Ce que la vocalisation ajoute au Coran

La loi de Zipf, le tracé et la récitation : un contrôle sur quatre-vingt-six textes arabes vocalisés

Rami Bechini · chercheur indépendant · ORCID 0009-0003-5798-2156 · version 1.0 (15 novembre 2026)

Résumé. Dans le Coran, la pente de la loi de Zipf vaut 0,944 sur le tracé consonantique (*rasm*) et 0,915 sur le texte vocalisé, et la même baisse se retrouve dans huit transmissions. Je teste si cette baisse est propre au Coran ou si elle résulte mécaniquement de la vocalisation, en mesurant de la même façon quatre-vingt-six textes arabes classiques entièrement vocalisés, avec un critère fixé avant la mesure. Sur les mille mots les plus fréquents, la baisse du Coran dépasse celle de 85 textes sur 86 ; sur l'ensemble du vocabulaire, le résultat s'inverse. L'analyse des formes montre que l'excès du Coran ne vient ni de son tracé ni de sa grammaire, mais de **la récitation notée dans ses voyelles** : la *shadda* d'assimilation en début de mot, présente dans 4,81 % des mots du Coran et absente des autres textes, et les voyelles de liaison en fin de mot. Sans ces deux marques, le Coran vocalisé redevient un texte ordinaire (il ne dépasse plus que 57 textes sur 86). La vocalisation du Coran est donc aussi la transcription d'une manière de réciter, ce dont toute mesure statistique du texte doit tenir compte.

Abstract. In the Qur'an, the Zipf slope is lower on the vocalized text than on the consonantal skeleton, in eight transmissions. Measured identically on 86 fully vocalized classical Arabic texts, with a decision criterion fixed in advance, this drop exceeds that of 85 texts on the most frequent words, but the result reverses on the whole vocabulary. The excess is traced to recitation marks in the vocalization: word-initial assimilation (shadda), found in about 5% of Qur'anic words and absent elsewhere, and liaison vowels at word ends. Without them, the vocalized Qur'an behaves like ordinary vocalized Arabic.

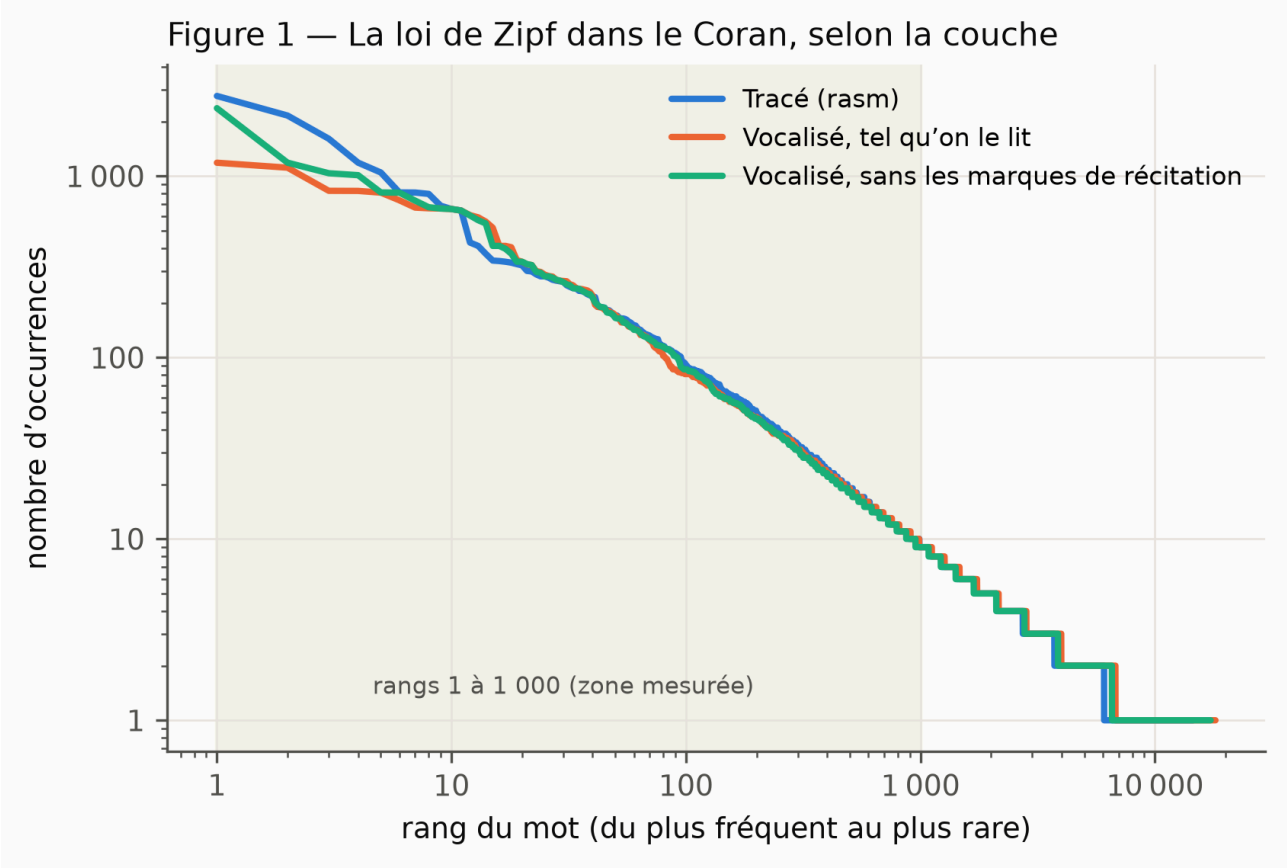
Mots-clés : loi de Zipf, Coran, rasm, vocalisation, récitation, tajwīd, linguistique quantitative.

