Journée d'études
« Recherche linguistique appliquée à des pratiques et à la production de ressources électroniques »

8 juin 2023, 9h30-17h, Bibliothèque du CRISCO (UniCaen)
Contact : mathieu.goux@unicaen.fr

Cette journée d'études est consacrée à la présentation des projets relevant de l'axe 2 du CRISCO, et à la production de ressources numériques pour la recherche en linguistique. Les six interventions de la journée présentent les projets, en cours et futurs, des membres permanents ou associés au laboratoire.

L'accès à la journée est parfaitement libre, et s'adresse tant aux enseignants-chercheurs qu'aux étudiants. Les créneaux horaires sont indicatifs : selon les discussions, il se peut qu'une présentation commence un peu plus tôt, ou plus tard, que l'heure indiquée sur le programme.

Programme

9h15 - 9h30 : accueil / mots introductifs

9h30 - 10h15

En dialogue avec les outils d'apprentissage automatique : une chaîne de traitement pour l'annotation syntaxique

- Rayan Ziane, ingénieur d'études (projet High-Tech)
- Natasha Romanova, coordinatrice de projet (projet MICLE)
- Manon Lavergne, stagiaire M2 (projets High-Tech et MICLE)

Ces dernières années, grâce à l'émergence d'outils d'intelligence artificielle par apprentissage automatique, l'analyse syntaxique automatisée devient de plus en plus accessible aux chercheurs tout en minimisant le travail d'annotation manuelle. Les analyseurs syntaxiques dont *HOPS (HOnest Parser of Sentences)* et *Stanford parser* rendent possible l'analyse syntaxique à une échelle sans précédent, proposant un programme de *tagging* (identification de parties du discours) et *parsing* (identification de fonctions syntaxiques). Il reste, néanmoins, nombre de défis à une intégration harmonieuse de ces outils à la pratique de constitution de corpus annotés. Pour pouvoir profiter de ces technologies, les données textuelles doivent être prétraitées d'une certaine façon et converties en formats compatibles et prérequis. Les analyseurs syntaxiques utilisent des formalismes particuliers qui peuvent diverger avec les pratiques adoptées par les majorités des communautés de chercheurs ou ne pas correspondre aux objectifs du projet. Malgré les taux de succès de l'annotation impressionnants, la sortie de l'analyse automatique, surtout dans le cas de corpus en diachronie, nécessite toujours un programme de vérification manuelle ne serait-ce que dans le but de réentrainement de modèles.

Dans cette intervention, nous présenterons une chaîne de traitement développée et testée dans le cadre des projets High-Tech et MICLE, actuellement en cours au Laboratoire CRISCO. Une partie de chacun de ces corpus sera mise à disposition via le portal TXM du CRISCO au printemps 2023 et sera disponible pour consultation et interrogation lors de la journée d'étude. Cette chaîne permet de relever une partie des défis esquissés ci-dessus et repose sur l'utilisation du *parseur* HOPS, basé sur l'analyse syntaxique en dépendances avec le formalisme Universal Dependencies (UD). Consistant en cinq phases : 1) prétraitement (segmentation et tokenisation), 2) *tagging* et *parsing* UD (HOPS), 3) lemmatisation, 4) conversion des jeux d'étiquettes (UPenn et Presto), 5) reparsing du texte annoté et vérifié (HOPS), cette chaîne prévoit une étape de correction manuelle entre chacune d'elles. Cette approche favorise la flexibilité et permet aux annotateurs de continuer à modifier, par exemple, la segmentation en *tokens* lors des étapes suivantes, dont la vérification de la lemmatisation ou conversion des jeux d'étiquettes.

La chaîne proposée privilégie la vérification et l'affinage de l'annotation en parties du discours (*tagging*) en proposant une analyse plus fine que celle de l'analyseur automatique, tout en permettant un requêtage qui combinerait les PoS et un nombre de fonctions principales (notamment, sujet, objet, oblique). Il est alors possible d'effectuer des requêtes complexes sur les textes du corpus, par exemple chercher des objets sans déterminant préposés au verbe. En outre, dans la phase (4) de conversion de jeux d'étiquettes, nous nous servons de l'annotation en dépendances produite par HOPS pour affiner l'annotation et désambiguïser des formes qui pourraient l'être par des scripts Python. Par exemple, un participe passé ou une forme conjuguée d'un verbe, ou pour faciliter le repérage des noms au pluriel (ce travail a été mené dans le cadre d'un stage de M2 au CRISCO).

