06/05/2021

L’emprunt à l’anglais dans le journal L’Est Républicain (1999-2011)

Projet du Dictionnaire et Néologisme

Yuyan QIAN

M1 Langue et Informatique, Sorbonne université

Plan

Introduction

Pourquoi ?

Marketing : « Les chercheurs sont parvenus à des mesures de bien-être beaucoup plus prédictives en matière de fidélité »

L'évolution de l'emprunt à l'anglais dans les journaux français :

- plus fréquent ou au contraire en diminution ?
- le pourcentage d'emprunts linguistiques par rapport à celui de la néologie en français ?

L'EST RÉPUBLICAIN

Nancy

ER **La start-up nancéienne Abby lève 1,2 million d'euros**

L'entreprise fondée à Nancy par quatre étudiants du CESI est spécialisée dans l'accompagnement des ...

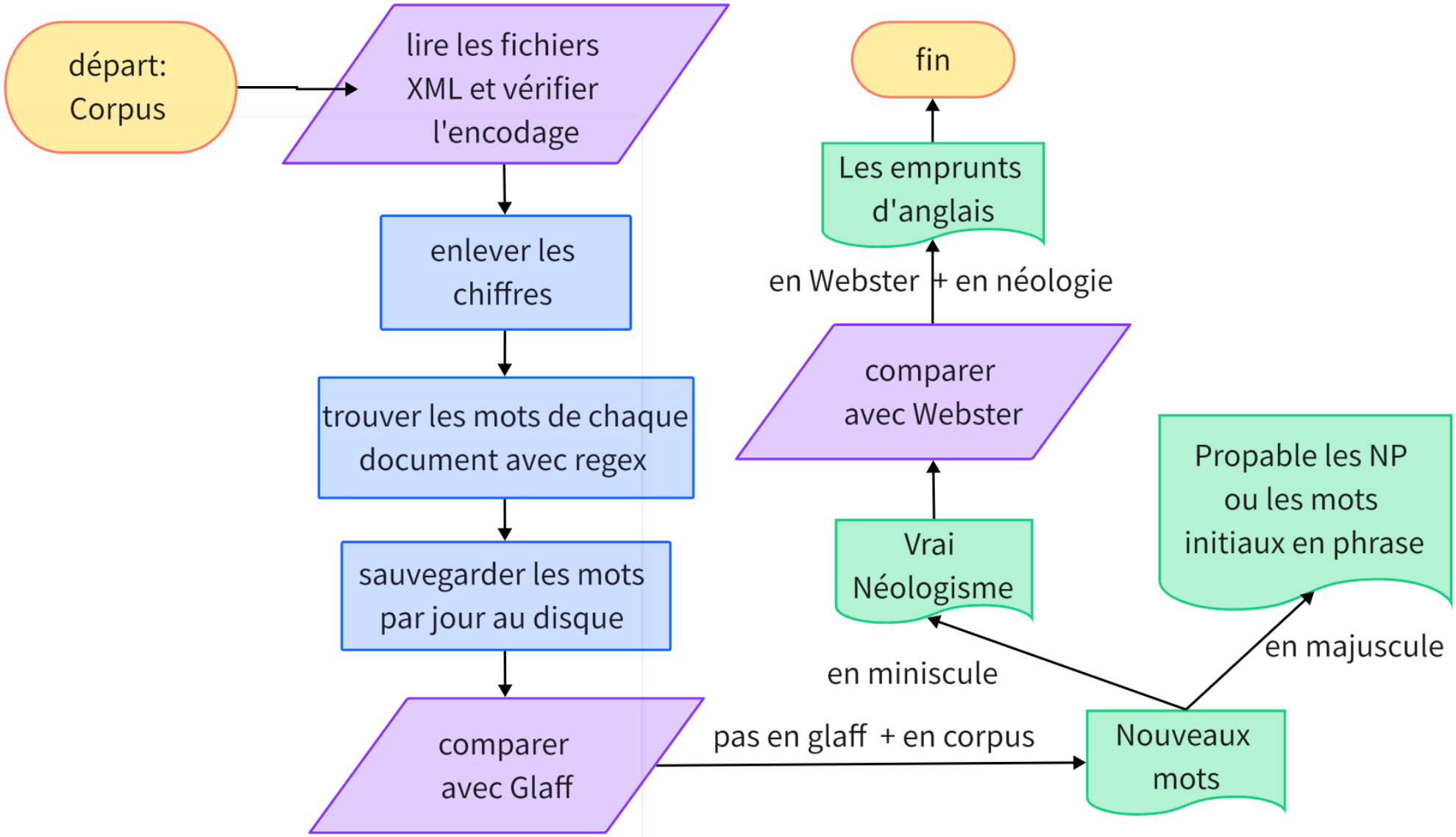
Matériel

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

Dictionnaires

Corpus journalistique issu de l'Est Républicain

1. Glaff (1, 406, 857 entrées) - un lexique du français construit à partir du Wiktionnaire, la branche francophone de Wiktionary, partagé dans la licence CC BY-SA 3.0.
2. Webster (86, 036 entrées) - un texte brut dans le fichier dictionary.txt. Son contenu est soumis aux termes de la licence du Projet Gutenberg.
   Ex: *"BAHAISM":"The religious tenets or practices of the Bahais."*


départ:
Corpus
lire les fichiers XML et vérifier l'encodage
enlever les chiffres
