Construire et interroger les corpus numériques à l'ère de l'intelligence artificielle

Colloque international

Faculté des Sciences Humaines et Sociales de Tunis – Université de Tunis

Laboratoire *Intersignes* (LR14ES01)

29 Février - 1er Mars 2024

La Faculté des Sciences Humaines et Sociales de Tunis et le Laboratoire *Intersignes* (LR14ES01) organisent un colloque international intitulé "Construire et interroger les corpus à l'ère de l'intelligence artificielle".

Les bases de données textuelles, qu'elles soient littéraires, journalistiques, juridiques ou orales, ouvrent la voie à ce que l'on appelle « la linguistique de corpus » et constituent un outil majeur dans les travaux de recherche et un support didactique dans l'enseignement des langues et cultures. Le besoin de constituer des bases de données et des corpus, de les explorer ou de les publier se fait ressentir de plus en plus aussi bien dans le domaine linguistique, que dans celui des sciences humaines et sociales ou dans les divers secteurs socio-économiques.

Si, comme le précise E. Brunet, "un corpus est toujours artificiel. La nature n'en produit pas spontanément", il est important de distinguer les corpus des amas de textes numériques, disponibles et accessibles sur l'espace virtuel. Ainsi à partir d'un "corpus de référence" (Rastier 2017), un chercheur ou un enseignant vise à construire un ou son "corpus de travail", ce qui peut être considéré comme la première étape du processus d'interprétation.

De plus en plus d'équipes de recherche pluridisciplinaires formées d'ingénieurs en informatique et de spécialistes en linguistique et/ou en littérature s'intéressent à l'exploitation des corpus numérisés, du recueil au traitement. Leur travail consiste de ce fait à :

- scanner, relire et catégoriser les textes,
- fournir une version numérisée des grands textes littéraires,

- constituer des sous-corpus ou des corpus personnalisés selon les attentes du chercheur et selon les critères d'identification qu'il a définis,
- comparer des versions successives d'une même œuvre (comme les manuscrits de Proust ou de Balzac) en appliquant l'analyse génétique des textes.

Ces bases de données sont destinées à plusieurs usages auprès des chercheurs et des enseignants. Des recherches sur les spécificités scripturales d'un écrivain ou sur les propriétés du discours (littéraire, politique, historique, etc.), peuvent, dans cette perspective, être menées. L'exploration, via des logiciels de textométrie ou des concordanciers, comme AntConc, TROPES ou TXM va de la simple visualisation des occurrences au traitement statistique des fréquences. (François J., Gherissi Y. (2012)

Dans le cas spécifique de la langue française, et à partir, par exemple, de la base de données textuelles Frantext (www.frantext.fr), des analyses diachroniques allant de l'ancien français jusqu'au français contemporain peuvent être menées. Les études portant sur l'époque actuelle permettent, en outre, de comparer des ressources littéraires, journalistiques (François J., (2019) politiques, orales (ESLO, Université d'Orléans), etc.

En interrogeant les corpus de textes aussi bien sur le plan quantitatif que qualitatif, en décelant *le grain et la mesure* selon l'expression de Rastier (Rastier, 2011), les outils informatiques et les logiciels font émerger de nouveaux observables. L'intérêt des données numériques réside dans l'élaboration de nouvelles hypothèses de travail et dans l'adoption d'une méthodologie réflexive de manière à infirmer ou confirmer ces hypothèses (Rastier 2017).

Ce colloque se veut :

- l'occasion de répondre aux attentes des étudiants, des enseignants et des chercheurs, linguistes ou littéraires, qui veulent soit s'initier aux travaux sur les corpus, soit consolider leurs acquis et leurs apprentissages, ou encore confronter les différentes méthodes et techniques de travail. Ce qui

constituerait le point de départ pour l'initiation d'un axe de recherche et d'enseignement autour des Humanités numériques et de l'analyse des corpus ;

- une rencontre de chercheurs de diverses disciplines autour de la question des corpus textuels, de leurs modes de construction et des implications culturelles, éthiques et épistémologiques qu'ils posent. Quelle visibilité et quelles représentations offre aujourd'hui « l'architexte » qu'est le web aux différentes cultures ? Quel rôle peut-on et doit-on jouer pour agir sur le savoir répandu par les intelligences artificielles et les représentations du monde qu'il offre ? Nous aspirons à rendre visibles les travaux menés en Tunisie et ailleurs sur des corpus francophones ou arabophones, locaux ou étrangers et sur les outils disponibles en traitement des langues et en traduction ainsi que des usages didactiques possibles ;

- une réflexion sur les actions à mener pour un usage optimal des bases de données, des corpus et des intelligences artificielles dans la recherche et l'enseignement. L'objectif étant de pallier à une faille épistémologique et technique : au moment où certains se posent des questions d'actualité sur le rôle à jouer dans la constitution de corpus et la visibilité des savoirs à l'ère de ChatGPt et autres intelligences artificielles, d'autres continuent à travailler de manière plus traditionnelle et s'inquiètent des risques que constituent ces nouveaux outils.

