Le corpus “ Machine Translation ” : une exploration diachronique des (méta)données Istex

Mathilde Huguin, Sabine Barreaux

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Le corpus *Machine Translation* Une exploration diachronique des (méta)données Istex

Mathilde Huguin[1] Sabine Barreaux[1]
(1) Inist - CNRS (UAR 76), 54 519 Vandœuvre-lès-Nancy, France
mathilde.huguin@inist.fr, sabine.barreaux@inist.fr

RÉSUMÉ

Le corpus *Machine Translation* se compose de publications scientifiques issues du réservoir Istex. Conçu comme un cas d'usage, il permet d'explorer l'histoire de la traduction automatique au travers des métadonnées et des textes intégraux disponibles pour chacun de ses documents. D'une part, les métadonnées permettent d'apporter un premier regard sur le paysage de la traduction automatique grâce à des tableaux de bord bibliométriques. D'autre part, l'utilisation d'outils de fouille de textes sur le texte intégral rend saillantes des informations inaccessibles sans une lecture approfondie des articles. L'exploration du corpus est réalisée grâce à Lodex, logiciel open source dédié à la valorisation de données structurées.

ABSTRACT

The *Machine Translation* Corpus. A diachronic exploration of Istex (meta)data

The *Machine Translation* corpus consists of scientific publications from the Istex repository. Conceived as a use case, it allows one to explore the history of machine translation through the metadata and full texts available for each of its documents. On the one hand, the metadata provide a first look at the machine translation landscape through bibliometric dashboards. On the other hand, the use of text mining tools on the full text brings out information that is inaccessible without a thorough reading of the articles. The exploration of the corpus is carried out using Lodex, an open source software dedicated to the valorisation of structured data.

MOTS-CLÉS : Traduction Automatique, Corpus, Istex, Bibliométrie, Fouille de textes.

KEYWORDS: Machine Translation, Corpus, Istex, Bibliometrics, Text Mining.

1 Introduction

L'atelier que nous proposons poursuit deux objectifs : **(i)** présenter la ressource Istex [1] (*Initiative d'excellence en Information Scientifique et Technique*), qui permet de construire des corpus de publications scientifiques, et **(ii)** montrer que les formats et outils accessibles depuis Istex offrent des solutions profitables pour la fouille de textes. Notre démonstration trouve son origine dans la dynamique nationale (ex. ANR-22-MaTOS-0033 [2] ; Fiorini *et al.*, 2020) et européenne actuelle (ex.

1. Istex (www.istex.fr) est né en 2011 dans le cadre des Programmes d'Investissement d'Avenir (ANR-10-IDEX-0004-02).
2. https://anr-matos.github.io

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 54–59.

OPERAS [3] ; Helsinki-Initiative, 2019) autour des nouvelles méthodes de traduction automatique (désormais TA). Istex renfermant plus de sept siècles d'archives scientifiques, nous avons choisi d'explorer l'aspect diachronique de la TA à travers un corpus. Nous montrons, d'une part, comment l'analyse des métadonnées des publications donne accès à un panorama des travaux réalisés dans ce domaine et, d'autre part, en quoi l'analyse du texte intégral contribue à retracer l'histoire des méthodes et des approches mises en œuvre en TA.

Notre article s'organise comme suit. En 2, nous présentons le contenu d'Istex et les outils que nous manipulons. En 3, nous expliquons la méthodologie appliquée pour constituer le corpus et présentons quelques-uns des résultats obtenus.

