

Analyse de textes : projecteur sur la collecte des corpus

Pierre Zweigenbaum¹ — dialogue avec Alexandra Bensamoun²

¹Université Paris-Saclay, CNRS, LISN

²Professeure de droit, Université Paris-Saclay

*Text and data mining :
pratiques scientifiques et droit applicable*

DATAIA Seminars, en ligne, 3/3/2021



→ LISN



- 1 Introduction: LISN, traitement automatique des langues
- 2 Analyse d'œuvres diverses
- 3 Sites web

1 Introduction: LISN, traitement automatique des langues

- LISN, CNRS, Université Paris-Saclay

2 Analyse d'œuvres diverses

- Analyse d'articles scientifiques
- Articles de presse grand public
- Romans libres de droits
- Articles encyclopédiques : Wikipédia

3 Sites web

- Forums
- Sites web variés

LISN: An Interdisciplinary Laboratory for Digital Sciences

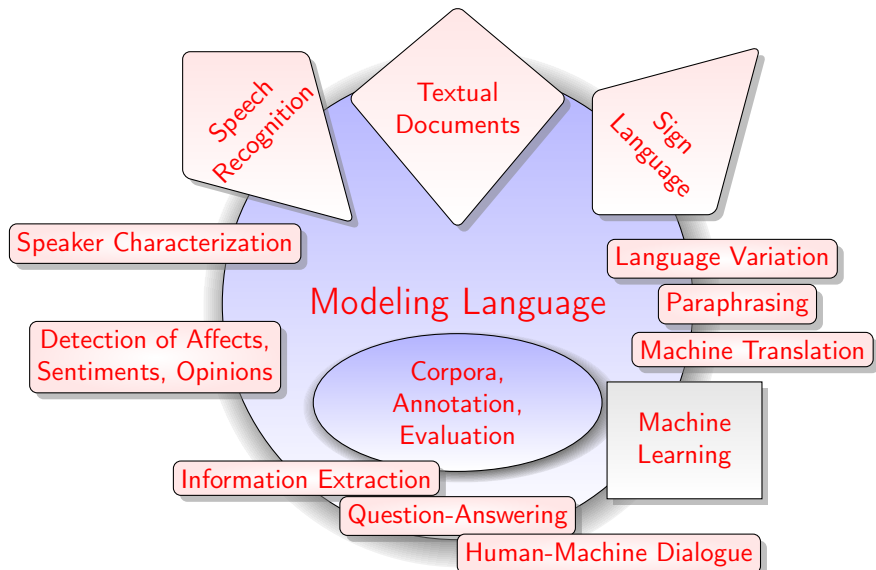
<https://www.lisn.upsaclay.fr/>

A joint **CNRS** Research Unit on the **Université Paris-Saclay** Campus: Belvédère buildings



LISN's Language Sciences and Technologies Department

2 teams, about 70 people including 28 permanent researchers



Annotated Corpora and Shared Tasks

Collection and annotation of text corpora for natural language processing

- Natural Language Processing needs annotated text
 - System development / training
 - System evaluation
 - Organization of shared tasks

1 Introduction: LISN, traitement automatique des langues

- LISN, CNRS, Université Paris-Saclay

2 Analyse d'œuvres diverses

- Analyse d'articles scientifiques
- Articles de presse grand public
- Romans libres de droits
- Articles encyclopédiques : Wikipédia

3 Sites web

- Forums
- Sites web variés

Titres d'articles : corpus Quaero médical

Entraînement de systèmes de détection d'entités médicales

- MEDLINE paper titles (and EMEA drug information)
- Entity mentions, 10 types, UMLS concepts (<https://quaerofrenchmed.limsi.fr/>)
- CLEF eHealth 2016 shared task



- Aurélie Névéol, Cyril Grouin, Thierry Hamon, Pierre Zweigenbaum *et al.*

¹ Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. Fourth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing -

- Entraînement de systèmes de traduction automatique
- Corpus parallèle
 - Trouver des articles avec des résumés en deux langues
 - WMT 2017–2020 biomedical shared tasks
 - Aurélie Névéol *et al.*
- Corpus monolingue
 - Collecte de résumés d'articles en français
 - Entraînement de modèles de langue
- Les résumés sont accessibles sans barrière

Articles en texte intégral

Constitution de plongements lexicaux pour le domaine biomédical

- Collecte d'articles scientifiques en texte intégral (texte brut)
- Base PubMed Central Open Access (PMC-OA)
 - <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>
- The PMC Open Access Subset some or all openaccess content is a part of the total collection of articles in PMC. The articles in the OA Subset are made available under a Creative Commons or similar license that generally allows more liberal redistribution and reuse than a traditional copyrighted work.
- Licences CC0, CC BY, CC BY-SA, etc.

Quaero Old Press Extended Named Entity corpus

L'un des plus grands corpus annotés en entités nommées pour le français

- Collecte d'articles sur Gallica
 - Le Temps, La Croix, Le Figaro
 - 76 journaux publiés en 1890–1891
- Au catalogue ELRA :
<http://catalog.elra.info/en-us/repository/browse/ELRA-W0073/>
- Détection d'entités nommées : personne, lieu, organisation, etc.

DEFT 2011 : de quand date ce texte ?

