

# Rapport du projet Dictionnaire et néologisme

Yuyan QIAN  
21108372  
Françoise Guérin et Gaël Lejeune

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Materiel</b>	<b>2</b>
<b>3</b>	<b>Méthodologies</b>	<b>3</b>
3.1	Lecture et prétraitement des données . . . . .	4
3.2	Intersection, union et difference . . . . .	4
3.3	Visualisation par jour . . . . .	5
<b>4</b>	<b>Limites</b>	<b>5</b>
<b>5</b>	<b>Résultats</b>	<b>6</b>
<b>6</b>	<b>Perspectives</b>	<b>7</b>

## 1 Introduction

Dans le domaine de la linguistique, le phénomène d'emprunt linguistique est un sujet intéressant qui consiste à intégrer des morphèmes d'une langue étrangère dans une langue donnée. Ce phénomène est particulièrement visible dans les journaux français, où l'on observe une utilisation fréquente de mots anglais en Une. Cependant, le français a également tendance à utiliser des calques, c'est-à-dire des traductions littérales de termes complexes provenant d'autres langues. Les deux méthodes sont utilisées pour enrichir le vocabulaire français, mais elles ont des effets différents sur la langue française. Ainsi, l'objectif de cette étude est de voir l'impact de l'anglais sur les journaux français. Je chercherai à déterminer si cette tendance est en augmentation ou en diminution au fil des années. Pour ce faire, je vais m'appuyer sur un corpus journalistique issu de l'Est Républicain, proposé par ORTOLANG (Open Resources and Tools for Language), de 1999 à 2011. Une méthode d'analyse statistique sera utilisée pour traiter le corpus en XML et extraire les informations pertinentes. Le corpus étudié sera ainsi libre d'utilisation sans but commercial.

Dans ce projet, Je vais étudier les emprunts linguistiques dans les journaux français et comparer les emprunts de l'anglais et les néologismes. Ensuite, je voudrais évaluer la fréquence des emprunts linguistiques par rapport à celle de la néologie. Je suppose que l'emprunt de l'anglais est de plus en plus important au fil des années, en raison de l'innovation croissante dans le monde anglophone.

Les morphèmes empruntés ne réfèrent pas seulement à une mutation linguistique, mais aussi une évolution sociale. En un mot, ma problématique porte sur l'évolution de l'emprunt de l'anglais dans les journaux français : est-il de plus en plus fréquent au fil des années ou au contraire en diminution ? Est-il possible d'observer cette tendance en évaluant le pourcentage d'emprunts linguistiques par rapport à celui de la néologie ?

## 2 Materiel

Avant de commencer la présentation, je voudrais d'abord définir de manière précisée le corpus que j'ai utilisé. ORTOLANG a conclu un partenariat avec la société du journal L'Est Républicain pour offrir un corpus journalistique aux chercheurs après son traitement informatique. Initialement, ORTOLANG a publié les données correspondant aux années 1999 à 2003, mais une nouvelle version enrichie de l'ensemble des éditions de 2006 à 2011, y compris la variante vosgienne de l'Est Républicain, Vosges Matin (VM), est désormais disponible. Le corpus est ainsi disponible en XML, libre d'utilisation sans but commercial, et peut être téléchargé sur le site d'ORTOLANG. Et il est monolingue en français qui me permet d'analyser l'influence de l'anglais.

La totalité des données de cette ressource soit un maximum de 15,1 Go, ece qui est considérable et peut être difficile à traiter et à télécharger. Pour mon projet, j'ai choisi de ne travailler qu'avec une partie de ce corpus, à savoir les données de 1999, 2002, 2003, 2006, 2008, 2010 et 2011. Il convient également de noter que la taille des données pour

