|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | | --- | | Dottorato in Civiltà  dell'Asia e dell'Africa | | |  | |

***Bruit de fond ou valeur ajoutée ?* Gérer le bruit lors des traitements informatiques des corpus linguistiques**

**Journée d'étude co-organisée à Grenoble par l'Université Grenoble Alpes et Roma La Sapienza**

L’influence croissante des méthodologies liées au TAL sur la linguistique de corpus oblige les chercheurs à réinterroger les pratiques de gestion du bruit et son impact dans les résultats de recherche (Fuchs & Habert, 2004 ; Léon, 2018 ; Zalmout et al., 2018). Qu’il s’agisse de corpus en diachronie longue (ex. français médiéval), de corpus dialectaux aux ressources limitées (ex. textes oraux ou écrits en arabe dialectal, cf. arabizi), ou encore de corpus de textes éloignés de la norme (ex. corpus d’apprenants), l’analyse du bruit est une étape nécessaire pour apprécier correctement la qualité des données de recherche (Molinelli & Putzu, 2015 ; Scaglione, 2018 ; Litosseliti, 2018). Cette journée d'étude sera l’occasion de réfléchir sur les méthodes de gestion du bruit dans les domaines du TAL et de la linguistique de corpus outillée, et à son impact sur la qualité des données linguistiques (Kraif et Ponton, 2007 ; Goutte et al., 2012 ; Zeroual, 2018).

Les questions sous-jacentes à toute étude linguistique concernent la définition de l'objet de recherche, la nature des données elles-mêmes, et la manière de préserver autant que possible leurs caractéristiques dans les différents traitements (lemmatisation, normalisation, etc.) (Sarrica et al., 2016). Ainsi, le choix des méthodes de gestion du bruit, de la phase de collecte à leur archivage, de la préparation des données à l'annotation, joue un rôle fondamental (Egbert & Baker, 2019). La journée d’étude stimulera une réflexion sur les pratiques des chercheurs pour réduire la portée des biais produits par le bruit et le silence, que ce soit durant la collecte, l’enregistrement ou l’annotation des données. De même, il s’agit de comprendre dans quelle mesure le bruit peut se révéler une source d’informations, notamment durant la phase d’annotation des corpus. A partir de quel seuil peut-on considérer le bruit comme acceptable ? comment différencier bruit et biais méthodologique ? comment estimer le bruit sans vérité de terrain ? Comment atteindre l’équilibre nécessaire pour que le bruit causé par les traitements des données ne compromette pas les résultats des recherches ? Pour tenter de répondre à ces questions, les personnes intéressées pourront soumettre une proposition de communication s’inscrivant dans l’un des trois axes ci- dessous :

1. **Le bruit pendant la collecte et l'enregistrement des données.** Si l’on accepte le postulat selon lequel « la donnée linguistique est un résultat » (Benveniste, 1966), comment décoder le bruit causé par le recueil des données et leur enregistrement ? En effet, en fonction des objets de recherche, il existe des facteurs potentiels de corruption des données, comme par exemple les croyances du chercheur, ou les biais introduits par un système OCR donné (Jentsch & Porada, 2020). L’enjeu consiste alors à prédire ou à déterminer les biais potentiels induits par ces facteurs lors de la sélection des données pour optimiser les phases de recherche successives.
2. **La préparation et le prétraitement des données**. Les méthodes choisies pour affiner les données brutes et les rendre disponibles pour des manipulations avancées peuvent représenter une importante source de bruit ou, au contraire, de silence : c’est notamment le cas du processus de normalisation des données (Al Sharou et al., 2021). Qu’il s’agisse de transcrire des données ou de corriger des erreurs, le chercheur fait des choix qui impactent nécessairement la nature des données, soit en les réduisant, soit en les enrichissant.
3. **Le processus d’annotation et les métadonnées**. À la base, l’annotation des corpus est un processus visant l’enrichissement des données : en fonction du modèle d’analyse mis au point, le chercheur tente de catégoriser des unités à travers un processus d’étiquetage (Péry-Woodley et al., 2011). Cependant, si d’un côté ce processus peut générer du bruit, de l’autre, il peut être une cause de silence fort préjudiciable aux résultats des recherches et à leur interprétation. La notion de *métadonnée* peut également être mise en cause : enrichir une donnée signifie-t-il la transformer en quelque chose d’autre ?