trouver les mots de chaque document avec regex
sauvegarder les mots par jour au disque
comparer avec Glaff
pas en glaff + en corpus
Nouveaux mots
en miniscule
en majuscule
Vrai Néologisme
Propable les NP ou les mots initiaux en phrase
comparer avec Webster
en Webster + en néologie
Les emprunts d'anglais
fin

Méthodologies

Traitements

Vérification d'encodage

```
import chardet
for path_file in glob.glob("Annee2008/*"):
    with open(path_file, 'rb') as f:
        result = chardet.detect(f.read())
    if result['encoding'] != 'utf-8’:
        print(path_file)
         #print(result['encoding’])
Par exemple : certains alias en byte
```

Lecture avec BeautifulSoup

Prétraitement

```
2. Lang (Sui) à 21'' ; 3. Taramarcaz (Sui) à 30''
ui) à 29'' ; 3. Brandau (All) à 37''... 8. Colin (
```

Tokenisation

```
re.findall("([A-ZÀ-ÿ][A-ZÀ-ÿ0-9.]+|
                [cCdDjJlLmnNsSt]’|
                [qQ]u'|\w+[\w'-]+\w+|
                \w\w+)", texte)

Par exemple : BW’s, contre-productive,
                publicité-marketing
```

Méthodologies

Sauvegarde

Réécriture en json

```
try:
    with open(path, 'r') as f:
        old_dict = json.load(f)
except:
    old_dict = {}
new_dict = {**old_dict, **dico}
```

Glaff et Webster

Différence et intersection

Mise à jour de vocabulaire néo

Vérification

Division les tokens en deux groupes :