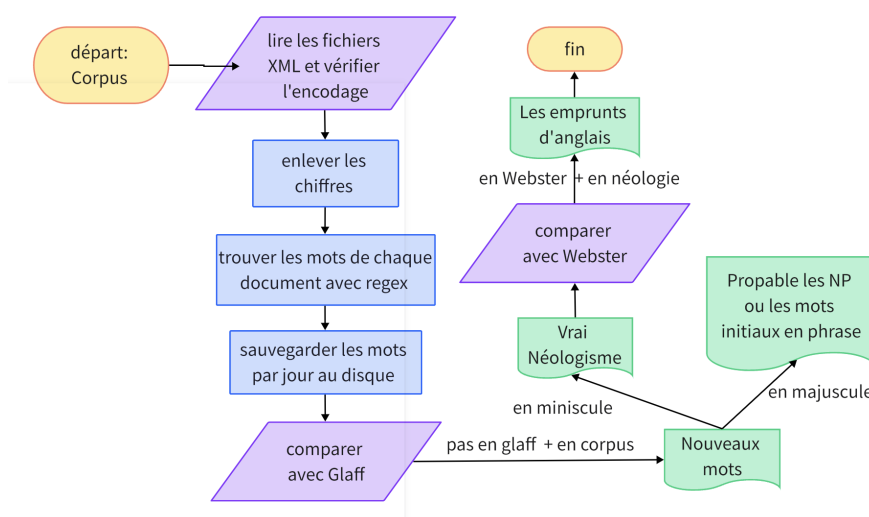
chaque année est très variable. Pour mieux visualiser les informations, j'ai utilisé une cellule de code pour calculer le nombre de jours de journal pour chaque année.

Voici les informations détaillées de ce corpus<sup>1</sup> utilisé dans mon projet :

Corpus	Collection	Le premier jour	Le dernier jour	Taille
Annee1999	132 jours	1999-05-17	1999-09-30	249,4 MB
Annee2002	169 jours	2002-01-02	2002-12-27	318,5 MB
Annee2003	54 jours	2003-01-02	2003-02-24	101,3 MB
Annee2006	357 jours	2006-02-01	2006-12-31	1,89 GB
Annee2008	361 jours	2008-02-01	2008-12-31	1,72 GB
Annee2010	170 jours	2010-02-01	2010-05-31	672,2 MB
Annee2011	13 jours	2011-02-01	2011-02-14	46,7 MB

Le tableau ci-dessus affiche que le corpus étudié est composé de 1256 jours de journal, pour une taille totale de 6,1 GB. Il est intéressant de noter que la taille des données varie considérablement d'une année à l'autre. Cette divergence peut avoir une incidence sur les résultats de l'analyse statistique. En effet, une plus grande quantité de données pour une année donnée peut accroître les possibilités de trouver des emprunts d'anglais. Toutefois, il est important de noter que les résultats attendus seront fiabilisés par le calcul de la fréquence, qui permet de réduire l'impact de la quantité de données sur les résultats.

### 3 Méthodologies



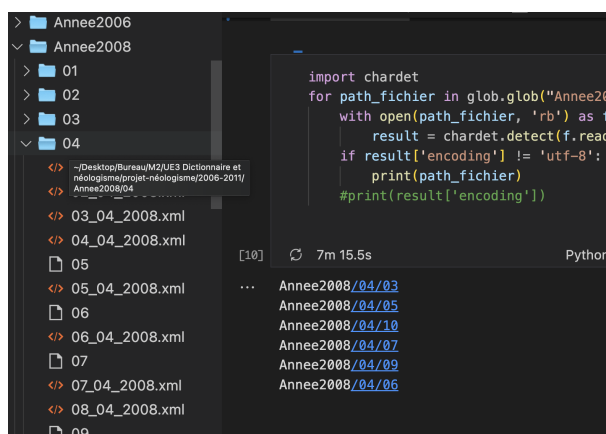
Voici un plan des étapes essentielles dans mon traitement des données. Je vais vous présenter les méthodes de chaque étape dans les parties suivantes. En bref, ce projet se compose de de trois tâches principales : la lecture et le prétraitement des données pour