1. La question

La loi de Zipf décrit une régularité de toutes les langues : la fréquence d'un mot est à peu près inversement proportionnelle à son rang, ce qui dessine, en coordonnées logarithmiques, une droite de pente voisine de 1 (Zipf 1949 ; Piantadosi 2014). Mesurée sur le Coran couche par couche, la pente est la plus proche de 1 sur le tracé, et chaque couche de voyelles l'en éloigne, dans les huit transmissions disponibles :

Transmission	Tracé	+ voyelles internes	+ voyelles finales	Texte vocalisé
Ḥafṣ	0,944	0,936	0,928	0,915
Shuʿba	0,944	0,936	0,928	0,915
Warsh	0,945	0,928	0,913	0,903
Qālūn	0,945	0,928	0,926	0,915
al-Bazzī	0,944	0,933	0,930	0,915
Qunbul	0,944	0,932	0,930	0,916
ad-Dūrī	0,943	0,924	0,926	0,913
as-Sūsī	0,945	0,920	0,924	0,907

Tableau 1. Pente de Zipf (rangs 1 à 1 000) du Coran, couche par couche, dans huit transmissions.



Les trois courbes ne se séparent qu'en tête de classement : la vocalisation éclate les mots les plus fréquents (la courbe orange s'écrase), et la retirer sans ses marques de récitation ramène la courbe (vert d'eau) vers celle du tracé (bleu).

L'objection est immédiate : la vocalisation distingue des homographes (عِلْم, la science, et عَلِمَ, il a su, partagent un tracé) et multiplie les formes distinctes, ce qui abaisse mécaniquement la pente dans n'importe quel texte arabe. La baisse dit-elle quelque chose du Coran, ou seulement de l'écriture arabe ?

2. Données et méthode

Le Coran : texte de Ḥafṣ et sept autres transmissions (éditions numériques publiques), et une édition en orthographe standard (Tanzil « Simple »). **Les textes de comparaison** : les 97 livres classiques vocalisés du corpus Tashkeela 0.3 (Zerrouki et Balla 2017), ses textes modernes, et neuf recueils de hadiths vocalisés (Open Hadith Data). Sont retenus les textes d'au moins 77 430 mots (la longueur du Coran) et dont au moins la moitié des lettres sont vocalisées : 86 textes classiques et hadiths.

Mesure. Chaque mot est normalisé (formes de l'alif, hamza, *tāʾ marbūṭa*) et pris en quatre états : le tracé, le tracé avec les voyelles internes, avec les voyelles finales, et complet. La pente est estimée par moindres carrés sur les rangs 1 à 1 000, comme dans la mesure d'origine. Chaque texte est mesuré sur cinq fenêtres continues de 77 430 mots tirées au hasard (graine fixée) ; on compare la baisse « tracé → vocalisé ». **Le critère, écrit avant chaque mesure** : un effet est dit propre au Coran si sa valeur dépasse celle d'au moins 95 % des textes. Les comparaisons sont strictes (quatre décimales).

