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

Atelier sur l'Analyse et la Recherche de Textes Scientifiques, CORIA-TALN 2023

Contexte

Objectif : Explorer l'histoire de la traduction automatique

Ressource utilisée : Istex (API, Istex-DL)

Outils : Lodex, web services Inist

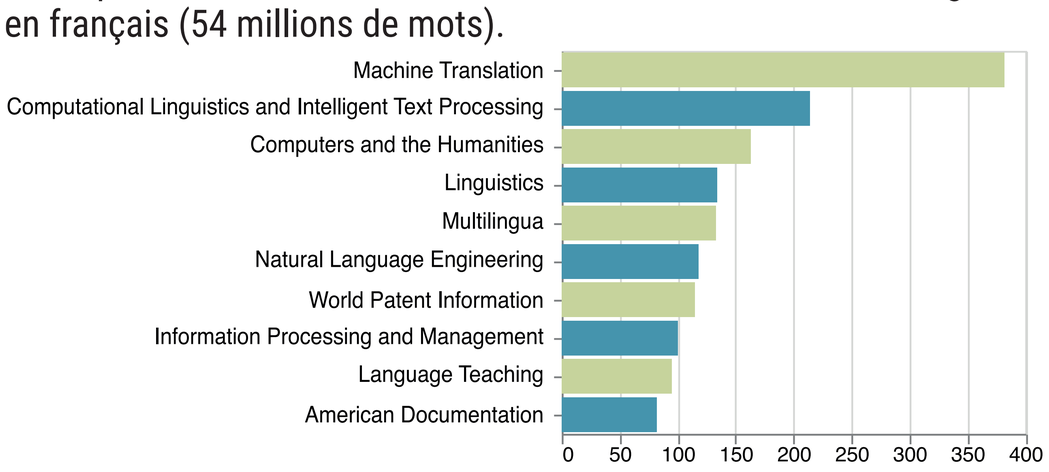
Motivations sociale, scientifique et politique : dynamique nationale et européenne autour de la traduction automatique (ANR-22-MaTOS-0033 ; Fiorini *et al.*, 2020 ; OPERAS ; Helsinki Initiative, 2019)

Constitution de corpus

Procédure itérative (de Salabert & Barreaux, 2020)

1. Interrogation d'Istex (syntaxe Lucene)
2. Détection du bruit dans Lodex
3. Révision de la requête

Le corpus *Machine Translation* contient **7 160 documents** en anglais et en français (54 millions de mots).



Dix revues les plus fréquentes

ISTEX

Initiative d'excellence en Information Scientifique et Technique

En chiffres

- **27 millions** de publications scientifiques
- **51** langues
- **41** corpus éditeurs
- **700** ans de publications

Son contenu

- Textes intégraux aux formats d'origine et standardisés (XML TEI)
- Textes nettoyés
- Métadonnées
- Enrichissements : entités nommées (Unitex), termes (Teeft), structuration du pdf (Grobid), domaines scientifiques (Nb, Multicat)

Linked Open Data EXperiment

Lodex est un logiciel open source créé pour les besoins du projet Istex afin de valoriser ses données structurées (Gregorio *et al.*, 2019).

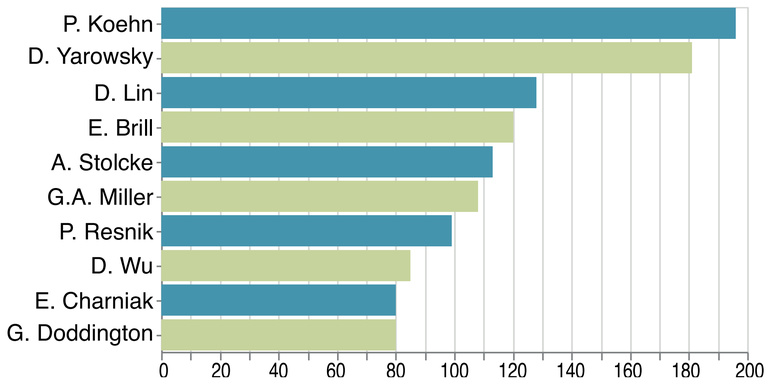
Il permet de concevoir des sites web offrant des interfaces pour explorer visuellement un jeu de données (CSV, JSON, etc.) au travers de tableaux de bord dynamiques présentant des indicateurs bibliométriques.

