English version below, p. 4

La constitution de corpus en diachronie longue, entre tradition philologique et analyse quantitative

ConCorDiaL 2024, Lyon

7-8 novembre 2024

Appel à communications

Depuis ses origines, la linguistique diachronique entretient des liens intimes avec la linguistique de corpus, les diachroniciens ne pouvant par définition faire appel à leur compétence de locuteurs et devant s'appuyer sur des données attestées et authentiques pour travailler (Marchello-Nizia 2004, Prévost 2020). Les corpus numériques diachroniques et/ou de textes anciens se sont ainsi considérablement développés, et en France, Frantext et la Base de français médiéval ont joué un rôle moteur dans ce mouvement initié dans les années 1980. Ces corpus se sont généralement construits sur des éditions imprimées et ont connu un développement distinct de celui des éditions numériques natives, plus focalisées sur la transposition numérique de l'apparat critique et de la représentation des sources primaires souvent manuscrites dans une perspective philologique. Cette dissociation, qui recouvre en partie les frontières disciplinaires entre linguistes et littéraires, s'est notamment traduite en France par la création de deux consortiums distincts dans le cadre de l'infrastructure nationale de recherche Huma-Num, à savoir un consortium pour les corpus linguistiques (aujourd'hui CORLI, *CORpus, Langues et Interactions*) et un autre pour les éditions de texte et l'analyse littéraire et stylistique (aujourd'hui ARIANE, *Analyses, Recherches, Intelligence Artificielle et Nouvelles Éditions numériques*). On peut se demander si la dynamique actuelle des approches quantitatives en littérature (Bernard et Bohet 2017, Diwersy *et al.* 2021, Barré, Camps et Poibeau 2023) et si la création de données linguistiques nouvelles sous forme numérique ne rendent pas cette bi-partition désormais en partie artificielle.

L'essor exponentiel des corpus numériques est par ailleurs à l'origine d'une surabondance, voire d'un « déluge » de données (Habert 2005 : 41), et même si cette tendance est moins forte dans le cas des corpus de langues anciennes – l'accès aux données primaires n'étant pas aussi immédiat que pour les données langagières contemporaines –, ces corpus ne cessent de grossir en taille et en diversité. Les outils permettant le traitement, l'annotation et l'interrogation des textes, ont en parallèle considérablement enrichi les corpus textuels et leur exploitation numérique. Toujours plus gourmands en données (*cf.* les avancées récentes de l'IA et des agents conversationnels), les outils du Traitement automatique des langues ne font qu'amplifier la demande d'accroissement et favorisent en même temps le développement des méthodes statistiques en linguistique de corpus et dans l'analyse des données textuelles (Lebart, Pincemin et Poudat 2019).

C'est dans ce contexte et dans la lignée du premier colloque ConCorDial (Grenoble 2022, https://concordial2022.sciencesconf.org) que cette seconde édition propose d'approfondir la réflexion sur les corpus numériques en diachronie longue, en articulant constitution et analyse de corpus et en poursuivant les échanges entre créateurs et utilisateurs de données langagières.

Axe 1 : Traitement des corpus numériques diachroniques

L'accumulation de données numériques oblige à faire face au défi de leur hétérogénéité interne. Cette hétérogénéité dérive de la diversité des sources qui peuvent avoir différentes provenances avant d'être réunies dans un corpus particulier. Elle peut concerner aussi bien la qualité de numérisation des textes, que leur format numérique (XML ou autre), les métadonnées qui permettent de les décrire et bien sûr aussi leurs annotations linguistiques. À ces facteurs généraux, peuvent s'ajouter pour les périodes les plus anciennes les variations graphiques et morphologiques qui compliquent la reconnaissance des formes et le travail des outils de TAL. On pourra s'intéresser aux différentes façons de traiter cette hétérogénéité en fonction des usages escomptés et des contraintes (techniques, financières, etc.) qui s'imposent.

