

Le Coran et la loi de Zipf

Chaque couche ajoutée au texte l'éloigne de la constante — mesure sur huit transmissions

Une mesure faite par AI-BED. Comptage réalisé le 23 octobre 2026 sur les éditions imprimées de huit transmissions du Coran (Complexe du Roi Fahd), avec un programme écrit pour l'occasion. Le site : ai-bed-coran.com. La méthode est donnée à la fin pour que chacun puisse la refaire.

1. La constante

Aucune langue n'a été conçue par un comité. Pourtant toutes obéissent aux mêmes régularités, qui sortent d'un simple comptage (G. K. Zipf, 1949 ; H. S. Heaps, 1978) : le rang d'un mot multiplié par sa fréquence donne à peu près une constante (pente 1) ; environ la moitié des mots n'apparaissent qu'une fois ; les mots les plus employés sont les plus courts ; le vocabulaire croît moins vite que le texte.

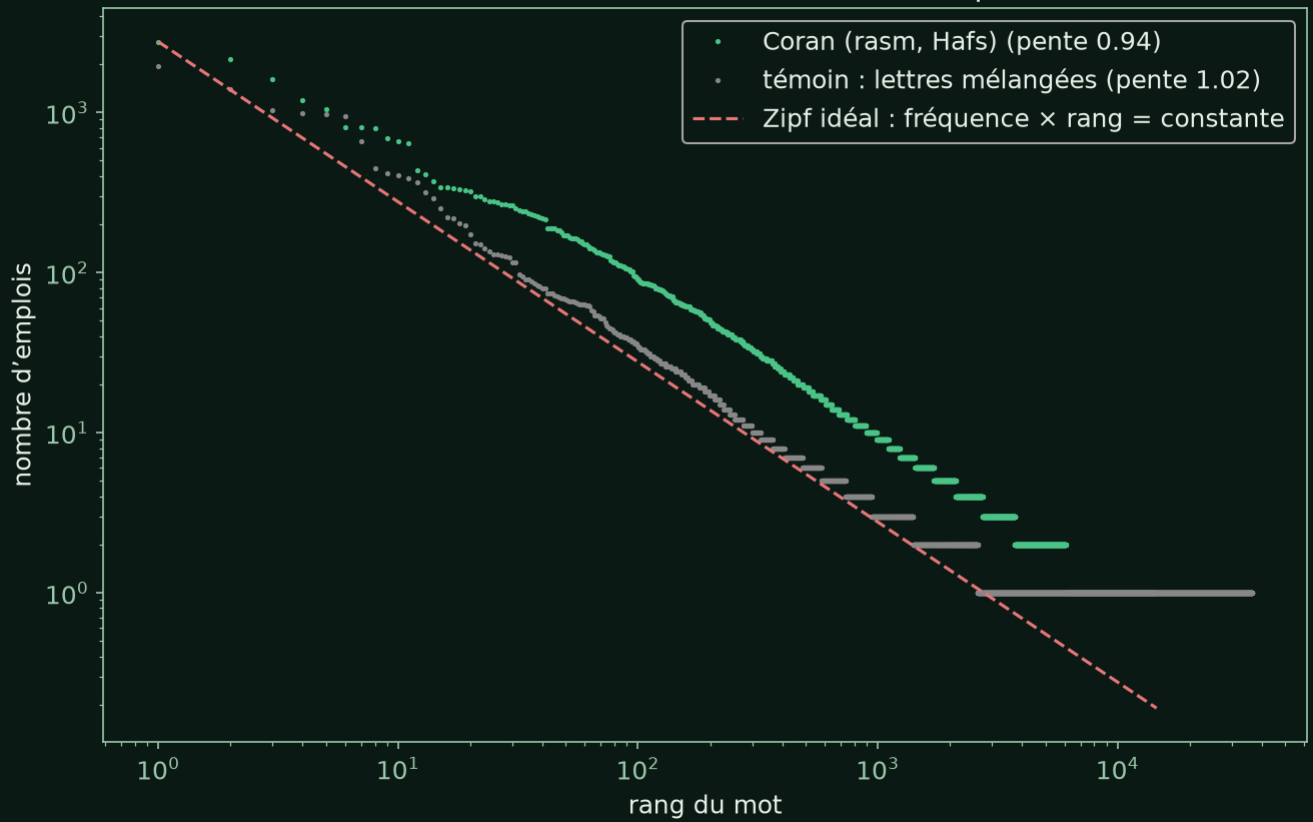
Un texte tapé au hasard imite la première loi (G. A. Miller, 1957). D'où un **témoin** : les mêmes lettres et les mêmes espaces que le Coran, mélangés au hasard. Une langue passe les quatre tests ; le hasard, non.