Lodex offre également la possibilité d'utiliser des web services afin d'enrichir les documents à l'aide de programmes d'analyse, de curation, d'annotation et d'indexation.

Résultats de l'exploration

Indicateurs bibliométriques

Exploitation des métadonnées



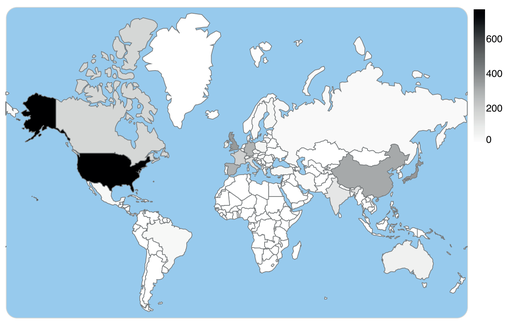
Auteurs les plus cités

Grobid permet de détecter les références bibliographiques et de les structurer dans un format interprété par Lodex.

Utilisation de deux web services :

1. Le web service de découpage d'adresses retourne une adresse au format texte en tableau de champs

2. Le web service *Pays et subdivisions* normalise les graphies des pays en exploitant le thésaurus Loterre

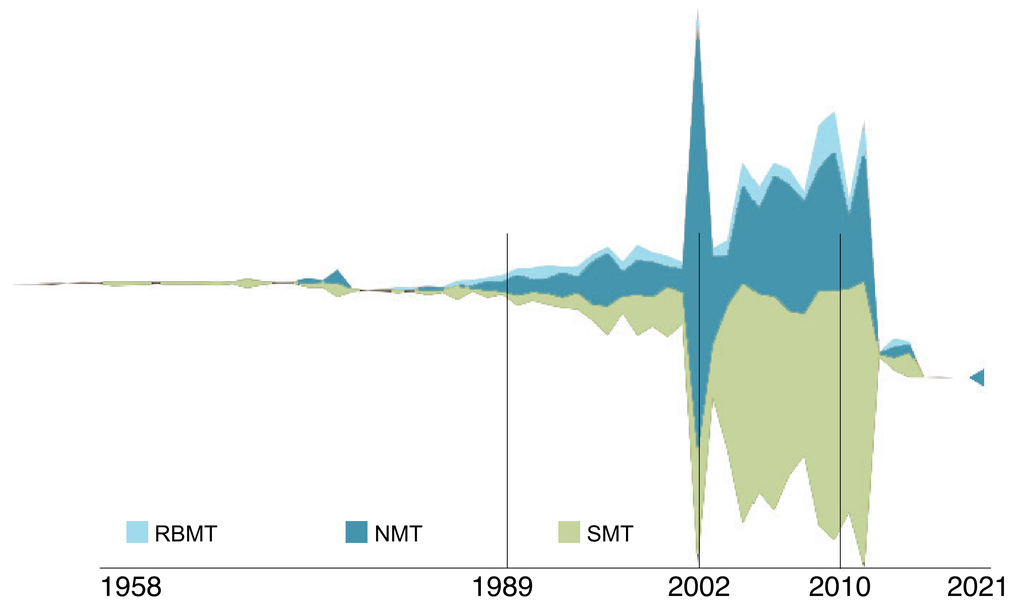


Cartographie des affiliations

Diachronie de la traduction automatique

Exploitation du texte intégral

- Création d'une ressource terminologique permettant d'annoter le corpus
- Annotation du corpus grâce à une feuille de style XSL



Évolution diachronique des approches de la traduction automatique

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. *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 »*. DOI : 10.52949/20.

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

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.