Cette journée d'étude, coorganisée par des jeunes chercheurs et des chercheurs expérimentés, s'adresse aux doctorant.e.s, jeunes chercheur.e.s et post-docs, ainsi qu'aux chercheur.e.s expérimenté.e.s. Les personnes souhaitant présenter leurs recherches sont priées d'envoyer un résumé en deux versions, l'une anonyme et l'autre avec le nom des auteurs, de 1000 mots maximum, références comprises, à [je-bruit-corpus@scienceconf.org](mailto:je-bruit-corpus@scienceconf.org) (voir le calendrier prévisionnel ci-dessous). Veuillez enregistrer les documents sous les noms suivants : "anonymous.doc" pour la version anonyme et "AUTHOR\_NAME.doc" pour la version identifiable par l'auteur. La langue d'échange de la journée d'étude sera de préférence le français (sans exclure l’italien et l’anglais) ; les propositions en italien et en anglais seront également acceptées. Toutes les propositions de communications feront l’objet d’une évaluation en double aveugle par les membres du comité scientifique de la journée d’étude. A l'issue de la journée d'étude, une publication dans un numéro thématique de revue sera envisagée.

L'inscription est gratuite.

# Calendrier prévisionnel

* Soumission : vendredi 25 novembre 2022
* Notification aux auteurs : vendredi 10 février 2023
* Soumission de la version finale : vendredi 24 mars 2023
* Journée d'étude : vendredi 28 avril 2023

***Rumore di fondo o valore aggiunto?* Gestire il rumore nell'elaborazione informatica dei corpora linguistici**

La progressiva influenza delle metodologie legate all’Elaborazione del Linguaggio Naturale (ELN) nella linguistica dei corpora sta portando un numero crescente di ricercatori a riesaminare le pratiche di gestione del rumore e del suo impatto sui risultati della ricerca (Fuchs & Habert, 2004; Léon, 2018; Zalmout et al., 2018). Che si tratti di corpora diacronici (e.g., di francese medievale), di corpora dialettali o di varietà con risorse limitate (e.g., arabo dialettale orale o scritto, cfr. arabizi), o di corpora di apprendenti, l'analisi del rumore è un passaggio necessario per valutare correttamente la qualità dei dati su cui si basa la ricerca (Molinelli & Putzu, 2015; Scaglione, 2018; Litosseliti, 2018). Questo workshop sarà un'occasione per riflettere sui metodi di gestione del rumore, in ELN e nella linguistica dei corpora, e sull’impatto che esso ha sulla qualità del dato linguistico (Kraif & Ponton, 2007; Goutte et al., 2012; Zeroual, 2018).

Alla base di qualsiasi studio linguistico c’è l’identificazione dell'oggetto di ricerca, la definizione della natura del dato e della metodologia atta a preservarne il più possibile le caratteristiche nelle fasi di elaborazione (lemmatizzazione, normalizzazione, ecc.) (Sarrica et al., 2016). Pertanto, le scelte metodologiche di gestione del rumore, dalla fase di raccolta all'archiviazione, dalla preparazione dei dati all’annotazione, gioca un ruolo fondamentale (Egbert & Baker, 2019). La giornata di studio incoraggerà la riflessione sul potenziale impatto del rumore nella fase di raccolta, registrazione o annotazione dei dati. Allo stesso modo, l'obiettivo è capire fino a che punto il rumore può essere una fonte di informazioni, in particolare durante la fase di annotazione dei corpora. Entro quali soglie il rumore dovuto all’elaborazione dei dati è tale da non compromettere i risultati della ricerca? Come si può distinguere il rumore dal bias metodologico? Come si può valutare il rumore senza disporre dei “dati di verità di base” (*ground truth*)? Per rispondere a queste domande, gli interessati possono presentare una proposta di contribuzione in una delle tre aree indicate di seguito:

1. **Il rumore durante la raccolta e la registrazione dei dati**. Se accettiamo il postulato che “Il dato linguistico è un risultato” (Benveniste, 1966) come possiamo interpretare il rumore generato in fase di raccolta o di registrazione dei dati? In effetti, a seconda dell’oggetto di ricerca, esistono potenziali fattori di corruzione dei dati, come le rappresentazioni mentali del ricercatore o le distorsioni introdotte da un determinato sistema OCR (Jentsch & Porada, 2020). La sfida consiste quindi nel prevedere o determinare i potenziali errori indotti da questi fattori durante la selezione dei dati, al fine di ottimizzare le fasi successive della ricerca.
2. **Preparazione e pre-elaborazione dei dati.** I metodi scelti per ripulire i dati naturali e renderli fruibili tramite sistemi informatici possono rappresentare una fonte importante di rumore o, al contrario, di silenzio: questo è il caso, in particolare, del processo di normalizzazione dei dati (Al Sharou et al., 2021). Sia che si tratti di trascrivere i dati o di correggerne gli errori, il ricercatore prende delle decisioni che hanno necessariamente un impatto sulla natura dei dati, operando una selezione o un arricchimento degli stessi.
3. **Il processo di annotazione e i metadati**. Fondamentalmente, l'annotazione dei corpora è un processo volto ad arricchire i dati: in base al modello di analisi adottato, il ricercatore cerca di categorizzare le unità attraverso un processo di etichettatura (Péry-Woodley et al., 2011). Tuttavia, se da un lato questo processo ha come obbiettivo l’arricchimento dei dati attraverso l’aggiunta di metadati, dall'altro introduce del rumore che può alterare l’interpretazione dei dati e i risultati della ricerca. Anche la nozione di *metadato* può quindi essere oggetto di discussione: categorizzare i dati significa trasformarli in qualcos'altro?

Questa giornata di studio, co-organizzata da ricercatori giovani ed esperti, è rivolta a dottorandi, giovani ricercatori e post-doc, nonché a ricercatori esperti. Coloro che desiderino presentare una loro ricerca sono pregati di inviare un abstract in due versioni, uno anonimo e l'altro riportante i nomi degli autori, di massimo 1000 parole (bibliografia inclusa), all’indirizzo [je-bruit-](mailto:je-bruit-corpus@sciencesconf.org)[corpus@sciencesconf.org](mailto:corpus@sciencesconf.org)(si veda il calendario provvisorio sotto). Si prega di salvare i documenti con i seguenti nomi “anonimo.doc” per la versione anonima e “NOME\_COGNOME.doc” per la versione riconducibile agli autori. La lingua della conferenza sarà preferibilmente il francese; saranno tuttavia accettate anche proposte in inglese e italiano. Tutte le proposte saranno sottoposte a un processo di doppia valutazione anonima da parte dei membri del comitato scientifico della giornata di studio. A completamento della giornata di studi, è prevista la pubblicazione in un numero tematico di rivista.

L'iscrizione è gratuita.

# Calendario provvisorio

* Presentazione dei contributi: venerdì 25 novembre 2022
* Notifica agli autori: venerdì 10 febbraio 2023
* Presentazione della versione finale dei contributi accolti: venerdì 24 marzo 2023
* Giornata di studio: venerdì 28 aprile 2023