Ces questions pourront également être abordées sous l'angle de la compatibilité et de l'interopérabilité entre différents corpus. Les référentiels communs (concernant les balises, les métadonnées, la segmentation lexicale, les lemmes, les jeux d'étiquettes morphosyntaxiques, les annotations syntaxiques ou sémantiques, etc.) sont une manière de répondre à cet objectif qui devient de plus en plus nécessaire à mesure que les corpus se multiplient. Dans ce cadre, les enjeux de la perennité et de la sauvegarde des données sont également à prendre en compte. On pourra notamment se demander comment concilier une exigence de normalisation avec le respect de la diversité et de la richesse des données d'origine (comment, par exemple, utiliser un jeu d'étiquettes multilingue sans appauvrir l'étiquetage d'une langue particulière ?).

La dimension historique sur le temps long pourra faire l'objet d'une réflexion spécifique, la variation diachronique étant d'autant plus importante que le corpus couvre une vaste période et se manifestant à tous les paliers de traitement. Comment gérer les évolutions qui touchent les genres textuels (apparitions/disparitions, évolutions à l'intérieur d'un genre donné, les genres étant historiquement situés et évoluant dans le temps, cf. Winter-Froemel 2023) ? Doit-on utiliser les mêmes lemmes quelle que soit la période ou se fonder sur des dictionnaires propres à chaque état de langue ? Comment traiter les changements dans la segmentation en unités lexicales et l'émergence de locutions grammaticalisées ?

Les questions soulevées ici ne sont pas exhaustives et toutes les propositions de communication abordant la constitution et le traitement de corpus diachroniques seront examinées.

Axe 2 : Méthodes quantitatives et qualitatives pour l'exploitation de corpus diachroniques

Les méthodes quantitatives étant de plus en plus utilisées sur tous les plans de l'analyse linguistique (lexique, phonologie, morphologie, syntaxe, etc.) et se diffusant dans le champ des études stylistiques (stylèmes, phraséologismes) et littéraires (topiques, motifs narratifs, etc.), on pourra interroger leur impact sur les corpus numériques diachroniques : comment tenir compte de ces usages dans la sélection, la préparation, la description et l'organisation des données ? quelles méthodes et quels outils employer pour le repérage et l'interprétation quantitative des données ?

Dans ce cadre, on pourra s'intéresser plus spécifiquement aux apports et aux limites de l'annotation linguistique et se demander quels types d'enrichissements privilégier pour faciliter les recherches diachroniques, quel niveau de granularité adopter, quel équilibre viser entre quantité et qualité des annotations, etc.

Les méthodologies quantitatives spécifiquement adaptées à l'analyse diachronique feront l'objet d'une attention particulière. On pourra notamment traiter des différents types de variation, des spécificités du facteur diachronique ou des manières de cibler ce facteur particulier ou au contraire de décrire la façon dont il interagit avec d'autres (Hilpert et Gries 2016). De même, les nouvelles possibilités offertes par les outils de périodisation automatique (Gries et Hilpert 2008, Diwersy *et al.* 2017), ou les méthodes permettant de mesurer et d'interpréter des tendances (Hilpert et Gries 2009), etc. pourront être présentées.

L'articulation entre méthodes quantitatives et analyse qualitative sera également prise en compte, de même que la dimension philologique des données construites pour une exploitation linguistique ou littéraire.

Conférences invitées

- Sascha Diwersy (Université Montpellier, UMR Praxiling)
- Thierry Poibeau (CNRS, UMR Lattice)
- Céline Poudat (Université Côte d'Azur, UMR BLC)

Modalités

La durée des présentations sera de 30 minutes suivies d'une discussion de 10 minutes. Le colloque se déroulera en mode hybride (présentiel souhaité pour les intervenants). Les langues de communication acceptées sont le français et l'anglais.

Les résumés doivent comprendre entre 300 et 500 mots (sans compter les références bibliographiques) et seront rédigés dans la langue de communication. Ils doivent être déposés sur le site de la conférence (https://concordial.sciencesconf.org) en deux versions : une version anonymisée (à copier-coller dans le formulaire) et une version précisant le nom et l'affiliation de l'auteur ou des auteurs dans un document Word ou PDF. Merci d'utiliser le modèle de document proposé.