3. Résultats

3.1 Sur les mots fréquents, le critère est rempli

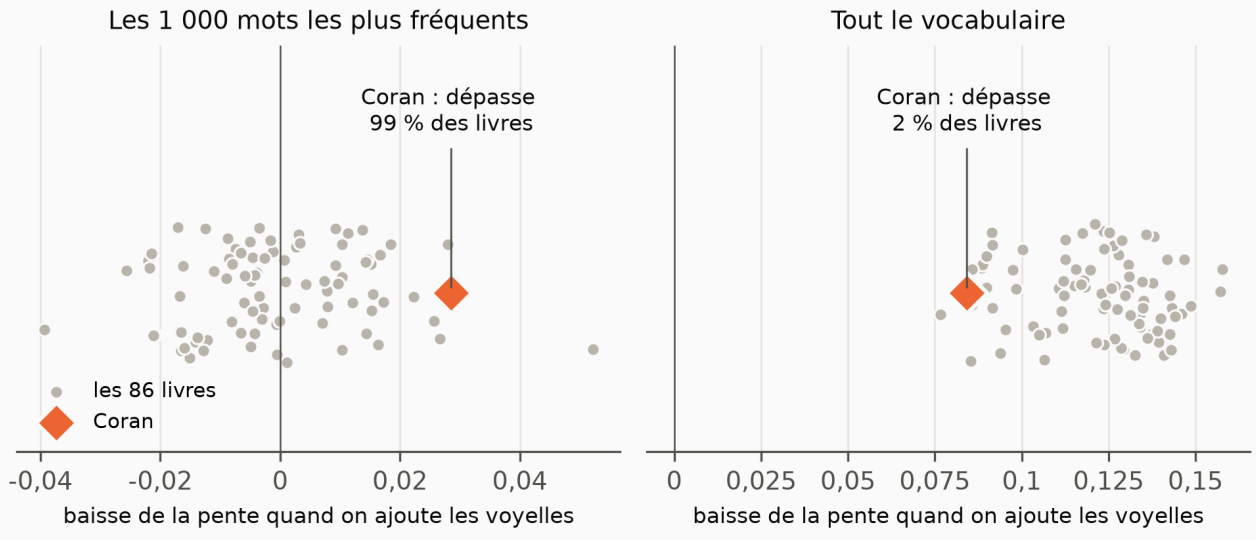
La baisse du Coran (0,0286) dépasse celle de 85 textes sur 86 (99 %) ; la médiane des textes est de -0,0029. Les textes les plus proches restent toutefois dans la marge de variation d'une fenêtre à l'autre.

3.2 Sur l'ensemble du vocabulaire, le résultat s'inverse

Mesure	Baisse du Coran	Médiane des textes	Le Coran dépasse
rangs 1-1000 (critère fixé d'avance)	0,0286	-0,0029	85/86 (99 %)
rangs 1-500	0,0144	-0,0228	82/86 (95 %)
rangs 10-1000	0,0157	-0,0032	77/86 (90 %)
rangs 1-100	0,0288	-0,0035	67/86 (78 %)
rangs 1-2000	0,0292	0,0279	45/86 (52 %)
droite sur tous les rangs	0,0752	0,1137	6/86 (7 %)
toute la distribution (vraisemblance)	0,0842	0,1239	2/86 (2 %)

Tableau 2. La baisse « tracé → vocalisé » selon la plage de rangs et la méthode d'estimation (vraisemblance : approximation discrète de Clauset et al. 2009).

Figure 2 — Le Coran parmi les 86 livres vocalisés : un résultat qui s'inverse



Chaque point est un livre (moyenne de cinq extraits de la longueur du Coran) ; le losange est le Coran.

Sur l'ensemble du vocabulaire, la vocalisation abaisse la pente de tous les textes, et le Coran est l'un de ceux qu'elle change le moins. Le profil est double : dans le Coran, la vocalisation agit sur les mots fréquents ; dans les autres textes, sur les mots rares.