1. [https://drive.google.com/drive/folders/1xemST2oD7VGd0Wc8dme0bx4rmPon4hMA?usp=share\\_link](https://drive.google.com/drive/folders/1xemST2oD7VGd0Wc8dme0bx4rmPon4hMA?usp=share_link)

trouver les mots, la comparaison avec Glaff<sup>1</sup> pour trouver les néologismes et la comparaison entre la néologie et la dictionnaire anglaise Webster<sup>2</sup>. Pour voir les emprunts, il me faut d'abord trouver une référence de mots. Le plus pratique est de consulter une dictionnaire numérique. Ainsi, la dictionnaire française que j'utilise dans ce projet est un lexique du français construit à partir du Wiktionnaire, la branche francophone de Wiktionary, paartagé dans les Mêmes Conditions 3.0 non transposé (CC BY-SA 3.0). Elle contient 1406857 entrées, soit 1082688 types. En plus, la dictionnaire anglaise Webster contient 86036 types. Le dictionnaire peut être trouvé en texte brut dans le fichier dictionary.txt. Son contenu est soumis aux termes de la licence du Projet Gutenberg.

### 3.1 Lecture et prétraitement des données

Tout d'abord, j'ai commencé à lire les données anuelle en codage UTF-8 sauf pour l'année 2008 qui se compose aussi des fichiers en byte. Dans ce cas, j'ajoute une condition pour vérifier et détecter l'encodage des fichiers.



```
import glob
import chardet

for path_fichier in glob.glob("Annee2008/*"):
    with open(path_fichier, 'rb') as f:
        result = chardet.detect(f.read())
        if result['encoding'] != 'utf-8':
            print(path_fichier)
            #print(result['encoding'])
```

[10] 7m 15.5s Python

...  
Annee2008/04/03  
Annee2008/04/05  
Annee2008/04/10  
Annee2008/04/07  
Annee2008/04/09  
Annee2008/04/06

Ensuite, j'ai utilisé la bibliothèque BeautifulSoup pour lire les données XML et extraire le texte de chaque article. La tokenisation des mots est réalisée par les expressions régulières de la bibliothèque re.

Après avoir vu le corpus, j'ai trouvé plusieurs mots composés, comme *publicité-marketing* et les mots qui contiennent les caractères spéciaux. Pour résoudre ce problème, j'ai utilisé un pattern ci-dessous.

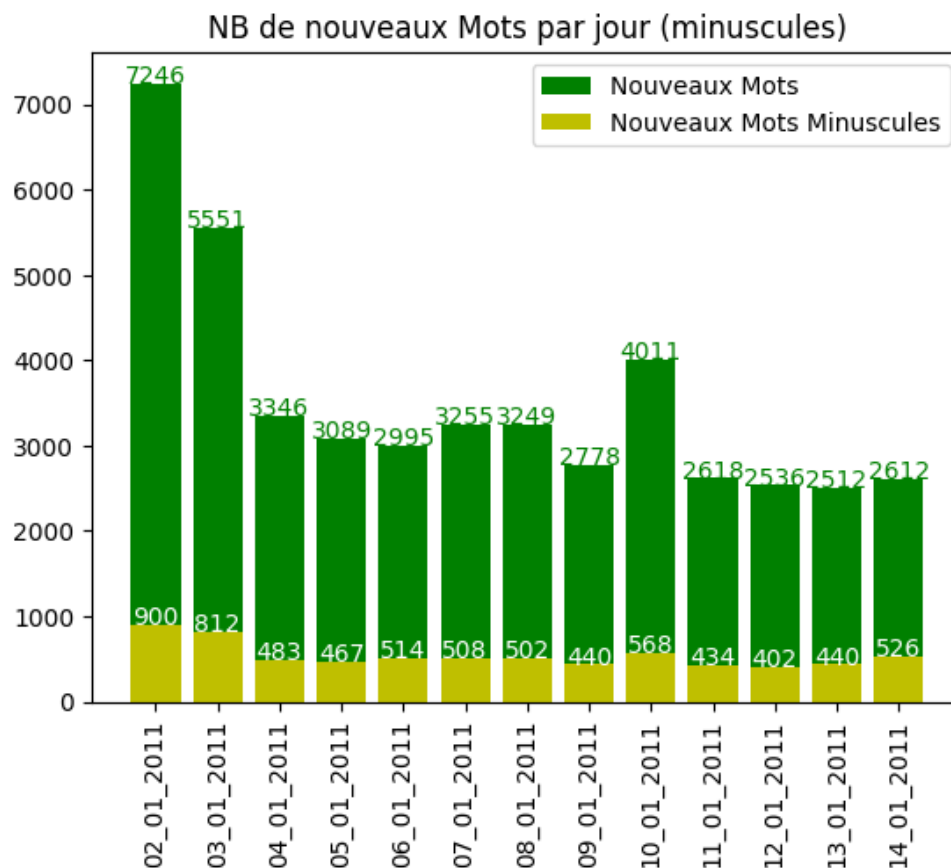
```
re.findall("[A-ZÀ-ÿ][A-Z0-9.]+|[cCdDjJlLmnNsSt]'|[qQ]u'|\w+[\w'-]+\w+|\w\w+)", texte)
```