- Collecte d'extraits d'articles de journaux de 1800 à 1944 sur Gallica
- Le Journal des Débats, Le Journal de l'Empire, Le Journal des Débats politiques et littéraires, La Croix, Le Figaro, La Presse, Le Temps
- Découpage en segments de 300 ou 500 mots
 - jeté à terre, alors les misérables qui répandent ces machines ont le temps de se sauver avant que l'alarme «oit donnée. » Oa écrit de Bruxelles, le 25 août « L'été de <annee/> marquera certainement dans les annales des années orageuses comme des grandes convulsions de la nature. La fréquence et l'intensité des orages, des ouragans.des tempêtes de toute espèce laissent peu ou point d'exemples sous ce rapport. » ...
- Déterminer automatiquement la décennie de publication de chaque segment
 - <https://deft.limsi.fr/2011/>

Le Monde, l'Est Républicain

Des corpus sur catalogue

- Exemples d'applications :
 - Annotation en parties du discours
 - Analyse syntaxique
- Le Monde
 - Catalogue ELRA
 - **Le Monde.fr**
 - <http://catalog.elra.info/en-us/repository/browse/ELRA-W0015/>
 - Diffusion restreinte
- L'Est Républicain
 - Catalogue Ortolang
 - **L'EST RÉPUBLICAIN**
 - https://www.ortolang.fr/market/corpora/est_republicain
 - Licence CC-BY-NC
 - Il est possible que l'accès à certains fichiers de cette ressource soit réservé uniquement aux utilisateurs connectés et/ou aux membres de l'ESR.

Alignement de phrases parallèles

Collecte de livres en français et de leur traduction en anglais

- Objectif (par exemple) : alignement des phrases entre *Cinq semaines en ballon* et sa traduction en anglais
- Livres en français : Association des bibliophiles universels
 - <http://abu.cnam.fr/>
 - [œuvres du domaine public francophone](#)
- Livres en anglais : projet Gutenberg
 - <http://www.gutenberg.org/>
 - Project Gutenberg eBooks may be freely used in the United States because most are not protected by U.S. copyright law. They may not be free of copyright in other countries. Readers outside of the United States must check the copyright terms of their countries before accessing, downloading or redistributing eBooks. We also have a number of copyrighted titles, for which the copyright holder has given permission for unlimited non-commercial worldwide use.

Recherche de traductions dans des corpus comparables

Collecte de textes dans deux langues

- Objectif (par exemple) : recherche de traductions de phrases dans les textes collectés
 - Campagnes d'évaluation BUCC 2017, 2018
 - <https://comparable.limsi.fr/bucc2018/bucc2018-task.html>
- Collecte de toutes les pages Wikipédia en français, en anglais
- Droit d'auteur : les textes sont disponibles sous licence Creative Commons attribution, partage dans les mêmes conditions ; d'autres conditions peuvent s'appliquer.

1 Introduction: LISN, traitement automatique des langues

- LISN, CNRS, Université Paris-Saclay

2 Analyse d'œuvres diverses

- Analyse d'articles scientifiques
- Articles de presse grand public
- Romans libres de droits
- Articles encyclopédiques : Wikipédia

3 Sites web

- Forums
- Sites web variés

DEFT 2013 : cette recette est-elle difficile ?

- Collecte de recettes de cuisine sur un grand site francophone
 - Séparation des méta-informations, du corps de la recette
- Organisation d'une campagne d'évaluation
 - Niveau de difficulté de la recette ?
 - Entrée, plat, dessert ?
 - Quel est le titre de cette recette ?
 - Quels sont les ingrédients de cette recette ?
- <https://deft.limsi.fr/2013/>

Vigi4med : pharmacovigilance

Quels effets secondaires les internautes rapportent-ils pour quels médicaments ?

- Collecte de messages sur des forums de santé francophones

20/02/2015
Homme, 48

Salazopyrine (sulfasalazine)

Spondylarthrite ankylosante

Satisfaction générale



Après deux ans de douleurs dans le dos attribuées par erreur à une possible hernie discale, un Rhumatologue m'a diagnostiqué une Spondylarthrite Ankylosante. Je suis porteur du HLAB27. Je prends depuis un an de la Salazopyrine (2x par jour, deux comprimés de 500 mg soit 2 grammes par jour). Dans les trois premiers mois de traitement, j'ai eu des

coups de fatigue soudains, des baisses de la concentration, quelques somnolences. Cette période passée, je n'ai plus eu d'effets secondaires (à part une coloration jaune intense de l'urine). La douleur au dos est encore présente mais a considérablement diminué, je ne prends plus d'antalgique (paracétamol) ou d'anti-inflammatoire (cartrex) (ou de façon très épisodique). La Salazopyrine a changé ma vie car je scénarisais une progression de l'inflammation et une gêne au quotidien de plus en plus présente.

Satisfaction sur

Efficacité



Quantité effets secondaires



Gravité effets secondaires



Facilité d'emploi



- Détection d'effets indésirables de médicaments

- nom de médicament
- signe, symptôme, diagnostic
- relation causale assertée entre les deux

- Source d'études pour les pharmacovigilants

Paracrawl : collecte de textes parallèles

Entraînement de systèmes de traduction automatique

- Détection de sites web qui ont des pages en plusieurs langues
- Identification de paires de pages traduction l'une de l'autre
- <https://www.paracrawl.eu/>
 - We do not own any of the text from which these data has been extracted
 - We license the actual packaging of these parallel data under the Creative Commons CC0 license ("no rights reserved").

- 1 Introduction: LISN, traitement automatique des langues
 - LISN, CNRS, Université Paris-Saclay
- 2 Analyse d'œuvres diverses
 - Analyse d'articles scientifiques
 - Articles de presse grand public
 - Romans libres de droits
 - Articles encyclopédiques : Wikipédia
- 3 Sites web
 - Forums
 - Sites web variés

Conclusion: Corpus pour le traitement automatique des langues

- Des corpus libres
- Des corpus sous contrat
- Des corpus au statut moins clair

Merci !