Frais d'inscription

Les frais d'inscription seront communiqués à l'ouverture de l'inscription (entre 40 et 60 €).
Exonération :

- participants en ligne
- membres des laboratoires organisateurs
- doctorants

Calendrier

- Date limite de soumission de résumé : 15 mai 2024
- Retour des évaluations : 1er juillet 2024
- Soumission de la version définitive des résumés : 1er octobre 2024
- Inscription au colloque : du 1er septembre au 1er octobre
- Colloque : du 7 au 8 novembre 2024

Corpus Building in Long Diachrony, between Philological Tradition and Quantitative Analysis

ConCorDiaL 2024, Lyon

November 7-8, 2024

Call for papers

Since its origins, diachronic linguistics has maintained close links with corpus linguistics, since diachronicists cannot, by definition, call on their competence as speakers and must rely on attested and authentic data to work. (Marchello-Nizia 2004, Prévost 2020). There has been significant development of digital diachronic corpora and historical texts in corpus linguistics and, as far as France is concerned, Frantext database and the BFM Old French Corpus have played a leading role in this movement, which began in the 1980s. These corpora have generally been built on printed editions, and have developed separately from native digital editions, which are more focused on the digital transposition of critical apparatus and the representation of primary sources, often handwritten, from a philological perspective. In France, this dissociation, which partly reflects the disciplinary boundaries between linguists and literary scholars, has resulted in the creation of two distinct consortia within the Huma-Num national research infrastructure: one for linguistic corpora (now CORLI, *CORpus, Langues et Interactions*) and the other for text editions and literary and stylistic analysis (now ARIANE, *Analyses, Recherches, Intelligence Artificielle et Nouvelles Éditions numériques*). In the light of the current dynamic of quantitative approaches to literature (Bernard and Bohet 2017, Diwersy *et al.* 2021, Barré, Camps and Poibeau 2023) and the creation of new digitally native linguistic data the relevance of this partition can be brought into question.

The exponential growth of digital corpora has also led to an overabundance, even a "deluge" of data (Habert 2005: 41). Although this trend is less marked in the case of historical language corpora - since access to primary data is not as immediate as for contemporary language data - these corpora continue to grow in size and diversity. At the same time, tools for processing, annotating and querying texts have considerably enriched textual corpora and their digital exploitation. Ever more data-intensive (*cf.* recent advances in AI and conversational agents), the tools of Natural Language Processing amplify the demand for growth, and at the same time foster the development of statistical methods in corpus linguistics and textual data analysis (Lebart, Pincemin and Poudat 2019).

In this context, and following on from the first ConCorDial conference (Grenoble 2022, https://concordial2022.sciencesconf.org), this second edition aims to deepen our reflection on digital corpora in long diachrony, by linking corpus creation and analysis, and continuing exchanges between creators and users of language data.

Topic 1: Processing diachronic digital corpora

The accumulation of digital data makes it necessary to face up to the challenge of their internal heterogeneity. This heterogeneity derives from the diversity of sources, which may have different origins before being brought together in a particular corpus. It may concern the quality of the digitized texts, their digital format (XML or other), the metadata used to describe them and, of course, their linguistic annotations. In addition to these general factors, for the earliest

periods we can add graphic and morphological variations that complicate form recognition and the work of NLP tools. Contributions could address different ways of dealing with this heterogeneity, depending on both the intended use of the corpus and the constraints (technical, financial, etc.) imposed.

These issues can also be addressed from the point of view of compatibility and interoperability between different corpora. Common standards (for markup tags, metadata, word segmentation, lemmas, morphosyntactic tag sets, syntactic or semantic annotations, etc.) are one way of meeting this objective, which is becoming increasingly necessary as corpora multiply. In this context, we also need to take into account the issues of data durability and backup. For example, how can the need for standardization be reconciled with respect for the diversity and richness of the original data: can a multilingual tag set even be used without compromising the granularity of the tags necessary for a particular language?)