Les communications pourraient s'articuler autour de 4 axes :

- Linguistique et analyse de corpus (en synchronie ou en diachronie) (François, 2010, 2019, Rastier 2011)

- Corpus littéraires (les mots d'auteurs, études contrastives, bases de données textuelles, lexicométrie, textométrie, etc. (Brunet 1981, Bernard et Bohet 2017 ; Legallois, 2016,)

- Enseignement et analyse de corpus (outils pour la classe, outils pour le FLE, FOS, FOU, etc.)(Cavalla, 2021)

- Traduction à l'ère du numérique : valorisation du patrimoine, visibilité de textes inédits, etc. (Desjardins R., Larsonneur C., Lacour Ph., 2021)

Les propositions de communications pourront prendre ainsi la forme de contributions critiques, de comptes rendus d'expériences ou d'ateliers d'initiation et d'apprentissage.

Merci d'envoyer votre proposition de communication (arabe/ français/ anglais) (entre 200 et 300 mots) accompagnée d'un titre, ainsi que d'une notice bio-bibliographique (précisant, entre autres, votre université, laboratoire et/ou unité de recherche de rattachement) à l'adresse suivante : colloque.corpusnumerique@gmail.com

Calendrier :

- Date limite de soumission des résumés : **15 octobre 2023**
- Date de notification d'acceptation : **15 novembre 2023**
- Date limite d'envoi des communications complètes : **31 Janvier 2024**
- Date de la tenue du colloque : **29 février- 1er mars 2024**

Le colloque se déroulera à la Faculté des Sciences Humaines et Sociales de Tunis, Tunis.

Les actes du colloque sous forme d'ouvrage collectif seront publiés courant 2024.

Comité scientifique :

- Jamil CHAKER (Université de Tunis, Faculté des Sciences Humaines et Sociales de Tunis- Laboratoire *Intersignes*)

- Sonia FITOURI-ZLITNI (Université de Tunis, Faculté des Sciences Humaines et Sociales de Tunis- Laboratoire *Intersignes*)

- Jacques FRANÇOIS (Université de Caen, Basse-Normandie)

- Samia KASSAB-CHARFI (Université de Tunis, Faculté des Sciences Humaines et Sociales de Tunis- Laboratoire *Intersignes*)

- Samir LABIDI (Académie militaire)

- Frédéric LANDRAGIN (CNRS, Laboratoire *LATTICE*)

- Dominique LEGALLOIS (Université Sorbonne-Nouvelle, Paris III)

- Badreddine HAMMA (Université d'Orléans, Laboratoire Ligérien de Linguistique)

- Yaacoub GHERISSI (Université de Carthage, Institut Supérieur des Langues de Tunis)

- François RASTIER, (CNRS)

Comité d'organisation :

- Dorra BASSI (Université de Tunis, Faculté des Sciences Humaines et Sociales de Tunis- Laboratoire *Intersignes*)

- Raja GMIR (Université de Tunis, Faculté des Sciences Humaines et Sociales de Tunis- Laboratoire *Intersignes*)

- Rania SAMET (Université de Tunis, Faculté des Sciences Humaines et Sociales de Tunis- Laboratoire *Intersignes)*

Références bibliographiques

Bernard M. Bohet B., (2017), *Littérométrie. Outils numériques pour l'analyse des textes littéraires*, Paris; Presses de la Sorbonne Nouvelle.

Brunet E. (1981), *Le vocabulaire français de 1789 à nos jours*: 3 volumes, Paris, Slatkine.

Cavalla C., (2021), "La formation de futurs enseignants de FLE à la phraséologie et aux corpus numériques", *in* Phrasis - rivista di studi fraseologici e paremiologici.

La revue *CORPUS,* notamment :

- n° 5 , 2006, *Corpus et stylistique,* *(https://journals.openedition.org/corpus/422)*
- n°8, 2009, "*Corpus de textes, textes en corpus*, (https://journals.openedition.org/corpus/1670)
- n°18 (2018), *Les petits corpus* (https://journals.openedition.org/corpus/3094)
- n° 24, 2023, *Les corpus numériques pour la didactique des langues : de la formation des enseignants à l'élaboration de dispositifs d'apprentissage* *https://journals.openedition.org/corpus/7438*

Desjardins R., Larsonneur C., Lacour Ph,(2021), When Translation Goes Digital: Cases Studies and Critical Reflections, éd. Polgrave Macmillan, mars 2021.