2. Le texte écrit (rasm de Hafs) passe les quatre tests

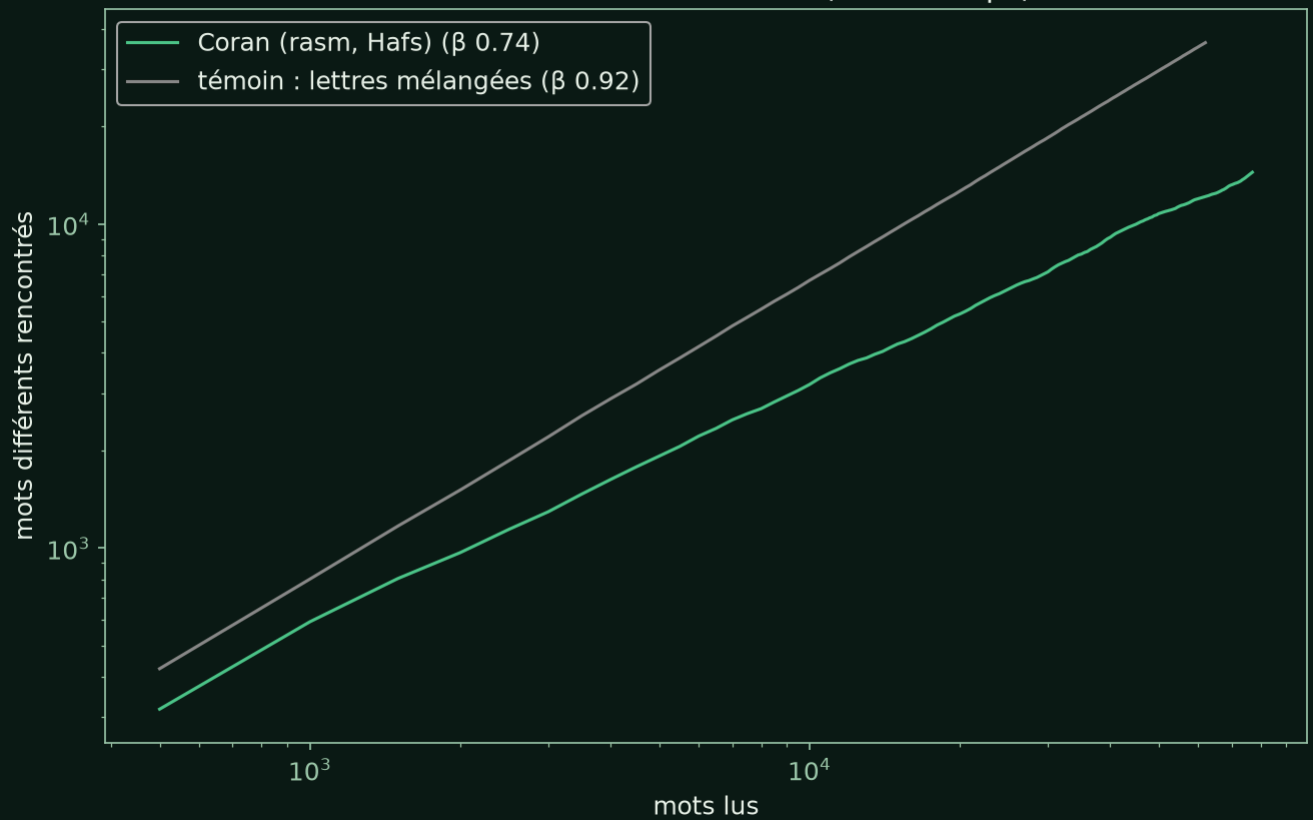
Loi	Coran (rasm)	Témoin mélangé
mots · mots différents	77 430 · 14 545	62 308 · 36 458
1. pente de Zipf (rangs 1-1000)	0,944	1,023
2. part des mots employés une fois	58,3 %	92,8 %
3. longueur : 100 mots les plus fréquents / hapax	3,4 / 5,5	1,7 / 8,0
4. exposant de croissance du vocabulaire	0,740	0,925

