Introduction

Elisa Gugliotta, Luca Pallanti, Olivier Kraif, Martina Barletta, Iris Fabry

Introduction

Elisa Gugliotta, Luca Pallanti, Olivier Kraif, Martina Barletta and Iris Fabry

AUTHOR'S NOTE

Tous les auteurs ont contribué de manière significative à l'élaboration de ce volume, participant à la révision des contributions et à la rédaction de l'introduction. Gugliotta et Pallanti ont coordonné et travaillé à la rédaction de l'introduction, la première s'occupant du bruit dans les corpus de langues peu dotées, le second des catégories de bruit dans les corpus d'apprenants. Kraif, Barletta et Fabry ont également rédigé certaines parties de l'introduction et contribué à la rédaction des résumés ainsi qu'à la révision de l'ensemble.

Contexte

1 Ce volume naît de la richesse des discussions et des échanges qui ont eu lieu lors de la journée d'étude intitulée « Bruit de fond ou valeur ajoutée ? Gérer le bruit lors des traitements informatiques des corpus linguistiques », qui s'est tenue à Grenoble le 28 avril 2023. Cet événement, organisé conjointement par l'Université Grenoble Alpes et l'Université Sapienza de Rome, a vu la collaboration des laboratoires LIDILEM (Univ. Grenoble Alpes) et ECP (Univ. Lumière Lyon 2) ainsi que de leurs jeunes chercheurs et doctorants. Cette dynamique fédérative a permis de réunir différentes générations de chercheurs et une pluralité de visions et compétences dans le domaine de la linguistique de corpus et du traitement automatique du langage (TAL). Les contenus de la présente publication sont donc le fruit d'un moment de rencontre scientifique qui a débouché sur une réflexion vaste et ouverte que nous détaillons dans les paragraphes qui suivent.

Les thèmes principaux et les perspectives ouvertes

2 L'importance croissante du TAL dans le domaine de la linguistique de corpus a soulevé de nouvelles questions concernant la gestion du bruit dans les données linguistiques. Le concept de « bruit » dans ce contexte est polysémique : il ne s'agit pas simplement d'un défaut à éliminer, mais il peut représenter une valeur ajoutée, une source d'informations supplémentaires qui, si elles sont gérées correctement, peuvent enrichir et affiner les analyses (Al Sharou *et al*. 2021).

3 Au cours de la journée d'étude, plusieurs aspects de la gestion du bruit ont été abordés, avec une attention particulière à la phase de collecte, de traitement et d'annotation des données linguistiques. Les discussions ont mis en évidence comment le bruit peut émerger sous différentes formes : du bruit introduit lors de la numérisation et de la normalisation des données, à celui intrinsèque aux données elles-mêmes, comme dans le cas des variations linguistiques ou des erreurs des apprenants (McEnery *et al.* 2019).

4 Un des points clés discutés a été le concept de bruit comme distorsion. En effet, dès l'introduction du concept de bruit dans la théorie de la communication de Shannon (1948), différentes dimensions ont été relevées : d'une part, le bruit comme information additionnelle – et indésirable – qui se superpose au message original, d'autre part le bruit comme distorsion, en lien avec le canal de transmission, qui engendre des déformations plus systématiques et potentiellement réversibles. Ainsi, le modèle de Shannon a inspiré des interprétations qui voient le bruit non seulement comme un défaut, mais aussi comme une partie du processus communicatif à gérer et, dans certains cas, à valoriser. Dans cette perspective, le bruit est perçu comme une information supplémentaire qui se superpose au message original, pouvant potentiellement le contaminer. Toutefois, comme il a été souligné lors de la clôture de la journée d'étude, il n'est pas toujours possible ou souhaitable d'éliminer complètement ce bruit. En effet, durant les processus de correction et de filtrage, le risque d'introduire d'autres erreurs ou bruits secondaires est élevé, surtout dans des phases délicates comme la reconnaissance optique de caractères (OCR). Dans ce volume, nous avons voulu relever le défi de considérer le bruit à la fois comme une source d'erreur et comme une source d'information qui peut améliorer les modèles TAL s'il est traité de manière adéquate (Bejan *et al.* 2023).