capitalisé ou pas

Connexion des documents

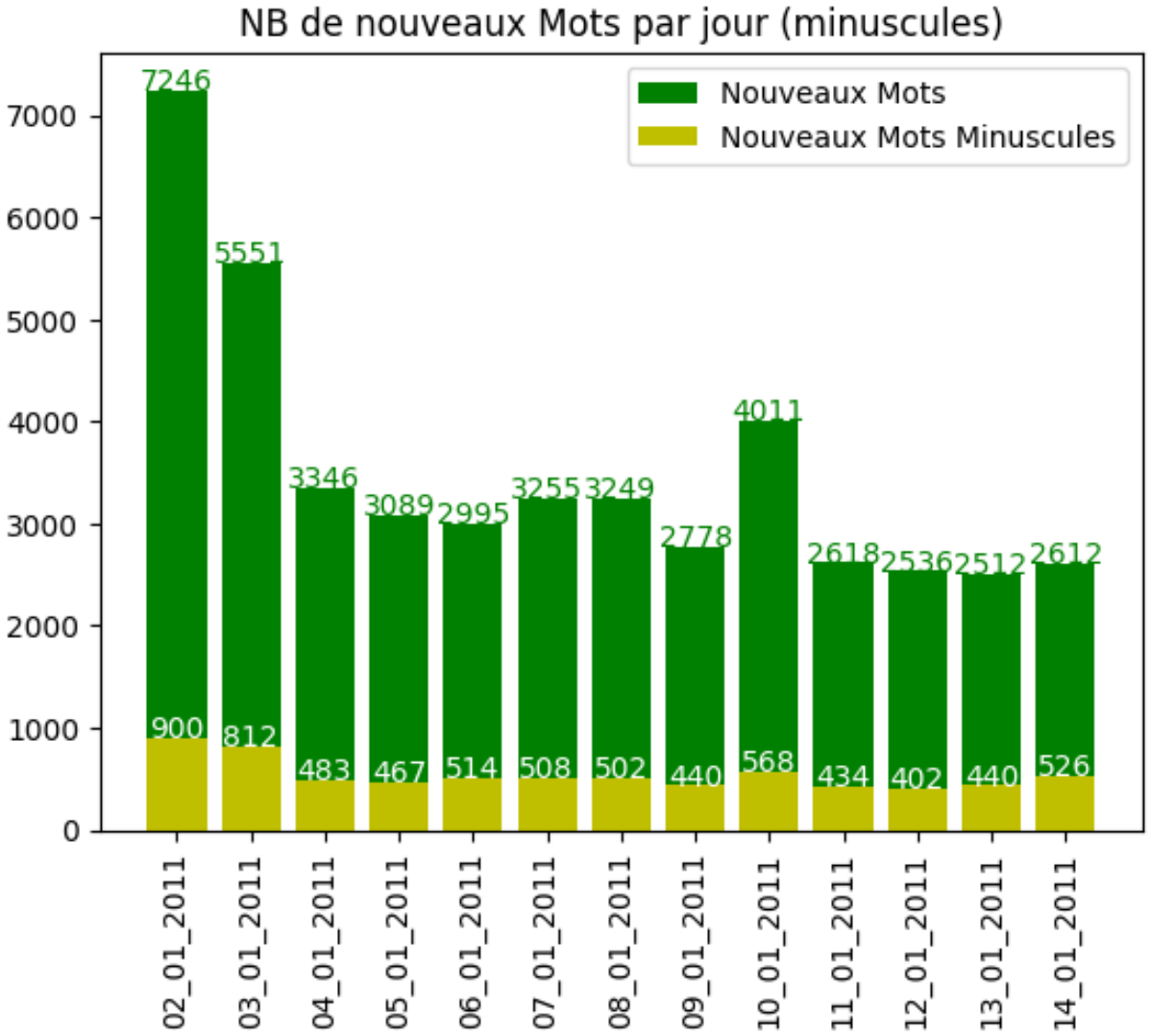
Actualisation des néologismes pour chaque année

