Le multilinguisme dans les références bibliographiques : une étude de corpus

Marc Bertin*, Iana Atanassova**,***

*ELICO, Université Claude-Bernard Lyon 1, France
marc.bertin@univ-lyon1.fr
**Université de Franche-Comté, CRIT, 25000 Besançon, France
iana.atanassova@univ-fcomte.fr
***Institut Universitaire de France (IUF)

Cet article présente les résultats de deux études publiées dans l'atelier BIR@ECIR2024 (Bertin et Atanassova, 2023) et la conférences internationale ISSI 2024 (Atanassova et Bertin, 2023).

Le statut de lingua franca de l'anglais dans la recherche est déjà bien établi. Dans quelle mesure les citations d'articles rédigés en anglais sont-elles prédominantes par rapport aux articles rédigés dans d'autres langues ? Dans cet article, nous abordons le problème du multilinguisme dans les publications par l'analyse de deux corpus différents : d'un côté le Semantic Scholar Open Research Corpus [1] , et d'un autre côté la base de données multilingue ISTEX [2]. Notre objectif est d'examiner les références en langues étrangères (différentes de la langue de publication) dans ces corpus, et d'observer leur nature et leur distribution.

Nous avons identifié les langues des articles et de leurs références sur la base de leurs titres. Dans notre étude nous utilisons l'ensemble de données S2ORC version 1, qui contient environ 81 millions d'articles en libre accès publiés jusqu'en 2020. ISTEX, quant à elle, fournit plus de 23 millions de documents en anglais, mais aussi des ressources dans plus de 50 langues. Nous avons traité l'ensemble des articles de toutes les langues présentes dans ISTEX, avec plus de 100 articles par langue. Pour cela, après avoir discuté de la multiplicité des écritures et de la translittération, notre méthodologie repose sur l'extraction des références, puis le titre de l'article de chaque référence, afin de détecter sa langue à l'aide de les bibliothèques python `spacy-langdetect2` et Compact Language Detector v3 (`gcld3`) de Google. Afin d'étudier la manière dont les différents groupes linguistiques sont représentés dans l'ensemble de données, nous avons classé les langues en groupes linguistiques. Nous avons évalué la détection des langues, et exclu les langues pour lesqueslles la précision observée était inférieure à 50%. Pour la majorité des langues conservées, la précision est supérieure à 75%.

Nous avons étudiés le nombre de références par langue pour les 9 langues examinées dans ISTEX. Nous observons que le nombre le plus important est celui des références de l'anglais vers l'anglais, suivi par les références du français vers l'anglais. Les figures 1 et 2 présentent les diagrammes de Sankey pour S2ORC et ISTEX. Nous observons que l'anglais est de loin le

1. `https://github.com/allenai/s2orc`, S2ORC version 1 contient environ 81 millions d'articles en libre accès publiés jusqu'en 2020, disponibles au format json.

2. `https://www.istex.fr/`, ISTEX : collections multilingues et multidisciplinaires de la littérature scientifique mondiale, en format XML TEI avec références en format Grobid et Bibtex

segment le plus important du côté droit, ce qui signifie que la grande majorité des articles ont tendance à citer des références en anglais.

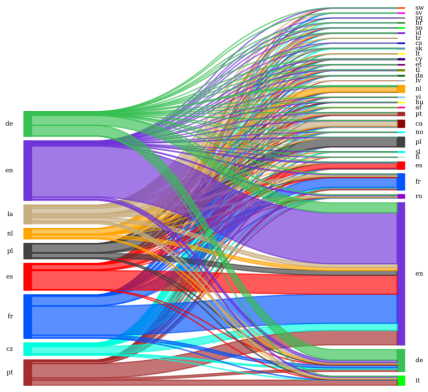

FIG. 1 – *Langues citantes et des langues citées dans ISTEX*

FIG. 2 – *Langues citantes et des langues citées dans les articles non anglophones de S2ORC*

Les résultats présentés dans cette étude peuvent être faussés par différents facteurs. Tout d'abord, la qualité de la détection des langues peut varier d'une langue à l'autre, en particulier pour les langues peu dotées qui peuvent avoir un faible taux de reconnaissance.

Les exigences éditoriales de certaines revues peuvent favoriser les références en langue anglaise ou exiger que les titres des références en langue étrangère soient traduits dans la langue de la publication. En outre, des différences disciplinaires sont à prévoir et nécessiteraient une étude plus détaillée.

Dans S2ORC, bien que la grande majorité d'articles soit en anglais, nous avons identifié des articles dans 44 langues différentes et des références dans 54 langues. Nous avons observé leurs groupes linguistiques et leur répartition dans le temps entre 1950 et 2020 et souligne la prédominance de l'anglais dans la science, où la grande majorité des citations est vers des sources anglophones. Cependant, nous avons également identifié des références non anglophones en lien avec un lieu ou une activité. Des connaissances produites dans une langue avec une expertise locale est mobilisée. Ces travaux trouverons échos avec les récentes évolutions des politiques éditoriales qui prennent en considération la dimension multilingue de la science.

Remerciements Ce travail est soutenu par l'ANR-20-CE38-0003-01

Références

Atanassova, I. et M. Bertin (2023). Multilinguism in References : a Study of the ISTEX Dataset. In *Proceeding of the 19th ISSI Conference 2023*, Indiana USA.

Bertin, M. et I. Atanassova (2023). Citing Foreign Language Sources : an Analysis of the S2ORC Dataset. In *Proceedings of BIR workshop at ECIR 2023.*, Dublin, Ireland.