2 L'écosystème Istex

Le réservoir Istex contient plus de 27 millions de publications scientifiques dans toutes les disciplines et dans plus de 50 langues. Il est alimenté en continu au travers des acquisitions pérennes des licences nationales [4] et de celles du GIS CollEx-Persée [5]. En 2023, il regroupe plus de 9 000 revues et 430 000 ebooks de 41 éditeurs différents, publiés entre le XIV^{e} siècle et aujourd'hui. Ces documents sont accessibles depuis l'interface Istex-DL [6] (*Istex Download*), connectée à une API [7], qui permet de télécharger les publications et de choisir les formats appropriés. En effet, si le réservoir constitue d'abord une ressource documentaire, Istex n'est pas exclusivement un outil de consultation. Il offre la possibilité de constituer des corpus à des fins de fouille de textes et d'analyse de contenu (ex. Bordignon & Maisonobe, 2022). Dans ce cas, il s'agit non seulement de rechercher des documents, mais aussi de vérifier leurs propriétés, de les télécharger massivement, pour finalement effectuer des traitements seuls ou les intégrer dans des outils. Tous les documents Istex sont disponibles à la fois sous forme de métadonnées et dans leur version en texte intégral, et ce, dans différents formats soit d'origine, soit convertis dans des standards (XML MODS et XML TEI) pour faciliter leur exploitation dans les outils de fouille ou de textométrie (ex. TXM Heiden *et al.*, 2010). Les documents Istex ont également été enrichis grâce à des outils, développés ou adaptés pour Istex (Cuxac & Thouvenin, 2017). Parmi ces enrichissements, nous exploiterons (§3.2) plus particulièrement la structuration des références bibliographiques obtenue avec l'outil GROBID [8] (*GeneRation Of Bibliographic Data*).

Lors du téléchargement, Istex-DL propose une passerelle vers l'outil Lodex [9] (*Linked Open Data EXperiment*), que nous utiliserons pour naviguer dans le contenu du corpus. Lodex est un logiciel open source [10] créé pour les besoins du projet Istex afin de valoriser ses données structurées (Gregorio *et al.*, 2019). Il permet de concevoir des sites web offrant des interfaces pour explorer visuellement un jeu de données (CSV, JSON, etc.) au travers de tableaux de bord dynamiques présentant des indicateurs bibliométriques. Lodex offre également la possibilité d'utiliser des services web [11] afin d'enrichir les documents à l'aide de programmes d'analyse, de curation, d'annotation et d'indexation.

3. https://operas-eu.org/projects/translations-and-open-science
4. https://www.licencesnationales.fr
5. https://www.collexpersee.eu
6. https://dl.istex.fr
7. https://api.istex.fr/document/?q=
8. https://github.com/kermitt2/grobid
9. https://lodex.inist.fr
10. https://github.com/Inist-CNRS/lodex
11. https://objectif-tdm.inist.fr/category/services

3 Le corpus *Machine Translation*

3.1 Méthodologie de constitution de corpus

L'élaboration du corpus *Machine Translation* suit une procédure itérative (de Salabert & Barreaux, 2020). La requête interrogeant Istex utilise la syntaxe Lucene [12] (pour les non-initiés, il est possible d'utiliser l'outil de recherche assistée). Notre requête initiale filtre les documents comportant des mots-clés relatifs à la TA dans tous les champs. Les documents sont téléchargés via Istex-DL, puis importés dans Lodex. Comme nous le montrerons lors de l'atelier, la visualisation et l'analyse de ces données aident à détecter le bruit et/ou le silence et amènent à une révision de la requête initiale. La requête corrigée cible les champs dans lesquels nous cherchons les mots-clés (ex. *title :"machine translation"*), restreint les langues cibles (ex. *language.raw :"eng" "fre"*) et exclut certains éditeurs provoquant du bruit (ex. *NOT corpusName :"nature"*). Dans sa version finale, le corpus *Machine Translation* se compose de 7 160 documents en anglais et en français, soit plus de 54 millions de mots. Il est accessible suivant ce lien : `http://traduction-machinetranslation.corpus.istex.fr`.

3.2 Indicateurs bibliométriques

Selon les choix de calculs, de curations ou de visualisations opérés, les métadonnées et les enrichissements apportés par Istex fournissent plus d'une quinzaine d'indicateurs bibliométriques. La Figure 1 présente, par exemple, les dix revues les plus fréquentes dans notre corpus. On y retrouve ainsi *Machine Translation*, la principale revue de TA.