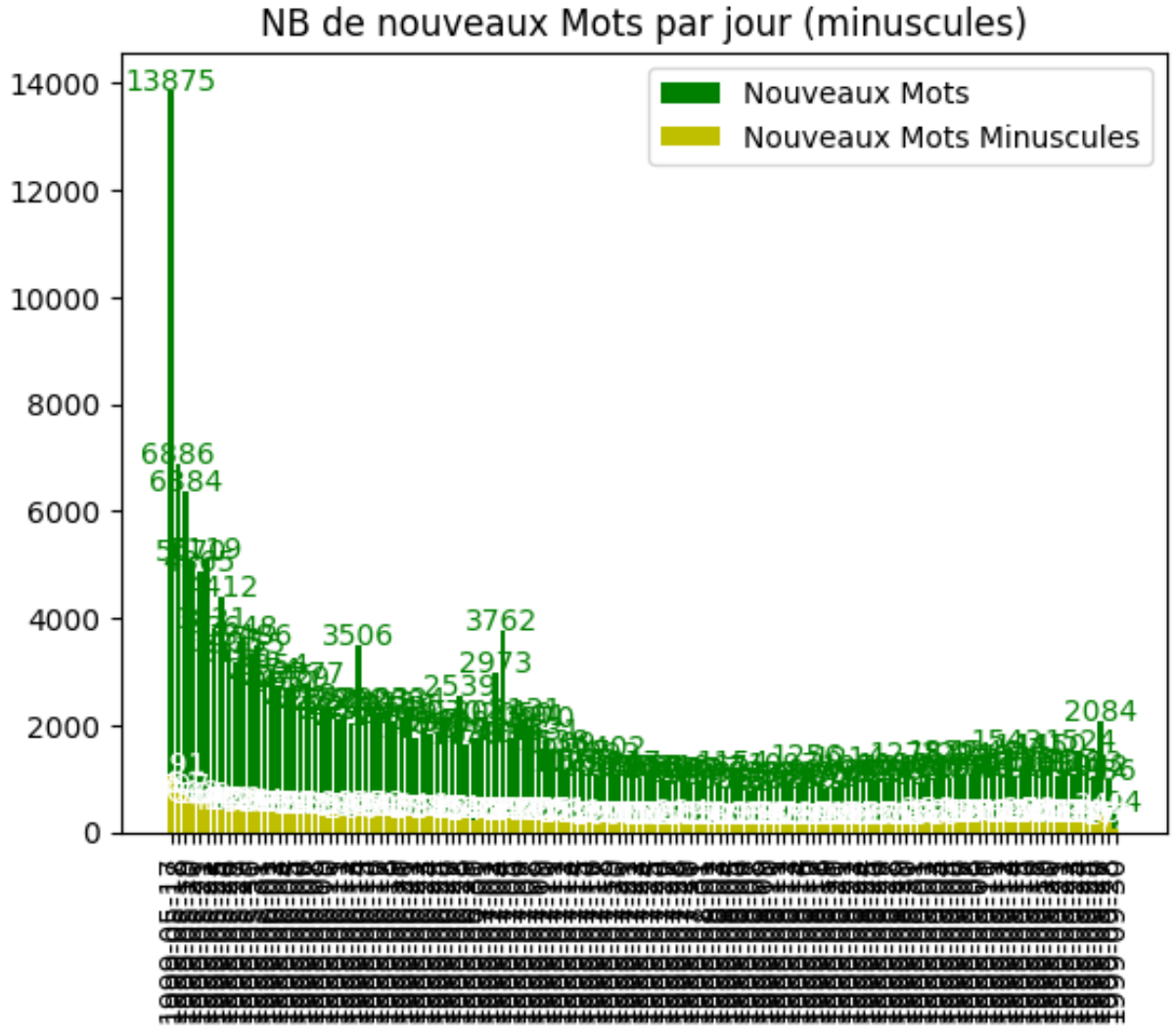
Visualisation

Matplotlib - histogramme et courbe

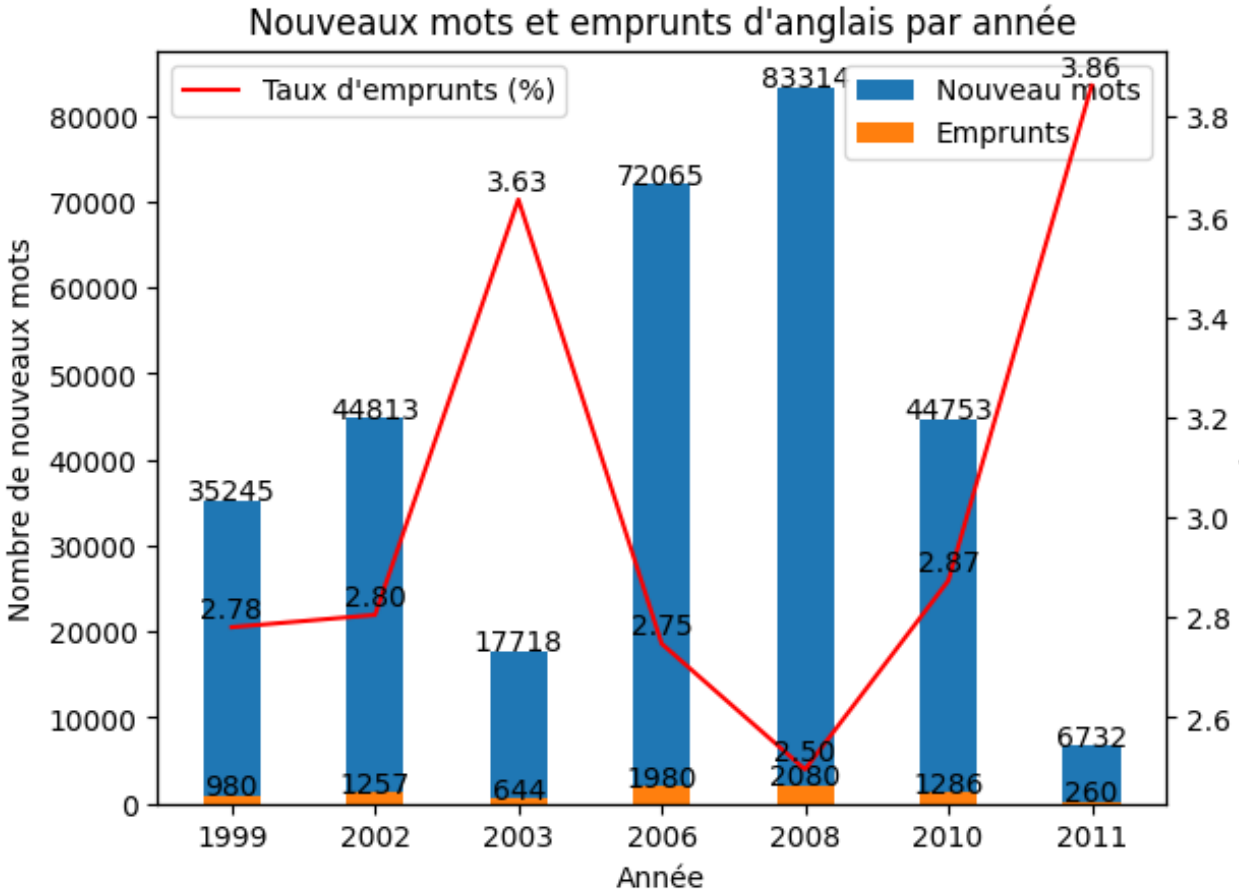
Wordcloud - nuage de mots

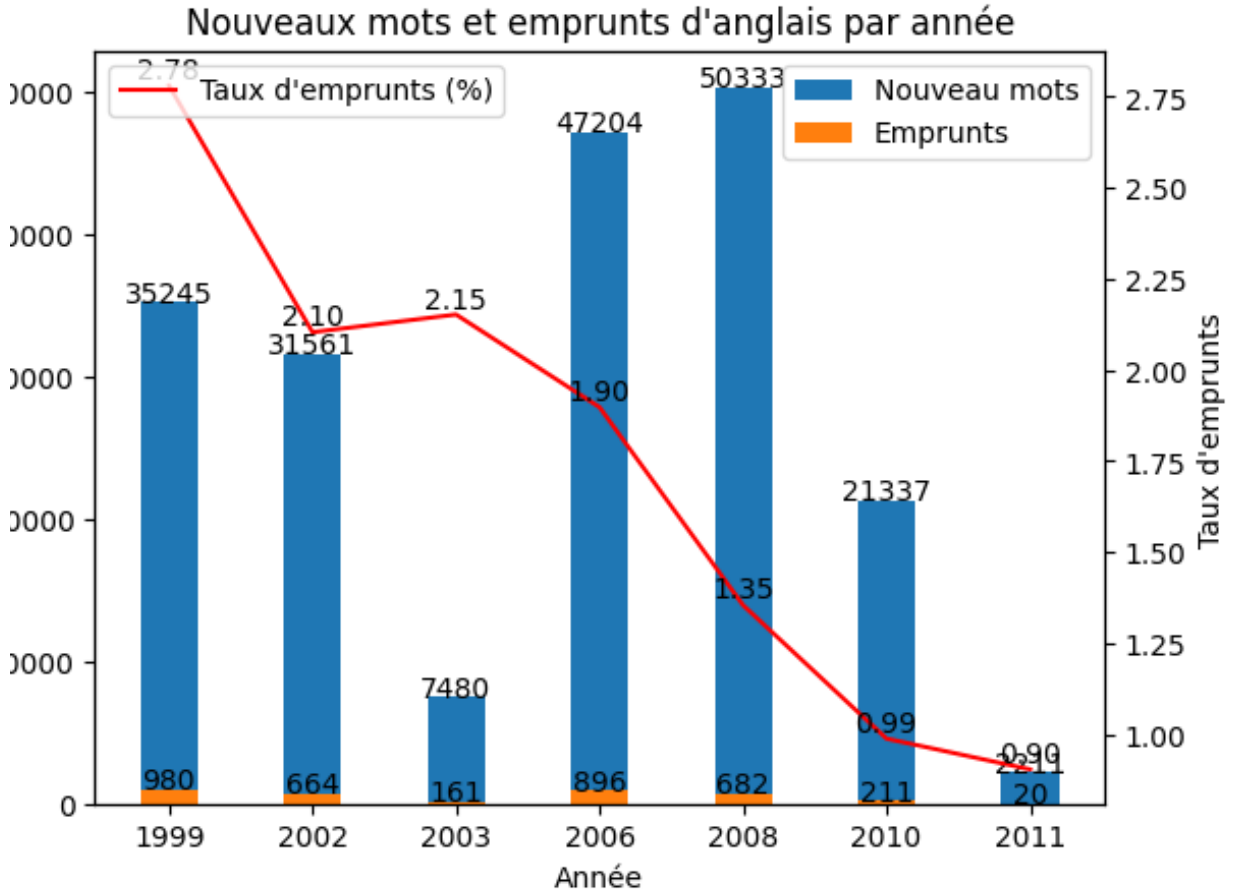
Résultats


NB de nouveaux Mots par jour (minuscules)
Nouveaux Mots
Nouveaux Mots Minuscules
0
1000
2000
3000
4000
5000
6000
7000
7246
5551
3346
3089
2995
3255
3249
2778
4011
2618
2536
2512
2612
900
812
483
467
514
508
502
440
568
434
402
440
526
02_01_2011
03_01_2011
04_01_2011
05_01_2011
06_01_2011
07_01_2011
08_01_2011
09_01_2011
10_01_2011
11_01_2011
12_01_2011
13_01_2011
14_01_2011


NB de nouveaux Mots par jour (minuscules)
Nouveaux Mots
Nouveaux Mots Minuscules
0
2000
4000
6000
8000
10000
12000
14000
13875
6886
6884
3506
3762
2973
2539
2084
91

Résultats


Nouveaux mots et emprunts d'anglais par année
Taux d'emprunts (%)
Nouveau mots
Emprunts