5 Le volume s'appuie sur une typologie du bruit qui vise à refléter l'articulation du phénomène dans ses différentes acceptions et nuances. Ainsi, nous avons identifié quatre principales catégories de bruit qui émergent lors du traitement des données linguistiques :

6 1. *Bruit comme distorsion communicative.* Ce type de bruit se manifeste lorsque le message original est altéré par des informations supplémentaires, qui peuvent être difficiles à éliminer sans compromettre l'intégrité des données. Ce phénomène est particulièrement visible dans le contexte de la recherche contemporaine, lié aux technologies de l'information et à la communication numérique (Fuller & Goffey 2012). À titre d'illustration, on peut citer les problèmes d'encodage de caractères, comme la présence d'entités html etc.

7 2. *Bruit comme défaillance de l'annotation automatique.* Dans les analyses linguistiques basées sur des modèles statistiques, le bruit se traduit souvent par des erreurs d'étiquetage qui influencent la précision des résultats. Une approche courante,

proposée par exemple par Manning *et al.* (2008) pour gérer ce type de bruit est le filtrage sélectif, qui vise à identifier et à réduire les portions de données les plus problématiques. Les auteurs discutent de diverses méthodes pour gérer et réduire le bruit, soulignant comment l'identification et la réduction des erreurs peuvent améliorer significativement les métriques de précision et de rappel dans les analyses linguistiques.

8 3. *Bruit comme déviation de la norme.* Dans des contextes comme les corpus d'apprenants, où les erreurs linguistiques font intrinsèquement partie de l'objet d'étude, le bruit devient l'objet même de la recherche, et constitue une donnée à part entière. Certaines erreurs peuvent être considérées comme centrales pour la recherche, tandis que d'autres peuvent revêtir un caractère aléatoire qui les rapprochent du bruit au sens de distorsion communicative (Granger, Hung & Petch-Tyson 2002). C'est le cas, en particulier, des erreurs de performance (mot oublié ou mal entendu dans une dictée, etc.) vis-à-vis des erreurs de compétence.

9 4. *Bruit comme variation intrinsèque*. Il s'agit d'un type de bruit particulièrement difficile à gérer, car il provient des fluctuations naturelles et des ambiguïtés présentes dans les langues elles-mêmes, qui ne peuvent pas toujours être facilement rattachées à une seule catégorie (Labov 1972). Le défi ici est de trouver un équilibre entre la nécessité de catégoriser des données continues et la préservation de leur variabilité intrinsèque (Alorifi *et al.* 2020). C'est notamment le cas des annotations effectuées sur une base interprétative, comme dans certains codages sémantiques, intrinsèquement variables d'un individu à l'autre (cf. article de Jonas Noblet, infra).

Vers une gestion critique du bruit

10 En 1986, Sperber et Wilson (1986), bien qu'ils ne traitassent pas du bruit au sens technique, introduisaient le concept de pertinence, étroitement lié à la gestion de l'information dans le langage. La pertinence est la mesure dans laquelle une information est significative ou utile pour un contexte ou un objectif donné. En d'autres termes, elle concerne la potentialité d'une donnée à contribuer de manière efficace à la compréhension ou à la résolution d'un problème (Saracevic 2022, Guyon & Elisseeff 2003).

11 Nous pensons que la notion de *pertinence* est naturellement associée au concept de bruit, car elle met l'accent sur la nécessité de caractériser le bruit de manière détaillée, en tant que phase préliminaire à toute tentative de normalisation et de catégorisation des données linguistiques. Dans cette perspective, l'un des principaux problèmes liés à la pertinence consiste à reconnaître et à documenter avec précision le bruit présent dans les données. Par exemple, en matière d'apprentissage automatique, la pertinence est souvent utilisée pour sélectionner les caractéristiques les plus significatives pour l'entraînement des modèles, contribuant à améliorer les performances et à réduire le surentraînement, ou *overfitting* (Lin *et al.* 2022). Une prise de conscience du degré de pertinence du bruit paraît donc cruciale pour une gestion critique des données linguistiques, en particulier dans une perspective écologique de celles-ci[1].

