TP1 – LTAL

Coste Charly

Introduction

Les documents étudiés sont tirés de Wikipédia, les 7 traitants de l'Église orthodoxe. Ce sujet a été choisi en parcourant la liste des pages longues du wiki Grec avec un script pour déterminer son existence dans les 6 autres langues désirées (find_long_multilingual_wiki_articles.py).

Les scripts d'analyse sont des réécritures en python3 des scripts fournis (analyse_html.py, htmlutils.py et linguistic.py).

L'étude des groupes de mots a été effectuée mais pas la partie coloriage.

Expressions régulières

On s'est servi des constantes du module «re» de python :

'\w' pour un caractère

'\w+' pour un mot

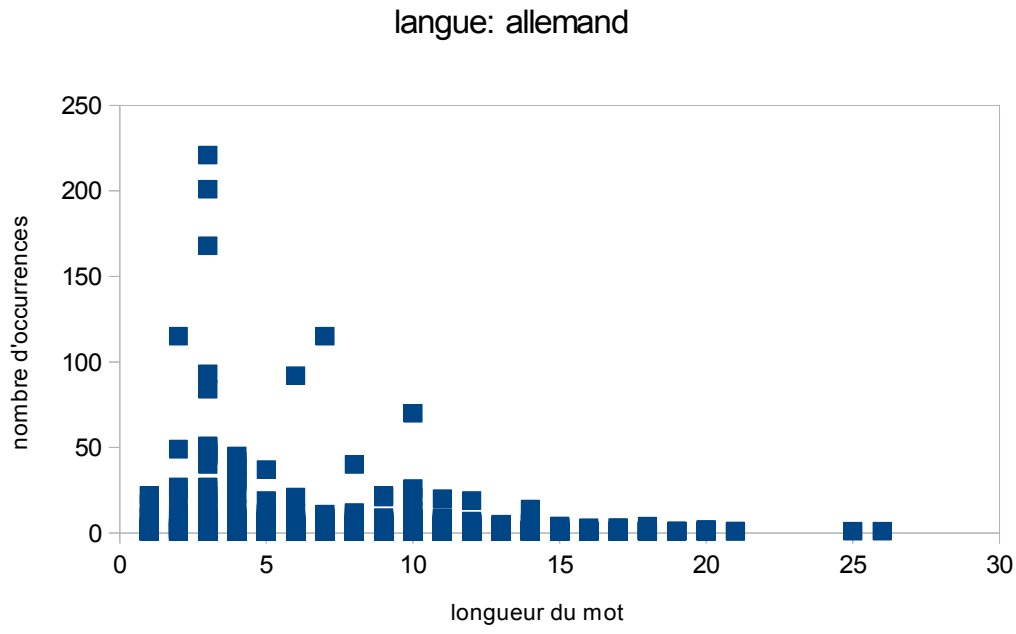
'\s+' pour les espaces blancs

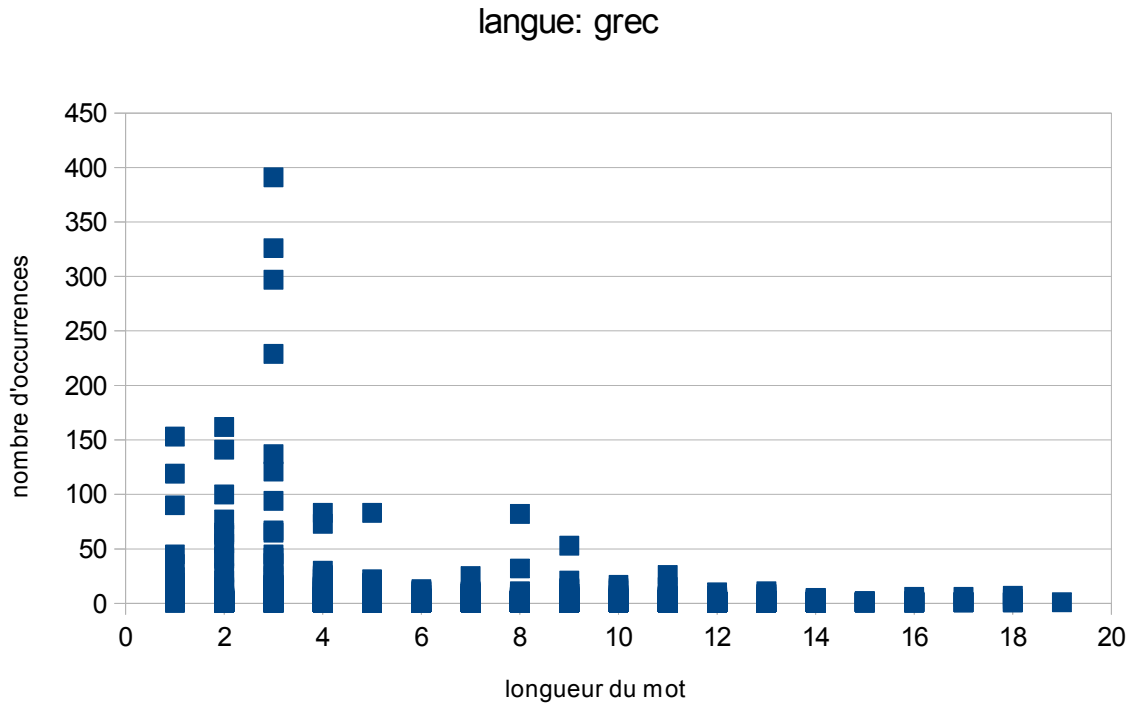
Fichiers en sortie

Le script d'analyse crée trois fichiers de sortie : $file.chars.csv, $file.words.csv et $file.word_groups.csv ; pour les caractères, mots et groupes de mots respectivement.

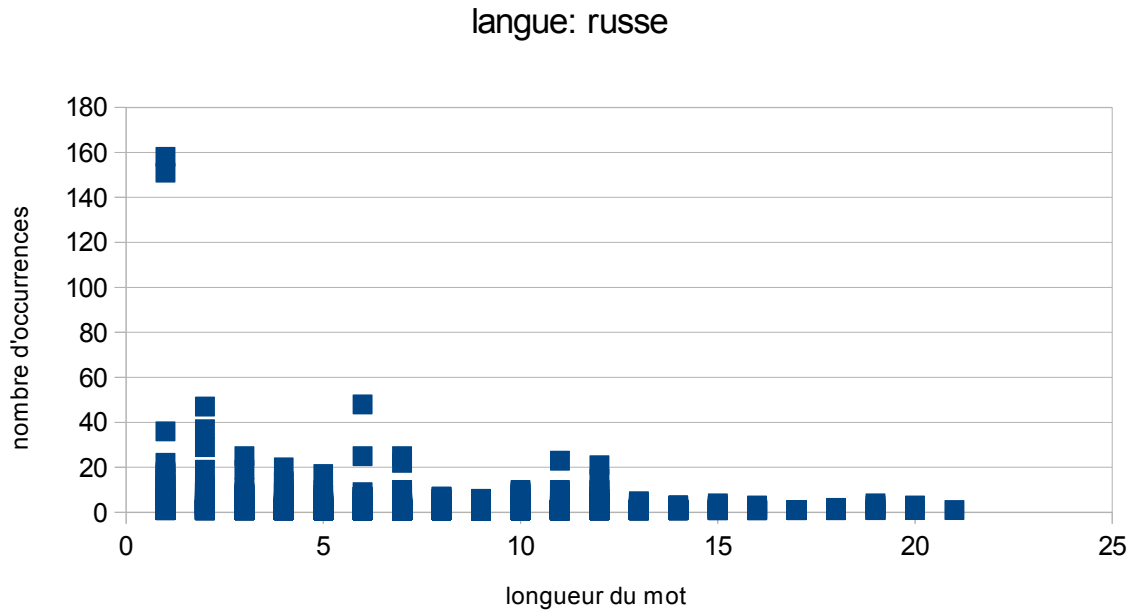
Relation entre longueur du mot et effectif