3.3 L'hypothèse de la normalisation

Une hypothèse en découle : les couches ajoutées au tracé (voyelles, orthographe standard) feraient entrer le texte dans le moule de l'arabe ordinaire. Elle prédit qu'à chaque normalisation, le Coran se rapproche des autres textes. Le tableau 3 donne sa distance au texte moyen, en écarts types :

Mesure	ʿUthmānī, tracé	ʿUthmānī, vocalisé	Orthographe standard, tracé	Orthographe standard, vocalisé
mots fréquents (1-1000)	-0,30	-0,76	-0,17	-0,53
tous les rangs	-1,00	-0,56	-1,00	-0,42
toute la distribution	-1,00	-0,67	-1,01	-0,54

Tableau 3. Écart du Coran à la moyenne des textes (en écarts types), selon l'orthographe et la vocalisation.

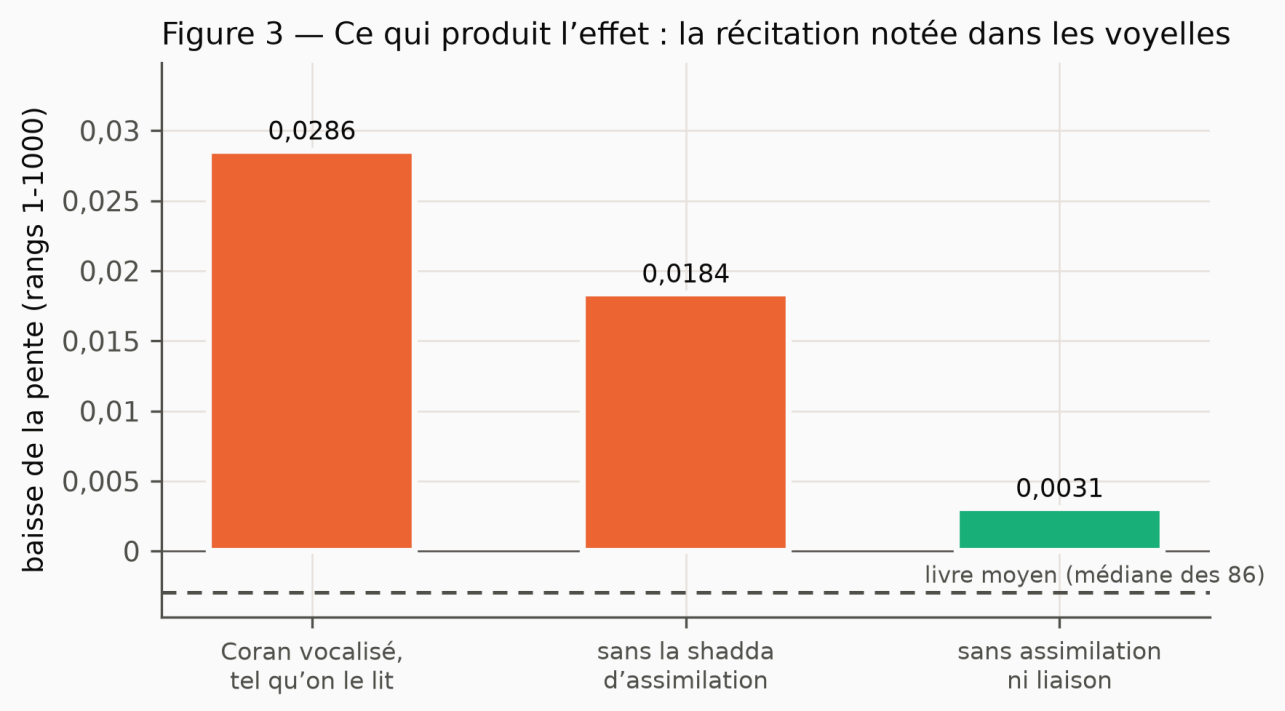
L'orthographe standard rapproche le Coran des autres textes dans quatre mesures sur six et ne l'en éloigne jamais ; les deux normalisations s'additionnent. Les effets restent modérés.

3.4 Le mécanisme : la récitation

Dans le Coran, 23 % des occurrences des mille formes les plus fréquentes quittent leur lecture principale une fois vocalisées, contre 17 % en médiane ailleurs. Les formes concernées sont des mots-outils : مِن، مَا، لَا، لَمْ، لَكُمْ. Beaucoup portent une *shadda* sur leur première lettre (مِّن، لَّكُمْ) : la notation de l'assimilation (*idghām*) du tanwīn ou du nūn du mot précédent. Cette marque touche 4,81 % des mots du Coran, et 0,00 % en médiane des autres textes (au plus 0,04 %).