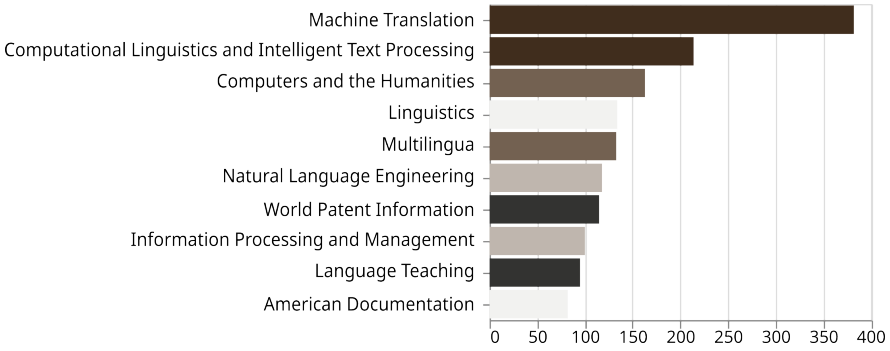

FIGURE 1 – Les dix revues majoritaires dans le corpus *Machine Translation*

Nous focaliserons notre présentation sur deux indicateurs bibliométriques obtenus grâce aux enrichissements et à l'utilisation de services web sur les métadonnées fournies par Istex.

(a) Les champs d'affiliations des auteurs sont restructurés et normalisés pour obtenir une cartographie des pays publiants, cf. Figure 2. Pour ce faire, nous utilisons deux services web développés par l'Inist et appelés depuis Lodex. Le service web de découpage d'adresses [13] retourne une adresse au format texte en tableau de champs. Le service web exploitant le thésaurus Loterre des noms de Pays [14] normalise les graphies des pays (regroupant par exemple *USA* et *États-Unis*). Comme attendu, les

12. `www.elastic.co`
13. `https://gitbucket.inist.fr/tdm/web-services`
14. `https://skosmos.loterre.fr/9SD/fr`

États-Unis sont le premier pays publiant sur la TA (en noir dans la figure), suivis de la Chine et du Japon qui sont les deux premiers co-publiants des États-Unis.

(b) La structuration des références bibliographiques permet notamment de détecter les auteurs et revues les plus cités, cf. Figure 2. Ces indications fournissent indirectement des éléments historiques sur la TA. L'auteur le plus cité est *Philipp Koehn*, considéré comme l'un des inventeurs de la méthode *phrase based* (Koehn *et al.*, 2003) et développeur de l'outil Moses (Koehn *et al.*, 2007). La détection renvoie également *Warren Weaver* et *Adrew Booth*, instigateurs des premières idées pour traduire les langues naturelles (Hutchins, 2007), ou encore *Peter Toma* développeur de Systran.