### 3.2 Intersection, union et difference

Il est pratique et facile de réaliser une comparaison entre les mots du corpus et ceux de la dictionnaire à l'aide des fonctions intersection, union et difference de Python. Pour

1. <http://redac.univ-tlse2.fr/lexiques/glaff.html>  
2. <https://github.com/adambom/dictionary>

vérifier les vrais néologismes, il me faut distinguer les mots capitalisés et les mots non capitalisés qui sont tous les deux identifiés comme la néologie. En effet, les mots capitalisés sont souvent des noms propres, des titres ou des sigles. Pour les différencier, j'ai utilisé la fonction `isupper()` de Python.



La figure présente le nombre des mots nouveaux et les vrais néologismes par jour en 2011. La décalage affirme qu'il est nécessaire de faire la distinction entre les noms propres ou les sigles et les vrais néologismes.

### 3.3 Visualisation par jour

Au début, j'ai essayé de visualiser les néologismes par jour. Cependant, il y a trop de données à afficher que les histogrammes deviennent invisibles. De sorte, j'ai décidé de visualiser les données par année qui est plus organisée.

## 4 Limites

La référence de deux dictionnaires n'est pas suffisante pour trouver tous les emprunts et les vérifier. Dans un côté, les deux ne sont pas complètes à inclure tous les mots français

ou anglais, particulièrement Glaff qui contient pas les complexes unitaires. Dans un autre côté, l'intégration de l'emprunt est ignorée dans ce projet parce que c'est un travail trop complexe à réaliser.

```
<div type="article">
  <div type="texte">
    <div>
      <div>
        <head type="h4">Le Dakar, c'est parti !</head>
        <p>Après les vérificationstechniques, les choses sérieuses
```

En plus, le corpus a des erreurs d'origine. Par exemple, en 2011 *vérifications* et *techniques* sont collés en un seul token *vérificationstechniques* dans le corpus d'origine. Même si cette nouvelle morphème est exclue en comparant avec Webster, il reste toujours à considérer comme une manque.

Enfin, en visualisation, la néologie de chaque année est séparée. Ici, j'ai pas pris en compte de l'intégration de l'emprunt. Donc, il est possible que les emprunts sont comptés plusieurs fois dans les années différentes.

## 5 Résultats

Le tableau ci-dessous affiche le nombre de néologismes trouvés dans chaque année. Il est intéressant de noter que le nombre de néologismes trouvés dans chaque année est très différent. Cela peut être dû à la quantité de données disponibles pour chaque année. Toutefois, il est important de noter que la taux d'emprunt à l'anglais est plutôt stable. L'emprunt est une matrice particulièrement productive qui participe activement à la néologie lexicale. Par exemple, entre 2006 et 2008, l'ajout de 10 000 nouveaux mots n'a entraîné qu'environ 200 emprunts. Et entre 1999 et 2002, moins de 10 000 nouveaux mots a entraîné presque 280 emprunts.

De plus, j'ai observé une lacune dans les traitements. J'ai collectionné les emprunts et les néologismes de chaque année, néanmoins, il est possible que les emprunts sont comptés plusieurs fois dans les années différentes. Par exemple, *stretching* est un emprunt en 1999, mais il est aussi considéré comme un néologisme en 2002. Ainsi, pour réduire l'impact de ce problème, j'ai fait une étape supplémentaire à créer les graphies en excluant les emprunts des années précédentes. Le résultat est affiché dans la figure 2.

