12/01/2018

GAO Shuai & SHE Yidian

Université grenoble alpes

**RAPPORT INFORMATIQUE**

***Sciences du langage***

Nos différents textes choisis qui constituent les corpus sont tous sur le thème de la laïcité. Comme demandé, nous possédons ainsi un corpus de type scientifique, un autre de type littéraire et un troisième de type grand public. Chaque corpus a presque la même longueur d’environ 10000 mots.

Pour chacun de ces trois types de corpus, on nous a demandé de comparer le nombre de phrases, de mots, de formes, de lemmes et de syllabes. D’abord, nous avons noté que, ce qui est le hasard, le nombre de phrases pour le corpus scientifique est pareil que celui pour le corpus littéraire : 269. Et pour le corpus grand public 460. Cela montre que notre corpus grand public est plus riche que les deux autres. Ainsi peut-on supposer qu’un corpus grand public, pour faciliter la lecture du public, est en générale plus fourni en phrases courtes.

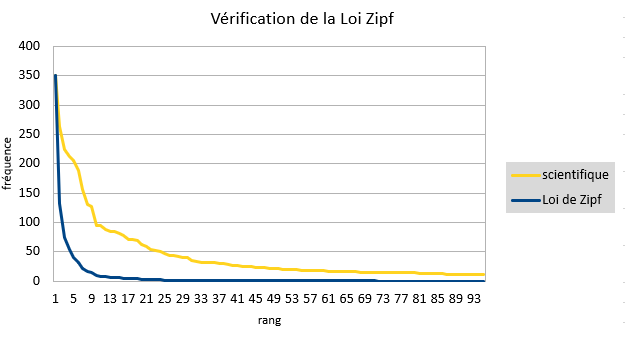
Concernant le nombre moyen de phrases par paragraphe, il y a 3,1 pour notre corpus scientifique, 5,5 pour le corpus littéraire et seulement 2,2 pour le corpus grand public. Cela nous permet de conclure que les textes littéraires ont tendance à s’écrire de façon plus étoffée dans chaque paragraphe.

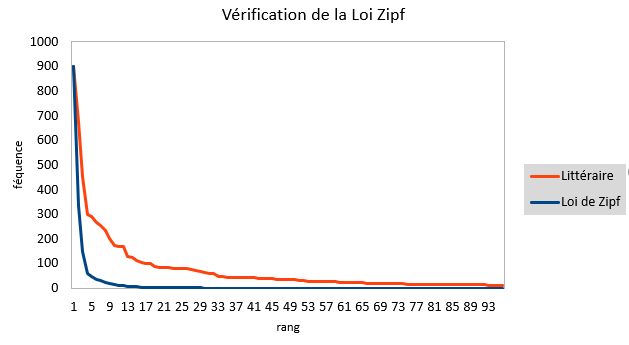
Quant à la longueur moyenne des phrases, nous constatons qu’il y a 39 formes dans le corpus scientifique, 41,8 au sein du corpus littéraire et 24,2 pour le corpus grand public, ce qui vient justement à justifier la supposition susmentionnée : les textes de type grand public se caractériseraient par un nombre moyen moins grand des formes par phrase. Cette observation démontre aussi que notre corpus scientifique et notre corpus littéraire sont à peu près égales à cet égard.

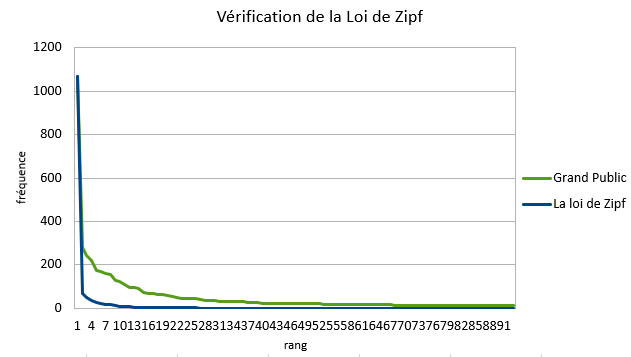
Ensuite, nous nous pencherons sur la longueur moyenne des mots, soit le nombre moyen des syllabes par forme. Et nous remarquons que cet indice s’avère quasiment pareil pour chacun de ces trois corpus : 1,6 pour le corpus scientifique, 1,4 pour le corpus littéraire et de nouveau 1,6 pour le corpus grand public.

Le nombre de lemmes nous informe la richesse lexicale du texte. En cela, nous en obtenons un classement : grand public (2337), scientifique (1967), littéraire (1449). Le corpus grand public est donc le plus riche lexicalement.

Pour ce qui est de la vérification de la Loi de Zipf, nous avons fait pour chaque corpus une figure qui donne des résultats globalement conformes. (Comme suit) Cependant, tous ces trois courbes faites de données se situent au-dessus de la courbe référentielle.







Après tout, nous nous intéresserons à la lisibilité de ces trois corpus. D’un côté, selon l’indice de Kandel & Moles, nous avons deux corpus qui sont normaux, le corpus littéraire et le corpus grand public, ayant respectivement 62,7 et 69,3 avec tous les deux 83% d’accès pour la population. Par ailleurs, le corpus scientifique a un indice de 53,2 avec seulement 54% d’accès populaire. D’un autre côté, d’après l’indice de Gunning Fog, notre corpus scientifique et notre corpus littéraire sont pourtant tous des textes difficiles avec respectivement un indice de 15,9 et de 16,8 tandis que le corpus grand public reste toujours un texte normal. Enfin, nous pouvons également faire un classement de lisibilité en synthétisant ces deux résultats : scientifique, littéraire, grand public, ce qui nous paraît raisonnable.