Combinant l'utilisation des outils d'apprentissage automatique, des scripts python basés sur des règles établies lors de l'analyse du corpus annoté et des étapes de vérification manuelle, nous proposons donc un protocole cohérent et facile à prendre en main qui permet d'effectuer une annotation syntaxique fiable, à coût réduit, tout en gardant le contrôle sur le déroulement du programme d'étiquetage.

Références

Blumenthal, P., Diwersy, S., Falaise, A., Lay, H., Souvay, G., Vigier, D., Descartes, P. R., Nancy, U., & de Lyon, U. (2017). *Presto, un corpus diachronique pour le français des XVIe-XXe siècles*.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics, 47*(2), 255-308. https://doi.org/10.1162/coli_a_00402

Grobol, L., & Crabbé, B. (2021). Analyse en dépendances du français avec des plongements contextualisés (French dependency parsing with contextualized embeddings). *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, 106-114. https://aclanthology.org/2021.jeptalnrecital-taln.9

Grobol, L., Prévost, S., & Crabbé, B. (2021). Is Old French tougher to parse? *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, 27-34. https://aclanthology.org/2021.tlt-1.3

Santorini, B. (2007) *Protocole d'étiquettage – Parties du discours (PDD)*. https://www.ling.upenn.edu/~beatrice/corpus-ling/annotation-french/pos/pos-index.html

Liens

Projet MICLE

Projet High-Tech
HOPS
Stanford Parser
Universal Dependencies

10h15 - 11h00

Défis interculturels et didactiques d'une collaboration transfrontalière

- Anne Prunet, directrice du Carré International
- Catrine Bang-Nilsen, enseignante-chercheure Norwegian University of Science and Technology

Notre exposé aura pour objectif de présenter les défis interculturels et didactiques d'une collaboration transfrontalière entre la France, la Norvège, l'Espagne et l'Allemagne. Nous exposerons le contexte éducatif norvégien et nous présenterons l'historique de la conception des programmes. Nous nous attarderons ensuite sur la dimension de l'ingénierie pédagogique et sur la conception de ressources électroniques et sur les défis rencontrés. Nous présenterons enfin les résultats des programmes (formation initiale et continue) et de l'axe de recherche.

Le projet d'Atlas linguistique numérique de la Normandie

* Stéphane Lainé, dialectogue (UniCaen)

S. Lainé nous présentera brièvement ce projet, avant la pause-café.

Publié en 5 volumes de 1980 à 2019, l'*Atlas linguistique et ethnographique normand* (ALN) restitue les résultats d'enquêtes effectuées de 1970 à 1977 par le linguiste Patrice Brasseur auprès de 697 informateurs dans les 5 départements normands et dans les îles anglo-normandes de Jersey, Guernesey et Sercq. Le projet vise à rendre facilement accessibles et à valoriser les données de l'ALN en les publiant sur une plateforme numérique et en les enrichissant de ressources interactives (enregistrements sonores et visuels, collections du Réseau des musées normands…).

11h00 - 11h15 : Pause café

11h15 - 12h00

Sur le *Dictionnaire Électronique des Synonymes* (DES)

- Laurette Chardon, ingénieure d'études (CRISCO)

Le *Dictionnaire Électronique des Synonymes* (DES) du Crisco, créé dans les années 1990, est un outil en accès libre devenu au fil des années de plus en plus utilisé. Il reçoit en moyenne 150.000 requêtes par jour.

Pourquoi un tel engouement ? Qui l'utilise ? Quelle est sa particularité ? Son originalité ? Comment est-il maintenu ? Quels sont les éléments le composant et les outils annexes

développés ? Quel est son apport en tant que projet de recherche permanent du laboratoire ? Quel est son positionnement par rapport au Plan national pour la science ouverte 2021-2024 ?

Nous allons lors de cette présentation répondre à ces questions et détailler deux points particuliers : d'une part la création des cliques avec le graphe d'adjacence et l'espace sémantique, et d'autre part l'insertion des catégories grammaticales.

Références :

Chardon, L. (2020), « Le Dictionnaire Électronique des Synonymes (DES) et ses graphes d'adjacence », sur HAL.
Chardon, L. (2022). « Insertion des catégories grammaticales dans le Dictionnaire Électronique des Synonymes (DES) », sur HAL.
Chardon, L. & François, J. (2020). « Les vedettes du Dictionnaire Électronique des Synonymes et les relations d'adjacence entre leurs synonymes », sur HAL.