Le deuxième mot de tout le texte est الله : 2 155 emplois, le seul mot plein parmi les dix premiers.

Les mots du Coran suivent la loi de Zipf



La croissance du vocabulaire (loi de Heaps)



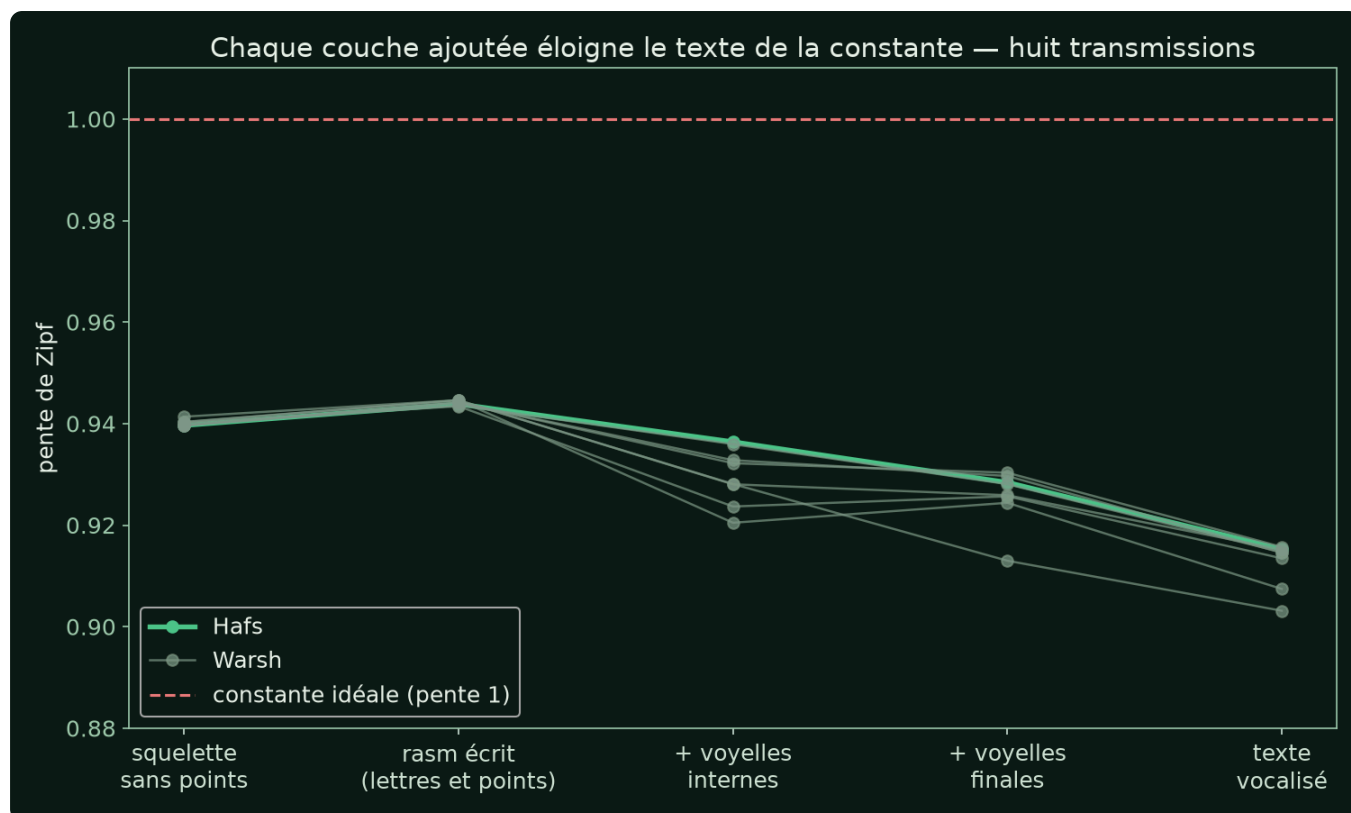
3. Les huit transmissions

Transmission	rasm : pente	rasm : hapax	rasm : Heaps	vocalisé : mots différents	vocalisé : pente
Hafs	0,944	58,3 %	0,740	19 179	0,904
Shou'ba	0,943	58,4 %	0,740	19 175	0,904
Warsh	0,944	58,2 %	0,740	20 888	0,868
Qâloûn	0,944	58,2 %	0,740	20 360	0,914
al-Bazzî	0,944	58,3 %	0,740	18 607	0,920
Qounboul	0,944	58,3 %	0,740	18 607	0,921
ad-Doûrî	0,943	58,2 %	0,740	20 001	0,903
as-Soûsî	0,944	58,2 %	0,741	20 023	0,913

Sur le rasm, les huit transmissions donnent les mêmes chiffres au millième près. Dès qu'on remet les voyelles, elles se séparent.

4. Couche par couche : la pente de Zipf

Transmission	squelette sans points	rasm (lettres et points)	+ voyelles internes	+ voyelles finales	texte vocalisé	témoin : voyelles au hasard
Hafs	0,940	0,944	0,936	0,928	0,915	0,697
Shou'ba	0,940	0,944	0,936	0,928	0,915	0,697
Warsh	0,940	0,945	0,928	0,913	0,903	0,686
Qâloûn	0,940	0,945	0,928	0,926	0,915	0,681
al-Bazzî	0,940	0,944	0,933	0,930	0,915	0,688
Qounboul	0,940	0,944	0,932	0,930	0,916	0,685
ad-Doûrî	0,940	0,943	0,924	0,926	0,913	0,681
as-Soûsî	0,941	0,945	0,920	0,924	0,907	0,684