12 De surcroît, le bruit peut être considéré comme une partie d'un continuum informationnel qui, s'il est interprété dans le bon contexte, peut contribuer à la communication en fournissant des informations sur des phénomènes linguistiques non

immédiatement visibles (Biber 1998). En revanche, une mauvaise gestion du bruit peut aplatir les niveaux d'information contenus dans un texte, en particulier dans le traitement de corpus dialectaux ou de langues non standards, comme l'arabizi, le judéo-arabe et plus généralement ce que l'on appelle les dialectes arabes (Besdouri *et al.* 2024). En sous-estimant la valeur du bruit, il y a un risque de perdre des nuances de signification, à travers, par exemple, des processus de normalisation trop agressifs qui entraîneraient un aplatissement des données originales. Dans ces cas, une stratégie prudente à adopter pourrait consister à conserver plusieurs niveaux d'annotation, permettant de préserver la valeur informationnelle du bruit à chaque étape.

13 Dans cette perspective multidimensionnelle du bruit, ce volume rassemble une série de contributions qui explorent les différentes facettes du bruit dans les données linguistiques, proposant des approches originales, à la fois théoriques et pratiques, pour sa gestion. Ces approches se déclinent sous la forme d'analyses théoriques ou d'études empiriques, offrant un aperçu élargi des défis et des opportunités liés au bruit dans les corpus linguistiques.

14 La conclusion de la journée d'étude nous a permis de faire une synthèse qui a constitué un point de départ pour de nouvelles explorations. L'invitation à contribuer à ce volume a été étendue à tous les participants et à l'ensemble de la communauté scientifique, dans le but de continuer à enquêter et à discuter sur la manière dont le bruit peut non seulement être géré, mais aussi valorisé dans le cadre de la recherche linguistique.

Composition du numéro

15 Dans le premier article du volume, intitulé *Des bruits dans mon corpus : des données à réduire au silence, à atténuer ou à écouter attentivement*, Loïc Liégeois propose au lecteur une vision globale des étapes méthodologiques les plus courantes où peut intervenir une procédure de gestion du bruit. En effet, qu'il s'agisse de la mise en forme des données brutes, de l'enrichissement des données secondaires ou de l'analyse, les interventions du chercheur peuvent générer des interférences dans les traitements successifs. Quels sont donc les enjeux liés aux différentes méthodes de gestion du bruit ? En s'appuyant sur des exemples concrets qui s'inscrivent dans le champ de l'acquisition du langage, l'article parvient à synthétiser le caractère multiforme du bruit associé à la manipulation des données textuelles.

16 À la suite de ce tour d'horizon, l'article *Navigating Noise : A Stratified Model for Scholarly Digital Editions of Arabic Manuscripts in Hebrew Script*, de Valentina Bella Lanza, donne une illustration concrète des différentes étapes liées à la gestion du bruit. Il traite de l'importance des éditions savantes numériques (ESN) dans l'étude des manuscrits arabes écrits en hébreu, en mettant l'accent sur la complexité orthographique du judéo-arabe. Les ESN représentent une avancée importante dans les sciences humaines, car elles combinent des outils numériques et des méthodes académiques pour créer et analyser le contenu textuel. L'article souligne que ces éditions doivent produire de nouvelles connaissances et ne pas se contenter de reproduire des sources primaires. L'un des principaux défis est la gestion du bruit dans le traitement des textes, bruit qui peut inclure des éléments à la fois nuisibles et utiles à la compréhension. Pour répondre à cette dualité, l'autrice propose un modèle stratifié, en quatre niveaux, pour la création de l'ESN, en visant à préserver l'intégrité orthographique et à réduire le bruit

nuisible lors de la manipulation des textes originaux. L'article présente également une étude de cas d'un manuscrit du 15e siècle, illustrant l'application du modèle en couches.