Lien :

Présentation en ligne du Dictionnaire Électronique des Synonymes

12h - 14h : Pause déjeuner

14h - 14h45

Relations entre rythme et sens dans la poésie versifiée

- Éliane Delente, enseignante-chercheuse (CRISCO)
- Stéphane Ferrari, enseignant-chercheur (CRISCO)

L'objet du projet présenté concerne les relations rythme / sens dans la poésie versifiée française. Nous étudions comment les constituants métriques, prosodiques, syntaxiques et discursifs progressent, souvent ensemble, mais parfois en décalé, dans la réception du discours versifié.

À ce jour, il n'existe pas véritablement de théorie de ces relations, pas de méthode d'analyse ni de corpus construits à cet effet. Le projet développe ces trois aspects simultanément, que nous présenterons un à un.

Une réflexion théorique

Le rythme est une succession d'événements temporels *perçus* par le lecteur ou l'auditeur. Cette dimension temporelle du traitement de tout discours, et à plus forte raison, de la poésie métrique, pose la question du type de grammaire utilisé. Nous verrons que les grammaires traditionnelle (normative), en constituants immédiats ou générative sont inaptes à rendre compte de ce caractère dynamique. Les grammaires dynamiques[1] sont, de fait, plus adaptées pour appréhender le discours, non comme un produit fini, statique, mais comme un flux structuré en cours de développement. Elles permettent de mieux rendre compte de cet aspect dynamique des structures métrique, prosodique ou syntaxique qui apparaissent linéairement

1 Voir, entre autres, les travaux de Sinclair, Mauranen, Auer, Auer-Couper-Khulen-Müller, O'Glady, Martin.

mais ne se développent pas toujours en synchronie. Elles sont également capables d'intégrer les attentes et prédictions du lecteur (auditeur), constitutives du processus de traitement.

Une méthode d'analyse

Parallèlement à cette réflexion, nous tentons d'établir une méthode d'analyse de ces décalages rythme / sens. Pour ce faire, nous avons dû constituer des sous-corpus ciblés afin d'observer la diversité des phénomènes étudiés.

Nous avons ainsi exploité le corpus TEI Malherbe (R. Renault, Criso Lab) pour constituer des corpus plus restreints, automatisés ou non, à partir des œuvres de Boileau (1636-1711), Chénier (1762-1794), Hugo et Verlaine.

De manière systématique, des extractions ont été réalisées sur l'ensemble des œuvres de ces trois auteurs en s'appuyant sur des configurations particulières de ponctuations au sein de couples de vers consécutifs. L'outil BaseX a permis de mettre en œuvre des requêtes Xquery exploitant l'encodage XML TEI du corpus pour repérer les paires de vers consécutifs répondant à une distribution de ponctuations préétablie. Les sorties ont été formatées en HTML, avec retour à la source, afin d'en faciliter l'analyse.

L'analyse de corpus

L'analyse du corpus sera présentée dans la troisième partie, limitée pour l'instant à Boileau et Chénier. Elle nous a permis de dégager un bon nombre de paramètres à l'œuvre, certains plus ou moins connus, d'autres inédits. La diversité de nature de ces paramètres (culturels, biologiques, neuronaux, psychologiques et linguistiques) complique le travail. Pour s'en tenir au seul aspect linguistique, chaque composante linguistique se trouve impliquée de manière spécifique et dans ses relations avec les autres composantes, ce qui rend difficile la tâche d'établir des contraintes hiérarchisées.

Dans une quatrième partie, nous comparons les deux auteurs afin d'assurer une perspective historique. La première phase de ce travail a déjà permis de mettre en évidence des éléments d'évolution entre Boileau et Chénier sur lesquels s'appuiera Hugo.

La conclusion, toute provisoire, tentera de rappeler quelques aspects utiles de ce travail mais s'attachera surtout à souligner les perspectives du projet.

14h45 - 15h30

Traitement automatique de textes versifiés : Bilan et poursuite

- Richard Renault, enseignant-chercheur (CRISCO)
- Stéphane Ferrari, enseignant-chercheur (CRISCO)

Ce projet est constitué d'un corpus de textes versifiés, de programmes d'analyse métrique et d'une base de données de relevés métriques générés automatiquement. Le projet est relativement ancien (débuté en 2007 avec E. Delente) et toujours actif. La présentation du projet mettra l'accent sur la poursuite du projet ; amélioration et extension du traitement automatique, et intégration de données linguistiques permettant l'étude de la convergence/divergence entre unités métriques et structuration syntaxique du texte.