Pour la majorité des langues étudiées les mots courts sont bien les plus fréquents, rejoignant l'économie d'effort proposée par Zipf :

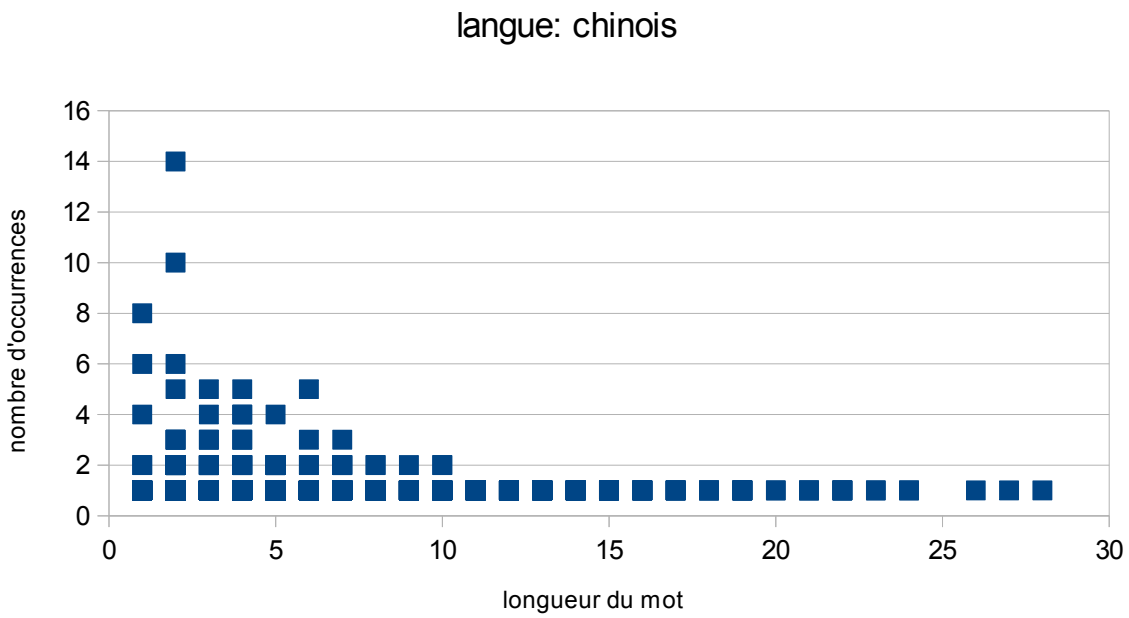




Le russe semble avoir un rapport plus homogène :