17 Dans son article *Numériser le patrimoine linguistique québécois : le traitement informatique des fiches dialectologiques de Gaston Dulong*, Wim Remysen décrit les solutions élaborées par son équipe pour gérer la numérisation et la revalorisation d'un ensemble massif de données dialectologiques conservées dans un format papier et dont la documentation est sporadique. L'auteur nous guide à travers ce processus en décrivant tout d'abord les données hétéroclites et volumineuses sur lesquelles il travaille et le bruit que celles-ci génèrent en transcription, pour nous amener ensuite aux décisions prises dans l'élaboration d'un outil de traitement adapté. En somme, c'est ce même ordre de préparation que W. Remysen et son équipe proposent, illustrant ainsi un processus de transcription intéressant, réalisé autour du bruit occasionné par l'informatisation plutôt que contre. Cette anticipation du besoin a permis à W. Remysen de considérablement faciliter et accélérer un travail autrement chronophage et fastidieux.

18 S'intéressant aux conséquences du bruit lié à la phase initiale d'acquisition des textes, Ljudmila Petkovic, Caroline Koudoro-Parfait, Marie-Sophie Demarest et Gaël Lejeune évaluent l'impact du bruit de reconnaissance optique des caractères (OCR) sur la tâche en aval de reconnaissance automatique des entités nommées (EN). Leur article, intitulé *Quelle solution pour améliorer les performances de la reconnaissance d'entités nommées sur des données bruitées, corriger l'entrée ou filtrer la sortie ?*, présente l'originalité de mener une étude empirique fine sur l'interaction entre deux types de bruit, bruit d'OCR et bruit de la détection des EN avec Spacy, sans vérité de terrain, en s'appuyant sur une référence elle-même partiellement bruitée, dite *Silver Standard* (par opposition à un *Gold Standard* manuellement vérifié). Ils montrent notamment qu'une procédure de remédiation du bruit issu de la phase d'OCR, par application d'un correcteur orthographique, ne présente un intérêt que dans le cas d'un OCR très bruité. La meilleure stratégie pour filtrer les EN erronées liées à l'OCR consiste plutôt à s'appuyer sur des critères tels que longueur des chaînes et fréquence documentaire. Les auteurs montrent ainsi qu'il peut être bénéfique de combiner plusieurs sorties d'OCR *différemment* bruitée pour améliorer le filtrage des sorties.

19 Du problème de l'acquisition de l'écrit par OCR on passe ensuite à celui de la transcription de l'oral vers l'écrit, avec la contribution de Thomas Bertin et Gwenolé Quellec, *Transcription automatique des interactions verbales - Limites observées et perspectives envisagées à partir d'un corpus de consultations médicales*, qui analyse les limites des outils de reconnaissance automatique de la parole (RAP) appliqués à la transcription automatique d'un corpus dans une perspective d'étude des interactions. Ces outils sont appliqués à des fins d'analyse d'un « échantillon de transcriptions de consultations pré-opératoires en vue d'une intervention de neurochirurgie en conditions éveillées ». Bien que permettant un gain de temps dans la transcription vers un écrit normé, ces outils s'avèrent peu adaptés pour capturer la complexité des échanges verbaux dans les interactions, en gommant des phénomènes qu'on pourrait qualifier de « bruit utile », comme les chevauchements de parole, les disfluences et autres marqueurs de l'oralité. L'étude propose des solutions pour améliorer ces outils, en intégrant mieux les particularités de l'oral, comme la délimitation correcte des tours de parole ou encore la restitution des marqueurs de l'oral dans les résultats de transcription.