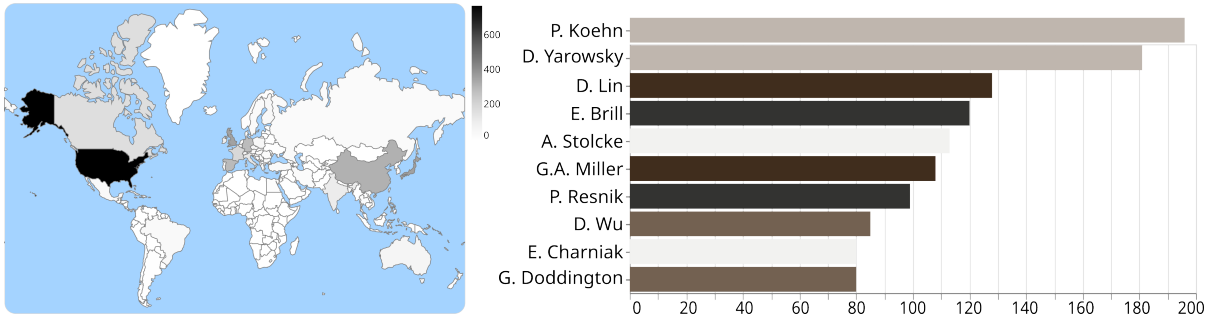

FIGURE 2 – Focus : **(a)** pays publiants et **(b)** auteurs citées

3.3 Diachronie de la TA

Nous nous servons de l'annotation du texte intégral pour retracer l'histoire de la TA et l'évolution des méthodes et des outils. Pour aboutir à ce résultat, nous avons initié une expérience en construisant une ressource terminologique bilingue d'environ 60 termes avec leurs variantes (sigles, alias). Ces termes désignent des modèles, outils, ou techniques de TA (ex. *lexical-functional grammar*, *Moses*). Ils sont regroupés selon leur appartenance à trois approches *Rule Based Machine Translation* (RBMT), *Statistical Machine Translation* (SMT) et *Neural Machine Translation* (NMT) (Hutchins, 2000, 2007).

La ressource terminologique est projetée à la fois sur les textes intégraux mais aussi sur les métadonnées (XML TEI) grâce à une feuille de style XSLT. Cette double annotation nous permet de vérifier l'apport de l'utilisation du texte intégral, cf. Tableau 1. Par rapport à une annotation des seules métadonnées, l'annotation du texte intégral permet d'obtenir 5 fois plus de documents dans lesquels des termes de la ressource sont détectés. Près de 4 200 documents sont ainsi catégorisés selon l'approche de TA utilisée [15].

	Métadonnées	Textes intégraux
Avec termes détectés	878	4 170
Sans termes détectés	6 282	2 990

TABLE 1 – Nombre de documents annotés selon l'input de l'annotation

Les occurrences des termes détectés dans les textes intégraux sont ensuite utilisées pour construire un graphique de flux montrant l'évolution temporelle des approches au sein des documents, cf. Figure 3.

15. Certains documents, hybrides, sont associés à plusieurs approches.

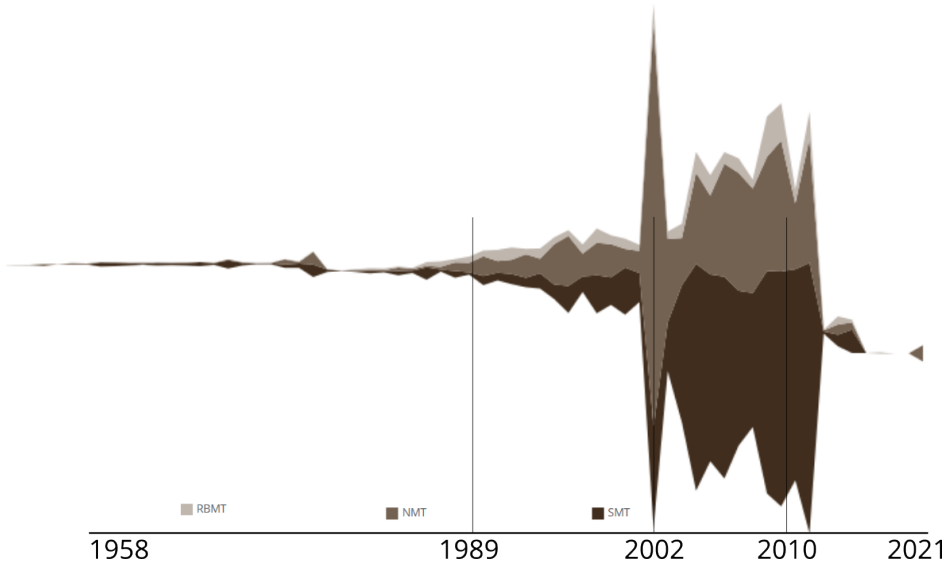

FIGURE 3 – Graphique de flux des approches en TA

La présence de certains termes dans plusieurs approches (ex. *Systran*) explique l'apparition de la NMT dès les années 80. Le graphique témoigne d'un essor de la TA dans les années 80 qui coïncide également avec l'invention de l'*example-based machine translation* (Song, 2022). La recherche en TA s'intensifie réellement depuis les années 90 (ce qui correspond à la création du modèle IBM de SMT). Deux valeurs sont particulièrement remarquables : le pic de 2002 coïncide peu ou prou avec la date de lancement du premier système de traduction sur internet et celui de 2010 à l'essor de la NMT.