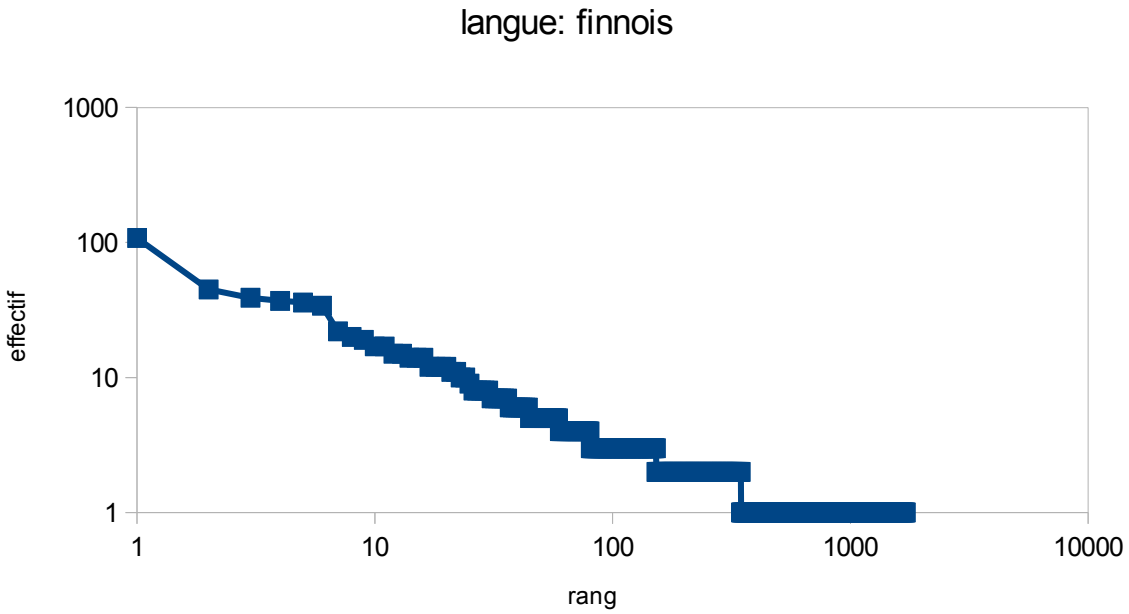


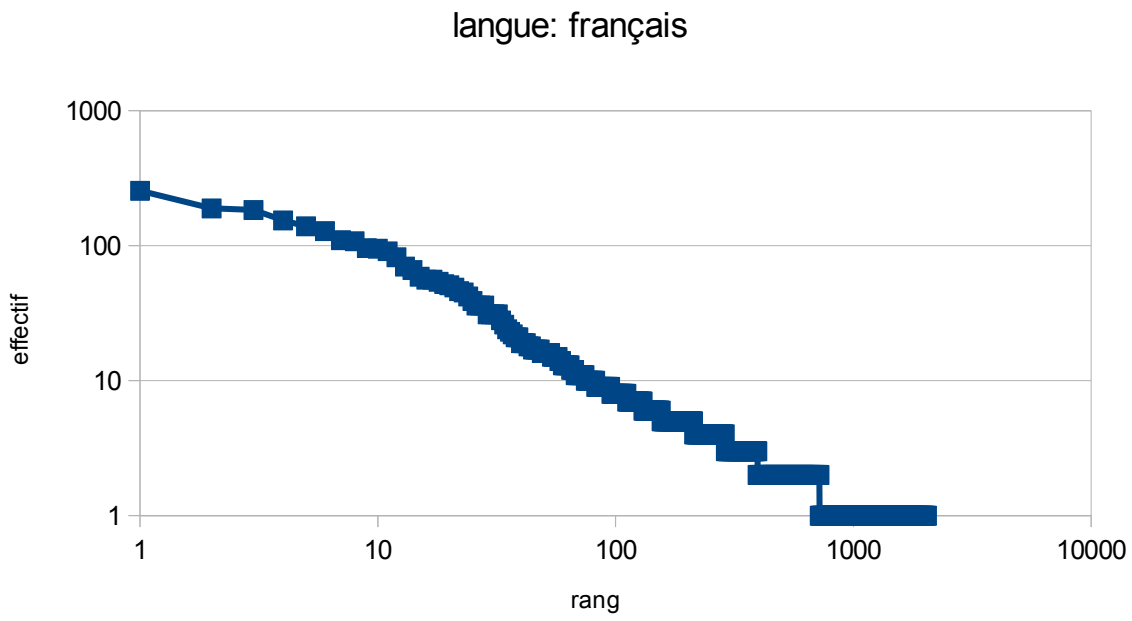
Le chinois est totalement différent, notamment à cause d'une séparation incorrecte des mots :