20 Les articles qui suivent abordent les étapes ultérieures d'annotation. Ainsi, dans la contribution de Delphine Bernhard et Joanna Dolińska intitulée *Managing Noise in Part-*

of-Speech Tagging for Extremely Low-Resource Languages : Comparing Strategies for Corpus Collection and Annotation in Dagur and Alsatian, l'étiquetage morphosyntaxique (*POS-tagging*) du dagur et de l'alsacien sont comparés pour évaluer des possibles stratégies de réduction du bruit. Le peu de ressources disponibles pour ces deux langues, ainsi que l'absence d'un système orthographique unifié, rendent nécessaire une normalisation en amont de tout traitement automatique. Deux corpus manuellement annotés en étiquettes de parties du discours (*POS-tag*), un pour chaque langue, sont comparés avec les annotations automatiques des différents modèles. Les expériences présentées, qui portent sur l'uniformisation orthographique dans le cas du dagur et sur le maintien de la richesse orthographique présente dans le corpus d'alsacien, exploitent le principe du transfert direct (*zero-shot transfer*), et montrent que les méthodes basées sur des langues syntaxiquement similaires peuvent être efficaces, en soulignant le rôle crucial de la proximité linguistique dans l'amélioration de l'étiquetage morphosyntaxique des langues peu dotées.

21 L'article de Jonas Noblet, *Le bruit dans la mesure de la composante cognitive de l'émotion pour l'évaluation de l'acceptabilité des innovations*, aborde les dimensions sémantiques de l'annotation, à travers une grille distinguant différentes dimensions émotionnelles. Il explore l'impact du bruit dans l'annotation de verbatims retranscrivant des jugements sur l'acceptabilité de certaines innovations. Il critique l'utilisation du désaccord entre annotateurs comme mesure de bruit, arguant qu'elle peut conduire à une surestimation de ce dernier. Noblet propose un modèle probabiliste pour mieux gérer la variabilité des annotations, bien que cela limite l'évaluation précise du bruit. L'étude s'appuie sur la méthode EMINOSA, qui utilise un cadre théorique multi-composant pour analyser les émotions liées aux innovations. Le corpus Yoomaneo, composé de 2 446 messages d'évaluateurs, est utilisé pour tester cette méthode. Plusieurs expériences pilotes révèlent des incohérences dans les annotations, avec des scores d'accord inter-annotateurs souvent faibles. Malgré des efforts pour améliorer la cohérence via des guides d'annotation, les résultats demeurent hétérogènes, remettant en question la validité de la mesure de la composante cognitive des émotions. L'article conclut sur la nécessité de repenser la modélisation du bruit pour optimiser l'analyse des annotations.

22 Dans leur article intitulé *La question de la normalisation des écrits scolaires pour leur traitement automatique. Le cas de l'omission de mots*, Martina Barletta et Claude Ponton traitent de la normalisation des écrits scolaires en vue de leur traitement automatique, en se concentrant sur l'omission de mots et le bruit causé par ces « absences » dans les textes. Les écrits scolaires, souvent éloignés de la norme linguistique, présentent des défis pour l'analyse automatique. Le projet Scoledit propose une solution par la normalisation des textes, permettant de comparer les productions des élèves à différents niveaux linguistiques. Cependant, ces omissions de mots peuvent introduire des erreurs dans les analyses. L'article explore trois méthodes pour aborder ces omissions : l'utilisation d'un token masque, d'un token spécifique à la catégorie grammaticale et l'utilisation du modèle Transformer FlauBERT pour prédire les mots omis. Le corpus Scolinter, un corpus trilingue d'écrits scolaires, est présenté, et les omissions sont analysées par catégorie grammaticale. Enfin, les trois méthodologies sont évaluées en fonction de leur efficacité, soulignant l'importance de traiter les omissions pour améliorer l'annotation morphosyntaxique ainsi que l'analyse de la cohérence textuelle.