Forme mesurée (pente, rangs 1-1000)	Baisse du Coran	Médiane des textes	Le Coran dépasse
Coran vocalisé tel quel	0,0286	-0,0029	85/86 (99 %)
sans la shadda initiale (assimilation)	0,0184	—	80/86 (93 %)
voyelles internes seules	-0,0023	-0,0063	68/86 (79 %)
voyelles finales seules	0,0155	0,0027	72/86 (84 %)
voyelles finales, liaison neutralisée (cas gardés)	0,0014	0,0032	41/86 (48 %)
voyelles finales en liaison seulement	0,0324	0,0097	85/86 (99 %)
vocalisé sans récitation (ni assimilation, ni liaison)	0,0031	-0,0032	57/86 (66 %)

Tableau 4. Décomposition de la baisse. Les mesures « voyelles internes / finales » et « liaison » sont faites sans la shadda initiale. Une fin de mot est « en liaison » si le mot suivant commence par un alif de liaison ; « liaison neutralisée » remet en liaison la voyelle finale que le mot porte d'ordinaire hors liaison, ce qui garde les cas grammaticaux.



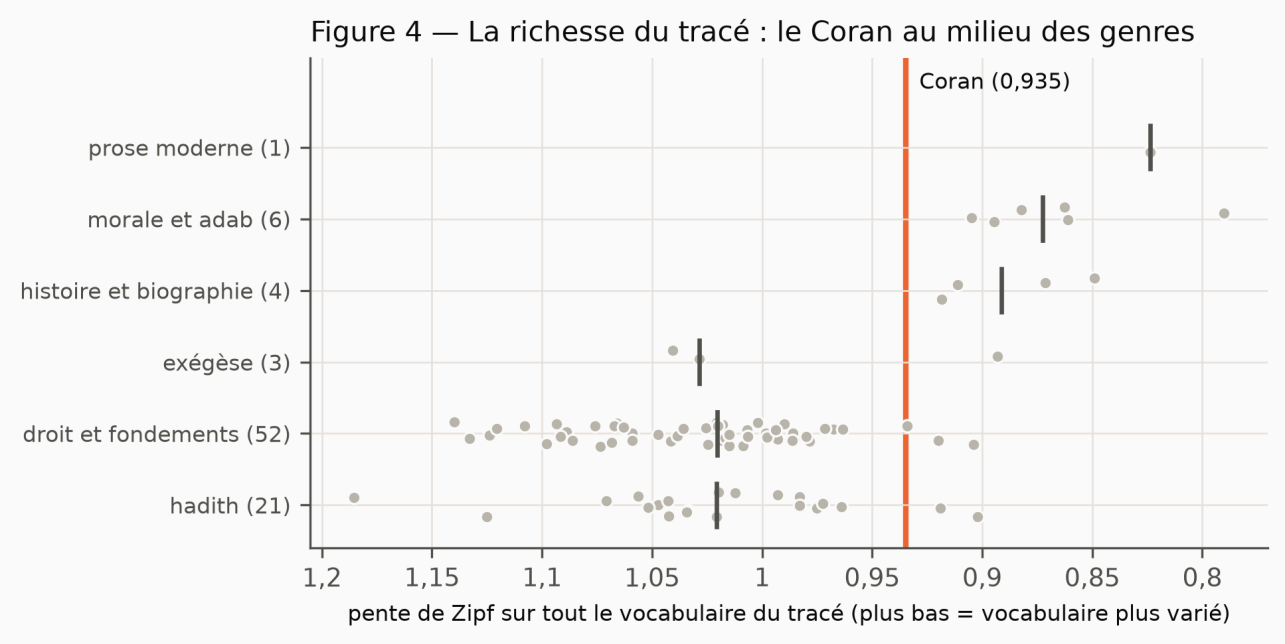
La baisse « tracé → vocalisé » du Coran, à mesure qu'on retire les marques de récitation ; la ligne pointillée est la médiane des 86 livres.