Nombre de nouveaux mots
Taux d'emprunts
Année
1999
2002
2003
2006
2008
2010
2011
35245
44813
17718
72065
83314
44753
6732
980
1257
644
1980
2080
1286
260
2.78
2.80
3.63
2.75
2.50
2.87
3.86


Nouveaux mots et emprunts d'anglais par année
Taux d'emprunts (%)
Nouveau mots
Emprunts
Taux d'emprunts
Année
1999
2002
2003
2006
2008
2010
2011
35245
31561
7480
47204
50333
21337
980
664
161
896
682
211
20
2.10
2.15
1.90
1.35
0.99
0.90

Résultats

blundering
take
verein
begonia
grating
ash
bonfire
dakotas
sell
velutina
bedside
weeding
buz
heliotrope
brush
atmo
steatite
lean
trumpets
double
poetry
dentist
harlequin
gentilesse
stating
parallel
decker
mazarine
primero
lanyard
quickstep
pollinose
parody
rove
forward
starling
misle
shifter
taint
innovative
oidium
perplexity
tod
miller
gid
analysis
arras
hereford
tablet
absentation
spree
speck
gentle
material
monopode
dude
fey
brond
spend
oscillatoria
weapon
coreopsis
crush
sudden
poplin
infinite

Limites

01 Les données du corpus

02 La mélange des vrais néologismes et des NP

03 Les manques d'entrées dans deux dictionnaires

04 La taille déséquilibre des données annuelles

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

EX: L'histoire de Tom-Tom

EX: Glaff ne contient pas les mots-composés

La fréquence servit à réduire l'impact de la quantité de données sur les résultats.

Perspectives

Période plus longue

d’autres langues européennes

Détection plus affinée

Références

[1] ATILF and CLLE. Corpus journalistique issu de l'est républicain, 2020. ORTOLANG (Open Resources and TOols for LANGuage) – www.ortolang.fr.

[2] Denis Jamet and Adeline Terry. Les néologismes anglais issus de l'emprunt : l'étude diachronique. ELAD-SILDA, 1, 2018.

Merci !