23 Dans la perspective de l'exploitation de corpus automatiquement annotés en étiquettes morphosyntaxiques, Christian Surcouf, dans son article *À pas de loup dans la bergerie… La problématique du silence et du bruit dans l'étiquetage automatique du subjonctif présent en français parlé*, pointe un phénomène rarement évoqué dans l'utilisation des analyseurs morphosyntaxiques : l'hétérogénéité du bruit en fonction des étiquettes. Prenant l'exemple du subjonctif présent, il effectue une analyse fine des ambiguïtés graphiques liées à la distinction entre le subjonctif et d'autres modes tels que l'indicatif ou l'impératif, puis il conduit une analyse empirique sur un corpus de transcriptions de l'oral. À partir d'une évaluation manuelle des sorties de trois étiqueteurs, il montre que les performances réelles liées à cette étiquette sont bien en deçà des performances habituellement affichées par les étiqueteurs, avec des taux de bruit et de silence pouvant dépasser les 50 % pour certains outils. Ce constat plaide pour une vision plus fine du bruit prenant en compte les disparités dans ses distributions : dans une étude qui s'intéresserait à l'usage du subjonctif à l'oral, ce bruit pourrait aboutir à un silence important dans les résultats de requêtes (silence d'autant plus insidieux qu'il passe inaperçu !), et biaiser les observations de manière rédhibitoire.

24 Enfin, le volume se clôt sur des pistes permettant d'intégrer le bruit dans les analyses, sans chercher à l'éliminer à tout prix. En s'appuyant sur un cas pratique d'analyse textométrique de la Base de français médiéval (BFM), dans l'article intitulé *Apprivoiser le « bruit » en linguistique de corpus : expérience d'une analyse factorielle et propositions*, Bénédicte Pincemin livre un protocole de gestion du bruit dans un corpus volumineux. L'enjeu consiste à contourner les erreurs d'étiquetage morphosyntaxique du corpus à travers des phases progressives de vérification et de filtrage des étiquettes. L'article montre comment le recours à l'analyse factorielle se révèle un choix pertinent, qui permet de filtrer et de détecter les éléments linguistiques « déviants » les plus fréquents, sans pour autant donner lieu à une manipulation des données du corpus (avec les risques de distorsion qu'elle implique).

Conclusion

25 À travers l'ensemble des articles du volume, on constate à quel point le bruit est une notion multi-dimensionnelle complexe en linguistique de corpus, recouvrant des concepts variés tantôt liés au problème de l'acquisition des données (dans le cliché d'un manuscrit ou l'enregistrement d'une interaction orale), à la question de la variation et des écarts à la norme (tout spécialement sur le plan orthographique), au problème des aléas liés à la performance des locuteurs impliqués (disfluences à l'oral, omissions à l'écrit), à la performance des outils automatiques (étiqueteurs, analyseurs), ou encore à la variabilité interprétative inhérente au matériau linguistique (quand plusieurs analyses syntaxiques ou plusieurs interprétations sémantiques sont possibles et également correctes). Loin d'être seulement une donnée quantitative, quand une référence existe, le bruit reste un aiguillon méthodologique et critique, qui doit sans cesse pousser le chercheur ou la chercheuse à s'interroger sur l'intégrité de ses données, la fiabilité de ses annotations, la possibilité des biais et le contrôle de ces derniers.

26 De toutes ces contributions, nous pourrons tirer deux enseignements : d'une part, que le bruit, s'il ne peut être éradiqué, peut être correctement géré et contrôlé ; et, d'autre part, que le bruit est intrinsèquement lié au point de vue que l'on a sur ses données. Ce

qui est du bruit pour une recherche peut représenter, pour une autre, le cœur même de la donnée.

27 En conclusion, nous espérons que ce volume contribuera à une compréhension plus profonde du bruit dans les corpus linguistiques et qu'il offrira de nouvelles perspectives pour aborder l'un des défis les plus complexes dans le domaine du TAL et de la linguistique de corpus. Nous sommes certains que les réflexions présentées ici susciteront d'autres discussions et recherches, enrichissant le débat académique sur un thème à la fois crucial et fascinant.