# Composition du comité d’organisation

Elisa Gugliotta, LIG, UGA

Luca Pallanti, ECP, Université Lumière Lyon2

Olivier Kraif, LIDILEM, UGA

Iris Fabry, LIDILEM, UGA Martina Barletta, LIDILEM, UGA

# Composition du comité scientifique

Barletta Martina, LIDILEM, UGA

Casentini Marco, Ca’ Foscari, Italie

Coulange Sylvain, LIDILEM, LIG, UGA

Dinarelli Marco, CNRS, LIG, UGA

Fabry Iris, LIDILEM, UGA

Gugliotta Elisa, LIG, UGA

Jacques Marie-Paule, LIDILEM, UGA

Kraif Olivier, LIDILEM, UGA

Lancioni Giuliano, Università degli Studi di Roma Tre

Lepadat Carmen, Roma Tre, Italie

Mion Giuliano, Università degli Studi di Cagliari Nguyên Minh Châu, LIG, UGA

Ormaechea Lucia, Université de Genève

Pallanti Luca, ECP, Université Lumière Lyon2

Ponton Claude, LIDILEM, UGA

Richard Ange, LIG, Pacte

Ruscio Valeria, Sapienza, Università di Roma (ISO, DIAG) Seddah Djamé, Université Paris-Sorbonne (Paris 4), Inria

# Bibliographie

Al Sharou, K., Li, Z., & Specia, L. (2021). Towards a Better Understanding of Noise in Natural Language Processing. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 5362. https://aclanthology.org/2021.ranlp-1.7

Benveniste, É. (1966). *Problèmes de linguistique générale*. Gallimard.

Egbert, J., & Baker, P. (Eds.). (2019). *Using corpus methods to triangulate linguistic analysis*. Routledge.

Fuchs, C., & Habert, B. (2004). Le traitement automatique des langues : Des modèles aux ressources. *Le Français Moderne - Revue de linguistique Française, CILF (conseil international de la langue française), LXXII: 1*, en ligne.

Goutte, C., Carpuat, M., & Foster, G. (2012). The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers.*

Jentsch, P., & Porada, S. (2020). From Text to Data : Digitization, Text Analysis and Corpus Linguistics. In S. Schwandt (Éd.), *Digital Humanities Research* (1re éd., Vol. 1, p. 89128). transcript Verlag / Bielefeld University Press. https://doi.org/10.14361/9783839454190-004

Kraif, O., & Ponton, C. (2007). Du bruit, du silence et des ambiguïtés : Que faire du TAL pour l’apprentissage des langues ? *TALN 2007*, 143152. https://hal.archives-ouvertes.fr/hal-01073706 Léon, J. (2018). Tal et linguistique : Application, expérimentation, instrumentalisation. *ELA. Etudes de linguistique appliquee*, *2*(190), 195203.

Litosseliti, L. (Ed.). (2018). *Research methods in linguistics*. Bloomsbury Publishing.

Molinelli, P., & Putzu, I. (2015). *Modelli epistemologici, metodologie della ricerca e qualità del dato.* *Dalla linguistica storica alla sociolinguistica storica*. FrancoAngeli.

Péry-Woodley, M.-P., Afantenos, S. D., Ho-Dac, L.-M., & Asher, N. (2011). La ressource ANNODIS, un corpus enrichi d’annotations discursives. *TAL*, *52*(3), 71101.

Sarrica, M., Mingo, I., Mazzara, B., & Leone, G. (2016). The effects of lemmatization on textual analysis conducted with IRaMuTeQ: results in comparison. *JADT2016: 13ème Journées Internacionales d’Analyse Statistique de Données Textuelles.*

Scaglione, F. (2018). « Lavorare »; il dato linguistico: Prospettive e limiti. Alcune considerazioni dall’esperienza dell’Atlante Linguistico della Sicilia (ALS). In G. Sampino (Éd.), *Atti del convegno internazionale dei dottorandi* (p. 101122).

Zalmout, N., Erdmann, A., & Habash, N. (2018). Noise-robust morphological disambiguation for dialectal Arabic. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 953-964).

Zeroual, I. (2018). *Building Arabic Corpora: Concepts, Methodologies, Tools, and* Experiments (Doctoral dissertation, University of Maryland, USA).