The long-term historical dimension may be the subject of specific reflection, diachronic variations being all the more important as the corpora cover vast timespans and are observable at all levels of processing. How can the evolution of textual genres, such as appearances/disappearances, changes within a given genre, genres being historically situated and evolving over time, cf. Winter-Froemel 2023) properly be taken into account? Should the same lemmas be used whatever the period, or should dictionaries specific to each language state be preferred? How should we deal with changes in the segmentation of lexical units and the emergence of grammaticalized phrases?

The questions raised here are not exhaustive, and all proposals for papers dealing with the constitution and processing of diachronic corpora will be considered.

Topic 2: Quantitative and qualitative methods for exploiting diachronic corpora

As quantitative methods are increasingly used in all areas of linguistic analysis (lexicology, phonology, morphology, syntax, etc.), and are spreading into the field of stylistic studies (stylemes, phrasemes) and literary studies (topics, narrative patterns, etc.), their impact on diachronic digital corpora can be examined. How are these practices into account in the selection, curation, description and organisation of the data? What methods and tools should be used to identify and quantitatively interpret the data?

In this context, we seek contributions analysing the added value as well as the limitations of linguistic annotation, and ask what types of enrichment should be favoured to facilitate diachronic research, what level of granularity should be adopted, what balance should be aimed for between the quantity and quality of annotations, etc?

Particular attention could be paid to quantitative methodologies specifically adapted to diachronic analysis. In particular, it will be possible to address the different types of variation, the specificities of the diachronic factor or the ways of targeting this particular factor or, on the contrary, describing the way it interacts with others (Hilpert and Gries 2016). Similarly, contributions considering the new possibilities offered by automatic periodization tools (Gries and Hilpert 2008, Diwersy *et al.* 2017) or methods for measuring and interpreting trends (Hilpert and Gries 2009) etc. are particularly welcome.

The link between quantitative methods and qualitative analysis will also be taken into account, as will the philological dimension of data constructed for linguistic or literary research.

Keynote speakers

- Sascha Diwersy (Montpellier University, UMR Praxiling)
- Thierry Poibeau (CNRS, UMR Lattice)
- Céline Poudat (Université Côte d'Azur, UMR BLC)

Format

Presentations will last 30 minutes, followed by a 10-minute discussion. The conference will be held in hybrid mode (in-person attendance preferred for speakers). The languages accepted for communication are French and English.

Abstracts should be between 300 and 500 words in length (not including bibliographical references) and should be written in the language of the paper. Abstracts must be submitted in two versions on the conference website (https://concordial.sciencesconf.org): an anonymized version to copy-paste in the submission form and a version specifying the author's name and affiliation in a Word or PDF document. Please use the provided document template.

Registration fees

Registration fees will be confirmed when registration opens (between €40 and €60).

Exemption :

- online participants
- members of the organizing laboratories
- doctoral students

Calendar

- Abstract submission deadline: May 15, 2024
- Confirmation of acceptance: July 1st 2024
- Submission of final abstracts: October 1st 2024
- Conference registration: from September 1st to October 1st
- Conference: November 7-8, 2024

Références / References

BARRÉ Jean, CAMPS Jean-Baptiste et POIBEAU Thierry (2023) « Operationalizing Canonicity: A Quantitative Study of French 19th and 20th Century Literature », *Journal of Cultural Analytics*, vol. 8, n° 3. ‹DOI : 10.22148/001c.88113›.

BERNARD Michel et BOHET Baptiste (2017) *Littérométrie : outils numériques pour l'analyse des textes littéraires*, Paris, Presses Sorbonne nouvelle.

DIWERSY Sascha, FALAISE Achille, LAY Marie-Hélène et SOUVAY Gilles (2017) « Ressources et méthodes pour l'analyse diachronique », *Langages*, vol. 206, n° 2, p. 21-44. ‹DOI : 10.3917/lang.206.0021›.