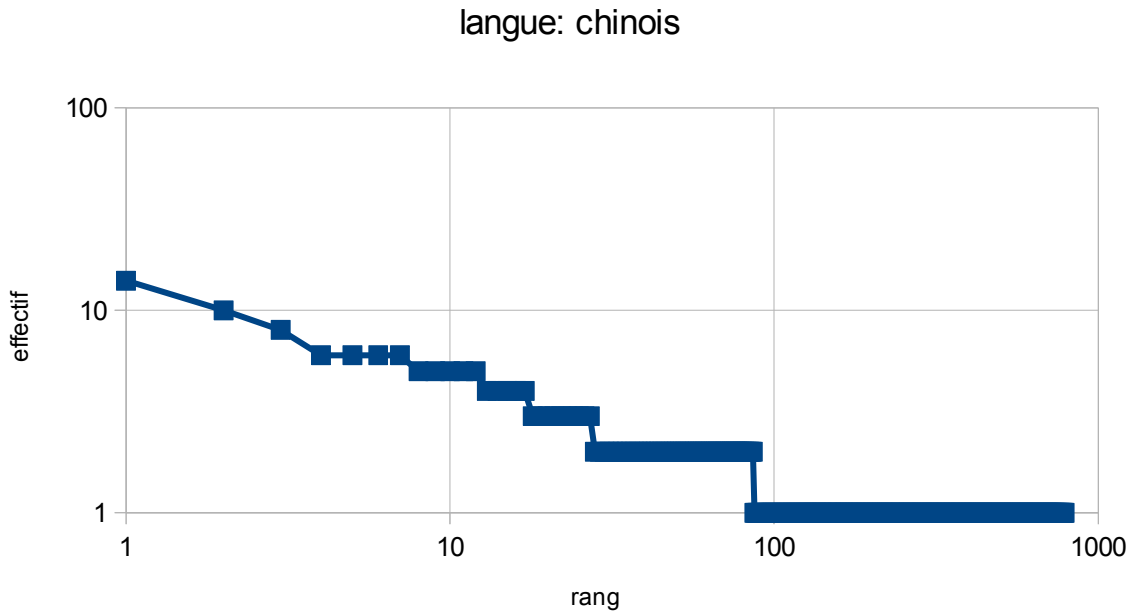
Loi de Zipf

La loi de Zipf est vérifiée pour les 6 langues à alphabets étudiées (de,el,en,fi,fr,ru), voici deux exemples :



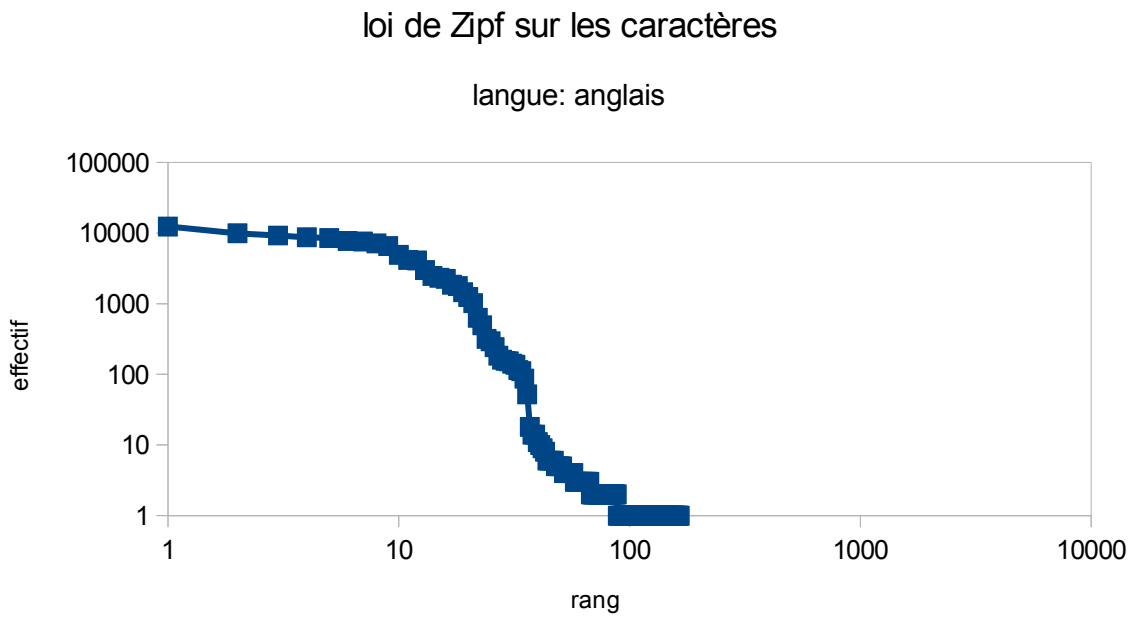


Le chinois ne vérifie pas la loi, anomalie notamment expliquée par une mauvaise séparation des mots :