15h30 - 16h00 : Pause café

16h00 - 16h45

La modélisation graphique de la polysémie évolutive à partir des entrées historiques du TLFi

- Laurette Chardon, ingénieure d'études (CRISCO)
- Justine Reynaud, enseignante-chercheuse (GREYC)
- Jacques François, professeur émérite (CRISCO)

Les seize volumes du *Trésor de la Langue Française,* le dictionnaire de langue le plus riche de toute l'histoire de la lexicographie du français, ont été publiés par l'*Institut National de la Langue Française* (INaLF, Nancy, CNRS et DGLF) entre 1971 et 1994. Dès la parution du dernier volume, l'informatisation du dictionnaire a été engagée par rétroconversion de l'édition papier (l'édition électronique du *Grand Robert* datait déjà de 1985) et poursuivie à partir du début du 21e siècle par le laboratoire qui a succédé à l'INALF, l'ATILF (*Analyse et Traitement Informatique de la Langue Française).*

Les articles du TLFi se composent d'une entrée lexicographique (renseignée notamment à l'aide de la base de données textuelles FRANTEXT constituée simultanément), de données phonétiques, orthographiques et statistiques, et d'une entrée « Étymologie et histoire ». Les entrées lexicographiques ont bénéficié d'une informatisation fonctionnelle (attribuant une fonction à chaque segment et permettant ainsi tout un jeu de recherches transversales), mais l'ATILF a renoncé à en faire de même pour les entrées historico-étymologiques (H-É), sans doute avec la conviction que le formatage de ces entrées était trop hétérogène, et s'est contenté d'une informatisation formelle, c'est-à-dire réduite à la délimitation des entrées et au format typographique. De ce fait, il est actuellement impossible de pratiquer sur les entrées H-É des recherches transversales similaires à celles que permettent les entrées lexicographiques.

Le projet qui sera exposé à la journée de l'axe 2 du CRISCO vise à convertir dans un format tabulaire l'essentiel des données fournies par les entrées H-É du TLFi dotées d'une « polysémie évolutive » (± 20 000 sur un total de ± 49 000), de leur associer un graphe historique (arborescent) et de publier dans un premier temps sur le site du CRISCO un banc d'essai sous la forme d'un couple graphe-tableau historique associé à chacune des entrées de 7 des 81 fichiers XML que nous a fournis l'ATILF, couvrant ainsi presque 10% du total des entrées. À moyen terme il s'agit de poursuivre l'informatisation fonctionnelle des ± 20 000 entrées à l'étude et de la soumettre pour publication dans le cadre des éditions électroniques du CNRTL (*Centre National de Ressources Textuelles et Lexicales,* CNRS). À long terme, la base de données ainsi constituée permettra des recherches transversales complétant celles sur les entrées lexicographiques.

Les bases linguistiques de ce projet ont été élaborées en 2020-21 par Jacques François. Justine Reynaud en a développé la dimension statistique et Laurette Chardon la préparation des tableaux de données et la modalisation graphique dans le prolongement de l'étude préliminaire de Triss Jacquiot (GREYC, 1er semestre 2022).

Références

Bernard P. / Dendien J. / Pierrel J.M. (2004), A computerized dictionary : Le Trésor de la langue française informatisé (TLFi). In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pp. 40–43, Geneva, Switzerland. COLING.

Chardon L. / François J. / Reynaud J. (2023), La polysémie évolutive du lexique français (XIIe-XXe siècle) - Projet d'informatisation fonctionnelle et de modélisation graphique des

entrées historiques du TLFi. Cahier du CRISCO n° 37 (Lien à venir sur la page web du CRISCO)

Dendien J. / Pierrel J.M. (2003), Le Trésor de la langue française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. In *Les dictionnaires électroniques*, M. Zock et J. Carroll (eds), *Traitement Automatique des langues*, Vol 44 – n° 2/2003 : 11-37.

François J. (2020), Pour un retraitement informatisé et dynamique des notices historiques du TLFi, *Cahiers de lexicologie* Varia, n° 117, 2020–2 : 55-92

François J. (2021a), Comment visualiser l'évolution historique des polysémies lexicales : l'itinéraire sémantique de *TERRE* et *MONDE*. *Zeitschrift für romanische Philologie* 137(3): 625–665

François J. (2021b), Les fluctuations historiques de la polysémie lexicale, *Travaux de linguistique* 2020/2 n° 81 : 57-98

Mazziotta N., / François J. / Kahane S. (dir. à paraître, 2023), Les diagrammes en sciences du langage. *Travaux de Linguistique*, n° thématique 1er semestre 2023. Éditions De Boeck Supérieur.

16h40 - 17h00 : mots conclusifs