DIWERSY Sascha, GONON Laetitia, GOOSSENS Vannina, *et al.* (2021) « La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase », *Corpus*, n° 22. ‹DOI : 10.4000/corpus.6101›.

GRIES Stefan et HILPERT Martin (2008) « The identification of stages in diachronic data: variability-based neighbour clustering », *Corpora*, vol. 3, p. 59-81. ‹DOI : 10.3366/E1749503208000075›.

HABERT Benoît (2005) « Face à la disette dans la profusion », *Scolia : Sciences Cognitives, Linguistiques et Intelligence Artificielle*, vol. 19, n° 1, p. 41-61. ‹DOI : 10.3406/scoli.2005.1065›.

HILPERT Martin et GRIES Stefan (2009) « Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition », *Literary and Linguistic Computing*, vol. 24, n° 4, p. 385-401. ‹DOI : 10.1093/llc/fqn012›.

HILPERT Martin et GRIES Stefan (2016) « Quantitative approaches to diachronic corpus linguistics », In M. Kytö et P. Pahta (éd.), *The Cambridge Handbook of English Historical Linguistics*, Cambridge University Press, p. 36-53. ‹DOI : 10.1017/CBO9781139600231›.

LEBART Ludovic, PINCEMIN Bénédicte et POUDAT Céline (2019) *Analyse des données textuelles*, Québec, Presses de l'Université du Québec.

MARCHELLO-NIZIA Christiane (2004) « Linguistique historique, linguistique outillée : les fruits d'une tradition », *Le français moderne*, n° 1, p. 58-70.

PREVOST Sophie (2020) « Une grammaire fondée sur un corpus numérique », In C. Marchello-Nizia, B. Combettes, S. Prévost et T. Scheer (éd.), *Grande grammaire historique du français*, Berlin, Mouton de Gruyter, p. 37-53.

WINTER-FROEMEL Esme (2023) « Discourse traditions research: foundations, theoretical issues and implications », In E. Winter-Froemel et Á.S. Octavio de Toledo y Huerta (éd.), *Manual of Discourse Traditions in Romance*, De Gruyter, p. 25-58. ‹DOI : 10.1515/9783110668636-002›.

Comité scientifique / Programme committee

- Corinne Denoyelle (U. Grenoble Alpes, LIDILEM, France)
- Sascha Diwersy (U. de Montpellier, PRAXILING, France)
- Mathieu Goux (U. de Caen, CRISCO, France)
- Céline Guillot-Barbance (ÉNS de Lyon, IHRIM, France)
- Serge Heiden (ENS de Lyon, IHRIM, France)
- Olivier Kraif (U. Grenoble Alpes, LIDILEM, France)
- Alexei Lavrentiev, IHRIM, CNRS, France)
- Raphaël Luis (ÉNS de Lyon, CERCC, France)
- Jean-Philippe Magué (ÉNS de Lyon, ICAR, France)
- Sophie Marnette (U. d'Oxford, Royaume-Uni)
- Pascale Mounier (U. Grenoble Alpes, LIDILEM, France)

- Bénédicte Pincemin (IHRIM, CNRS, France)
- Céline Poudat (U. de Nice, BCL, France)
- Sophie Prévost (LATTICE, CNRS, France)
- Matthieu Quignard (ICAR, CNRS, France)
- Thomas Rainsford (U. de Stuttgart, Allemagne)
- Adam Renwick (U. Grenoble Alpes, LIDILEM, France)
- Amalia Rodríguez Somolinos (U. Complutense de Madrid, Espagne)
- Alexandra Simonenko (U. de Gand, Belgique)
- Carine Skupien Dekens (U. de Neuchâtel, Suisse)
- Julie Sorba (U. Grenoble Alpes, LIDILEM, France)
- Gilles Souvay (ATILF, CNRS, France)
- Denis Vigier (U. Lumière Lyon2, ICAR, France)

Comité d'organisation / Organising committee

- Céline Guillot-Barbance
- Alexei Lavrentiev
- Tanguy Lemoine
- Raphaël Luis
- Matthieu Quignard.