Remerciements

28 Nous souhaitons remercier l'Université Grenoble Alpes et l'Université Sapienza de Rome pour avoir soutenu l'organisation de la journée d'étude « Bruit de fond ou valeur ajoutée ? Gérer le bruit lors des traitements informatiques des corpus linguistiques ». Cet évènement, qui a obtenu le label de l'Université Franco-italienne, a pu compter sur le soutien financier des laboratoires LIDILEM (Univ. Grenoble Alpes) et ECP (Univ. Lumière Lyon 2) ainsi que de l'institut MIAI (Univ. Grenoble Alpes) et de UGA-INP. Nous remercions ces institutions de la confiance accordée. Nous tenons enfin à remercier la revue *Corpus* d'avoir relevé le défi d'une publication autour de la question du bruit.

BIBLIOGRAPHY

Al Sharou K., Li Z. & Specia L. (2021, septembre). « Towards a better understanding of noise in natural language processing », *in Proceedings of the International Conference on Recent Advances in Natural Language Processing* (*RANLP 2021*), 53-62.

Bejan I., Sokolov A. & Filippova K. (2023, décembre). « Make Every Example Count : On the Stability and Utility of Self-Influence for Learning from Noisy NLP Datasets », *in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10107-10121.

Besdouri F. Z., Zribi I. & Belguith L. H. (2024). « Arabic Automatic Speech Recognition : Challenges and Progress », *Speech Communication* 163 : 103110.

Biber D. (1998). « Corpus linguistics : Investigating language structure and use », *Cambridge University Press google schola* 2 : 230-239.

Granger S., Hung J. & Petch-Tyson S. (éd.). (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. John Benjamins.

Guyon I. & Elisseeff, A. (2003). « An introduction to variable and feature selection », *Journal of Machine Learning Research* 3 : 1157-1182.

Fuller M. & Goffey A. (2012). *Evil media*. MIT Press.

Labov W. (1972). *Sociolinguistic patterns*. University of Pennsylvania.

Lin J., Nogueira R. & Yates A. (2022). *Pretrained transformers for text ranking : Bert and beyond*. Springer Nature.

Manning C. D., Raghavan P. & Schütze H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

McEnery T., Brezina V., Gablasova D. & Banerjee J. (2019). « Corpus linguistics, learner corpora, and SLA : Employing technology to analyze language use », *Annual Review of Applied Linguistics* 39 : 74-92.

Saracevic T. (2022). *The Notion of Relevance in Information Science : Everybody knows what relevance is. But, what is it really ?* Springer Nature.

Shannon C. E. (1948). « A Mathematical Theory of Communication », *The Bell system technical journal* 27(3) : 379-423.

Sperber D. & Wilson D. (1986). *Relevance : Communication and cognition*. Blackwell.

Wilkinson M. D., Dumontier M. *et al.* (2016). « The FAIR Guiding Principles for scientific data management and stewardship », *Scientific data* 3(1) : 1-9.

NOTES

1. Ici, la référence concerne les principes *FAIR* de l'*open science* (Wilkinson *et al.* 2016), sous-entendant qu'il serait possible d'éviter d'éliminer de manière indiscriminée des informations qui pourraient s'avérer précieuses pour des études futures.

AUTHORS

ELISA GUGLIOTTA

Laboratoire LIDILEM (Univ. Grenoble Alpes) & Università degli Studi di Sassari, Dipartimento di Storia, Scienze dell'Uomo et della Formazione

LUCA PALLANTI

Laboratoire ECP (Univ. Lumière Lyon 2)

OLIVIER KRAIF

Laboratoire LIDILEM (Univ. Grenoble Alpes)

MARTINA BARLETTA

Laboratoire LIDILEM (Univ. Grenoble Alpes) & "Riccardo Massa" Department of Human Sciences for Education (Université Milan-Bicocca)

IRIS FABRY

Laboratoire LIDILEM (Univ. Grenoble Alpes)