Eshkol-Taravella I, Lefeuvre-Halftermeyer A., (2017), Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage, (https://doi.org/10.4000/corela.4797)

François J., (2010), "L'attestation des combinaisons lexicales à l'aide de la base de données textuelles FRANTEXT", in *Cahier du CRISCO*, N°29, Université de Caen, Basse-Normandie, https://hal.archives-ouvertes.fr/hal-01834424 , [consulté le 25 juin 2023]

François J., (2019), Didacticiel FRANTEXT-2, https://www.interlingua.fr/didacticiel-frantext-2/, [consulté le 25 juin 2023]

François J., Gherissi Y., (2012), Pour une linguistique orientée outils : la polysémie du verbe compter et les genres, Cahier du CRISCO, N°34, Université de Caen, Basse-Normandie. (https://hal.archives-ouvertes.fr/hal-01811292)

Garric N. & Longhi J. (2012). Analyse de corpus face à l'hétérogénéité des données. *Langages* 2012/3 (n° 187)

Gmir R., (2021), Connaître: de la désémantisation à la constructionnalisation, in Syntaxe et Sémantique, N°22, *L'expansion pluridisciplinaire des grammaires de construction*, pp 97-120

Habert B., (2005), *Instruments et ressources électroniques pour le français*. Paris /Gap : Ophrys

Habert B., (2009), *Construire des bases de données pour le français : Tome 1*, Notions: Paris /Gap : Ophrys

Larsonneur, C. (2021), « Intelligence artificielle ET/OU diversité linguistique : les paradoxes du traitement automatique des langues », *Hybrid* [En ligne], 7 | 2021, mis en ligne le 08 avril 2021, URL : http://journals.openedition.org/hybrid/650 ; DOI : https://doi.org/10.4000/hybrid.650 [consulté le 25/06/2023]

Legallois D., Charnois Th. et Poibeau Th. (2016), "Repérer les clichés dans les romans sentimentaux grâce à la méthode des « motifs »" *in* : Frédérique Sitri et Agnès Tutin (dir.), LIDIL 53, *Phraséologie et genres de discours : Patrons, motifs, routines*, p. 95-117

Poudat, C., Landragin F. (2017), *Explorer un corpus textuel : méthodes, pratiques, outils*, Paris, De Boeck.

Quiniou S., Cellier P., Charnois Th., Legallois D., (2012), "Fouille de données pour la stylistique : cas des motifs séquentiels émergents", Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12), Liège, Belgique. pp.821-833. [ffhal-00675586f]

Silberztein, M., (2019) "Les outils informatiques au service des linguistes : présentation" , Dans *Langue française* 2019/3 (N° 203), pages 7 à 14

Rastier F. (dir.) (1987), *Sémantique et Intelligence artificielle*, Langages, 87.

Rastier F. (dir.) (1995) : *L'analyse thématique des données textuelles – L'exemple des sentiment*s, Paris, Didier.

Rastier F.,(2000), « L'accès aux banques textuelles : des genres à la doxa », *in* Cabré T. et Gelpi, C., éd., *Lèxic, corpus i diccionaris*, Cicle de confèrencies i seminaris 97-98, IULA, Université Pompeu Fabra, Barcelone.

Rastier F., (2001), « Genres et variations morphosyntaxiques », *Traitements automatiques du langage*, 42, 2, 2001, pp. 547-577. En collaboration avec Denise Malrieu.

Rastier F. (2011) : *La mesure et le grain. Sémantique de corpus*, Champion, Paris.

Rastier F., (2017), *Corpus numériques et accès à la culture*, Questions Vives [En ligne], N° 28 | 2017, mis en ligne le 06 novembre 2018, URL : http://journals.openedition.org/questionsvives/2421 ; DOI : https://doi.org/10.4000/questionsvives.2421 [consulté le 13 juin 2023]

Rastier F. (2021), « Data vs Corpora », *in L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, sous la direction de Damon MAYAFFRE et Laurent VANNI, Paris, Champion, 2021, pp. 203-249.

Samet R., (2021), Les constructions collocatives à noms prédicatifs invariants, *Syntaxe et Sémantique du Français ; Actes de la journée scientifique du 12 Avril 2019 à l'Université de Carthage (ISLT) en hommage au Professeur Jacques François*, Les Éditions SAHAR-ISLT ÉDITION.

Zufferey S., (2020), *Introduction à la linguistique de corpus*, Genève : ISTE éditions.