Une tendance remarquable dans la figure 2 : le taux d'emprunt par rapport aux néologismes trouvés chute au fil des années. Afin de confirmer les résultats de l'analyse statistique, j'ai lu une autre analyse qualitative des néologismes : *Les néologismes anglais issus de l'emprunt : l'étude diachronique*. Cela correspond à la figure deux : à partir des années 1870, le pourcentage d'emprunts à l'anglais diminue.

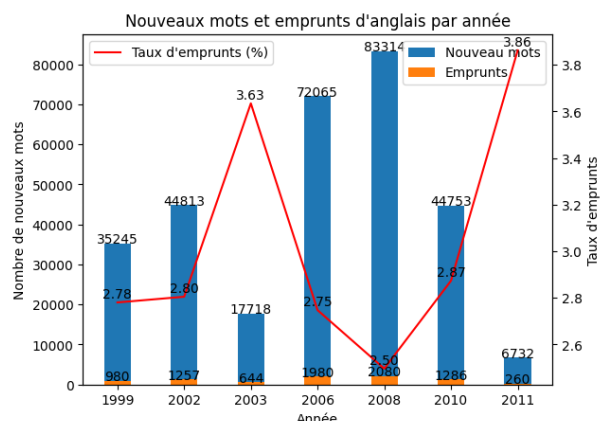


FIGURE 1 – Avec les emprunts des années précédentes

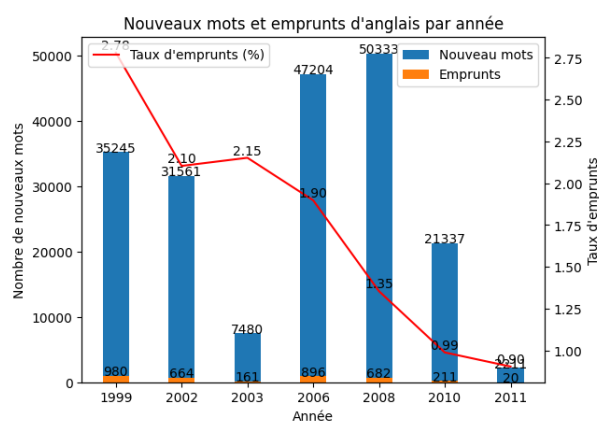


FIGURE 2 – Sans les emprunts des années précédentes

## 6 Perspectives

Dans ce projet, j'ai utilisé des techniques de traitement de texte pour analyser l'impact de l'anglicisme sur la langue française à travers les journaux français. Pour cela, j'ai collecté des données à partir des journaux français de l'année 2008 et j'ai extrait des informations telles que le nombre de mots par jour, le nombre d'emprunts à l'anglais et le nombre total de néologismes. J'ai ensuite utilisé des techniques de visualisation de données pour analyser ces informations. Les grandes données ont été traitées avec beaucoup de temps et d'efforts. Et je voudrais continuer à apprendre comment économiser du temps et de l'effort face à un grand corpus.

Pour améliorer ce projet, il serait intéressant d'élargir l'analyse à une période plus longue et plus complète, par exemple sur plusieurs décennies, afin d'observer les tendances à long terme. Il serait également intéressant de comparer les résultats avec d'autres langues européennes pour voir comment elles sont influencées par l'anglais. D'ailleurs, l'affinement de la détection des néologismes et des anglicismes pourrait être amélioré en ajoutant les étapes plus avancées. Ce que j'ai fait dans ce projet est seulement un commencement diachronique à voir l'emprunt linguistique.

## Références

- [1] ATILF and CLLE. Corpus journalistique issu de l'est républicain, 2020. ORTOLANG (Open Resources and TOols for LANGuage) –[www.ortolang.fr](http://www.ortolang.fr).
- [2] Denis Jamet and Adeline Terry. Les néologismes anglais issus de l'emprunt : l'étude diachronique. *ELAD-SILDA*, 1, 2018.