4 Conclusion

À travers un cas d'usage, notre atelier montre qu'Istex est une ressource puissante pour créer des corpus de publications scientifiques et pour exploiter toute la richesse des métadonnées et du texte intégral des publications. L'écosystème Istex, dans sa globalité, offre une boîte à outils pour transformer, nettoyer, enrichir et visualiser ses données en vue de les analyser plus finement.

Pour aller plus loin, le corpus *Machine Translation* pourra (i) être enrichi avec des documents provenant d'autres sources, afin d'étoffer sa couverture chronologique, et (ii) être utilisé pour repérer automatiquement des définitions afin de compléter la ressource terminologique sur la TA.

Remerciements

Nous remercions chaleureusement la mastérante en traduction Manon Delorme pour son soutien dans la constitution de la ressource terminologique de TA et le responsable de ressources terminologiques Majid Khayari pour son aide technique (plus que précieuse) dans l'annotation du corpus *Machine Translation*.

Références

BORDIGNON F. & MAISONOBE M. (2022). Researchers and their data : A study based on the use of the word data in scholarly articles. *Quantitative Science Studies*, **3**(4), 1156–1178. DOI : 10.1162/qss_a_00220.

CUXAC P. & THOUVENIN N. (2017). Archives numériques et fouille de textes : le projet ISTEX. In P. CUXAC, V. LEMAIRE & J.-C. LAMIREL, Éds., *Atelier TextMine - EGC 17*, p. 43–51, Grenoble, France.

DE SALABERT C. & BARREAUX S. (2020). Vers un corpus optimal pour la fouille de textes : stratégie de constitution de corpus spécialisés à partir d'ISTEX. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*.

FIORINI S., BARBIN F., GARNIER-RIZET M., MORIN K. H., HUMPHREYS F., JOSSELIN-LERAY A., KÜBLER N., LOOCK R., MARTIKAINEN H., NOMINÉ J.-F., PLAG C., ROSSI C. & YVON F. (2020). *Rapport du groupe de travail "Traductions et science ouverte"*. report, Comité pour la science ouverte. Pages : 44 p., DOI : 10.52949/20.

GREGORIO S., COLLIGNON A., PARMENTIER F. & THOUVENIN N. (2019). LODEX : des données structurées au web sémantique. In *Atelier Web des Données de la 19ème Conférence sur l'Extraction et la Gestion des Connaissances (EGC 2019)*, Metz, France.

HEIDEN S., MAGUÉ J.-P. & PINCEMIN B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In *Statistical Analysis of Textual Data - Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, volume 2, p. 1021–1032, Rome, Italie : Edizioni Universitarie di Lettere Economia Diritto. Issue : 3.

HELSINKI-INITIATIVE (2019). *Helsinki Initiative on Multilingualism in Scholarly Communication*. Rapport interne, Federation of Finnish Learned Societies ; The Committee for Public Information ; Publishing, The Finnish Association for Scholarly ; Universities Norway ; European Network for Research Evaluation in the Social Sciences and the Humanities, Helsinki. Publisher : figshare.

HUTCHINS J. (2007). Machine translation : A concise history. *Computer aided translation : Theory and practice*, **13**(29-70), 11. Publisher : Chinese University of Hong Kong.

HUTCHINS J. W. (2000). Early years in machine translation. *Early Years in Machine Translation*, p. 1–411. Publisher : John Benjamins Publishing Company.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, p. 177–180, Prague, Czech Republic : Association for Computational Linguistics.

KOEHN P., OCH F. J. & MARCU D. (2003). *Statistical phrase-based translation*. Rapport interne, University of Southern California Marina Del Rey Information Sciences Inst.

SONG R. (2022). Analysis on the Recent Trends in Machine Translation. *Highlights in Science, Engineering and Technology*, **16**, 40–47.