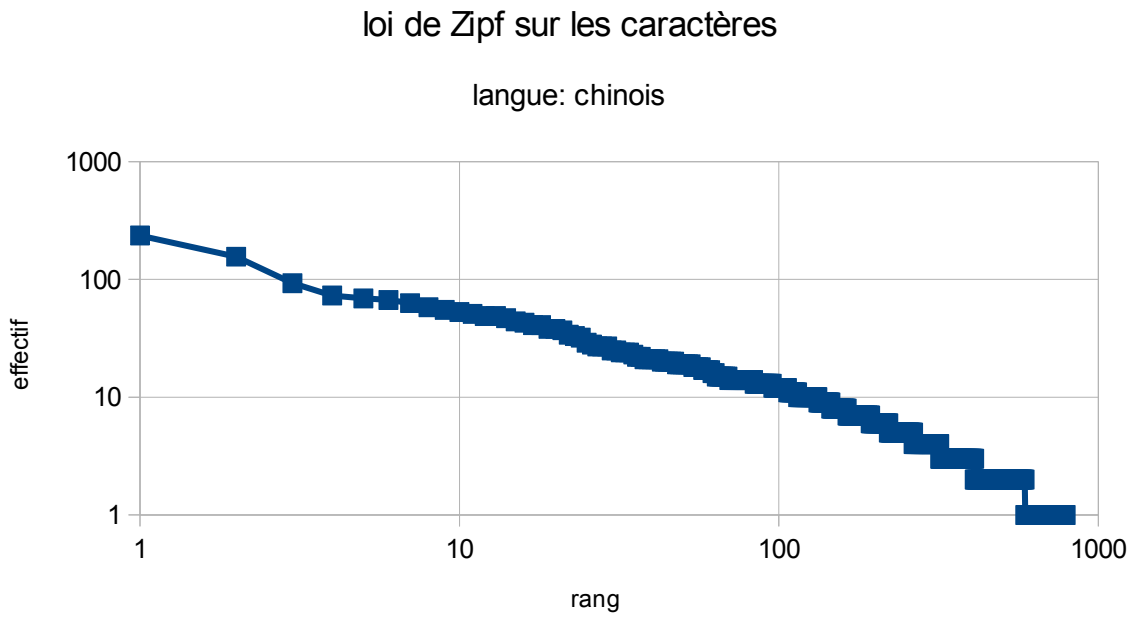


Loi de Zipf sur les caractères

La loi de Zipf ne semble pas s'appliquer aux caractères pour les langues alphabétiques (de,el,en,fi,fr,ru) :



Par contre elle semble fonctionner pour le chinois :



Rapport entre nombre d'occurrences et nombre de mots différents

Voici les rapports calculés pour les différentes langues :

de: 2.72634384718

el: 2.71118574012

en: 4.65810150775

fi: 1.64613596746

fr: 3.28459657702

ru: 1.87420915519

zh: 1.20580808081

Ce rapport nous indique le niveau de richesse du vocabulaire du document (à part pour le chinois pour la même raison que précédemment). On devine que le rapport est plus ou moins lié à la langue, cependant il faudrait étudier plusieurs documents (peut-être aussi en variant les sources) pour savoir s'il est «constant» ou non.

Traitement des groupes de mots

On a calculé le rapport «occurrences/nombre de groupe de mots» (donc plus le rapport est petit plus il y a de répétitions) pour des groupes de 2 à 5 mots :

de: 9.29848484848

el: 8.10424710425

en: 5.11543843284

fi: 17.5962732919

fr: 5.10022779043

ru: 10.1124497992

zh: 18.3653846154

On retrouve des résultats similaires aux précédents, c'est à dire qu'il y a plus de répétitions dans les versions anglaise et française. La version finnoise est celle qui se répète le moins. Les résultats pour le chinois sont toujours affectés par le mauvais découpage des mots.