Les voyelles internes ne jouent aucun rôle : le tracé du Coran ne contient pas plus d'homographes séparés par la lecture que les autres textes. L'excès se loge en fin de mot, et plus précisément dans la **liaison** (*waṣl*) : elle modifie 5,98 % des fins de mot du Coran contre 3,67 % ailleurs, et sa neutralisation fait disparaître l'excès. Retirées ensemble, l'assimilation et la liaison ramènent la pente vocalisée du Coran à 0,9408, contre 0,9439 pour le tracé : il ne dépasse plus que 66 % des textes. La récitation explique environ neuf dixièmes de l'effet.

3.5 Le tracé lui-même est-il singulier ?

Mesure du tracé	Coran	droit et fondements (52)	hadith (21)	exégèse (3)	histoire et biographie (4)	morale et adab (6)	prose moderne (1)
formes distinctes	**14 565**	11 932	9 634	12 017	14 501	16 489	19 577
part des hapax	**0,583**	0,525	0,553	0,550	0,632	0,615	0,618
entropie (bits)	**11,00**	10,74	9,21	10,78	10,48	11,14	11,71
pente, tous les rangs	**0,935**	1,020	1,021	1,029	0,891	0,872	0,824

Tableau 5. Richesse du vocabulaire du tracé : le Coran et la médiane de chaque genre (nombre de textes entre parenthèses).



Chaque point est un livre ; les traits sombres sont les médianes par genre ; la ligne orange est le Coran.

Le tracé du Coran est plus varié que les traités de droit et les recueils de hadiths, mais ordinaire face aux livres d'histoire, de morale et à la prose moderne. Sa singularité apparente sur l'ensemble du vocabulaire tenait à la composition du corpus de comparaison.

4. Discussion

Le texte vocalisé du Coran n'est pas une simple version grammaticale du tracé : c'est aussi la transcription d'une **récitation**, avec ses assimilations et ses liaisons, que les autres textes arabes ne notent pas. C'est cette couche, et non le tracé ni la grammaire, qui distingue le Coran dans la mesure de Zipf. Le

tracé, lui, suit la loi commune des langues, avec une richesse lexicale qui le place entre les textes savants et les récits.

Deux conséquences pratiques en découlent. Toute étude statistique du Coran doit préciser la couche mesurée : tracé, texte vocalisé, ou texte vocalisé sans marques de récitation ne donnent pas les mêmes résultats. Et les couches ajoutées au tracé ne sont pas neutres : l'orthographe standard et la vocalisation rapprochent le texte de l'arabe ordinaire, et la récitation lui donne une forme qui n'appartient qu'à lui.

5. Limites

- Le corpus de comparaison est dominé par le droit et le hadith ; un seul texte moderne atteint la longueur du Coran, et la poésie vocalisée manque.
- Le repérage de la liaison repose sur l'alif de liaison noté (ٱ dans le Coran, alif nu ailleurs) ; il ignore les liaisons non marquées.
- La pente par moindres carrés dépend de la plage de rangs (tableau 2) ; les conclusions sur le mécanisme sont formulées sur la plage fixée d'avance.
- Les critères ont été fixés avant chaque mesure dans un carnet de recherche daté, mais non déposés publiquement.

6. Conclusion

La baisse de la pente de Zipf qu'entraîne la vocalisation du Coran n'est pas un effet mécanique de l'écriture arabe, ni une propriété de son tracé : elle vient de la récitation que sa vocalisation transcrit. Le tracé est le texte ; la vocalisation y ajoute une grammaire, commune à la langue, et une manière de réciter, propre au Coran.

Données et code

Le code (Python, NumPy) et les résultats bruts permettent de refaire chaque chiffre ; graines fixées. Données : corpus Tashkeela 0.3 (GPL-2.0), Open Hadith Data (ODbL), éditions numériques publiques du Coran.

Références

Clauset, A., Shalizi, C. R. et Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661-703.

Heaps, H. S. (1978). *Information Retrieval: Computational and Theoretical Aspects*. New York : Academic Press.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin and Review*, 21(5), 1112-1130.

Zerrouki, T. et Balla, A. (2017). Tashkeela: novel corpus of Arabic vocalized texts, data for auto-diacritization systems. *Data in Brief*, 11, 147-151.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge (Mass.) : Addison-Wesley.

Open Hadith Data (mhashim6), recueils de hadiths vocalisés, Open Database License.