Le sommet est le rasm écrit, avec ses points. Chaque couche de voyelles éloigne le texte de la constante, dans les huit transmissions sans exception, et les voyelles de fin de mot (la grammaire) l'en éloignent le plus.

Les points des lettres ne sont pas une couche comme les autres. Absents des inscriptions d'avant l'islam, présents dès 643, ils distinguent des lettres ; ils rapprochent le texte de la constante, là où les voyelles l'en écartent.

Limite de la mesure. Des voyelles tirées au hasard détruisent la loi : les vraies voyelles ne sont donc pas du bruit, elles suivent une lecture. Ce que la mesure établit, c'est que toute couche de voyelles écarte le texte de la régularité qu'il a seul.

5. La méthode, pour la refaire

Texte : éditions imprimées de Hafs, Shou'ba, Warsh, Qâloûn, al-Bazzî, Qounboul, ad-Doûrî, as-Soûsî ; 6 236 versets chacune. **Normalisation** : voyelles, signes d'arrêt et hamza retirés ; ؤ lu ٥ ; formes de l'alif unifiées ; un mot = une suite de lettres entre deux espaces (les morceaux de mot séparés par erreur dans les fichiers sources, comme l'alif du tanwîn, sont recollés). **Pente** : régression de $\log(\text{fréquence})$ sur $\log(\text{rang})$, rangs 1 à 1 000. **Heaps** : vocabulaire relevé tous les 500 mots. **Témoins** : lettres et espaces mélangés (graine fixe) ; voyelles tirées au hasard sur autant de lettres que dans le texte réel.

document (fichier CSV).