entextualisation, une pratique courante dans la vie quotidienne. Dans le cadre scientifique, cette nouvelle entextualisation sert de modèle d'un événement de performance. A partir d'exemples de modèles textuels de discours dits et chanté, je présenterai le type d'analyse visé dans la création d'une standardisation de représentation visuelles en XML-TEI d'expressions stylistiques vocales et verbales du discours tirées de contextes de performance.

Intervenant : **Elisio RENOUF**

Titre : Constitution et analyse d'un corpus multimodal à partir du *subreddit r/memes* : publications et commentaires

Résumé : Cette étude se concentre sur la constitution de corpus multimodaux, faisant état de différents modes de communication, i.e. le texte, l'image ou encore le son. Il s'agit de répertorier des publications issues de *r/memes*, une page du site *Reddit* qui est dédiée aux mèmes, soit ces images, fixes ou mobiles, surmontées d'un texte que les internautes partagent massivement sur internet dans le but de faire rire leur auditoire virtuel. Notre but est de découvrir les façons dont les internautes peuvent communiquer sur internet grâce aux mèmes. Cette présentation portera sur les critères de sélection, sur la méthodologie de la récolte et sur l'analyse des données. Nous utilisons la Grille d'Analyse Systémique Mémétique (ou GASM) de Wagener (2022), permettant une analyse des mèmes sous les angles de l'image et du texte. Les commentaires, à leur tour, sont soumis à une analyse syntaxique et sémantique.

Séance 4 : Dialectes

Intervenants : **Patrice LAJOYE, Stéphane LAINÉ**

Titre : Paroles de Normands: bilan et perspectives

Résumé : Après quatre ans de travail, une première tranche du corpus de textes dialectaux normands "Paroles de Normands" a été mise en ligne. Quelles ont été les difficultés rencontrées pour parvenir à ce résultat? Quelles sont les points à améliorer, et surtout quels sont les travaux restant à effectuer pour que le corpus puisse être considéré comme pleinement efficient?"

Intervenante : **Doriane MUSY**

Titre : La perception et la représentation fictive des (multi)ethnolectes en Norvège

Résumé : Notre recherche s'intéresse à l'usage des (multi)ethnolectes en Norvège et à sa représentation dans les médias, plus précisément les séries télévisées. Elle repose sur une analyse d'un corpus de dialogues extraits de deux séries télévisées norvégiennes « 17 » et « Førstegangstjenesten » et des entretiens réalisés avec des étudiants norvégiens. D'un côté, l'étude des dialogues télévisuels vise à identifier la manière dont les (multi)ethnolectes sont représentés dans les productions culturelles, notamment grâce aux emprunts lexicaux à d'autres langues. Ces séries jouent un rôle dans la construction des imaginaires sociaux autour des pratiques linguistiques associées aux communautés multiculturelles. D'un autre côté, les entretiens avec des étudiants permettent de recueillir des témoignages sur la façon dont ils perçoivent les (multi)ethnolectes dans leur quotidien, ainsi que le rôle du discours urbain dans les questions d'identité, d'intégration et de rapports multiculturels. Notre étude cherche donc à comprendre comment les (multi)ethnolectes sont perçus et vécus en Norvège, avec une mise en lumière des différences entre appréciation et stigmatisation

Séance 5 : Corpus oraux

Intervenants : **Natasha Romanova et Maxence Multin**

Titre : Corpus CAENNAIS : les défis de la transcription d'un corpus d'apprenants

Résumé : Le corpus Corpus Audio d'Étudiants Natifs et non-NAtifs en InteractionS (CAENNAIS) est un projet pédagogique mené par une équipe d'enseignants, ingénieurs et étudiants au CRISCO. Dans sa première phase, il a eu pour but de collecter, entre novembre 2023 et mai 2024, des enregistrements de

conversations libres entre les locuteurs francophones natifs et non-natifs (étudiants norvégiens lors de leur année en France) (Pinsault, 2024). Chacun des six groupes ayant été enregistré trois fois entre novembre et mai, les dix-huit heures d'enregistrements recueillis sont actuellement, dans la deuxième phase du travail, en cours de transcription. Dans cette intervention nous présenterons le corpus et les avancées actuelles dans l'application des logiciels de transcription automatique sur les données d'interaction.