Compression de Huffman

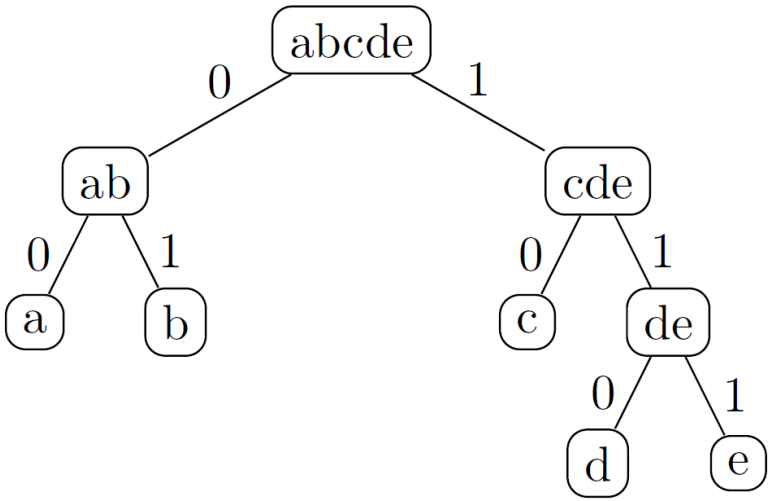
Benguezzou Mohamed
Fiszer Andrea

Licence 2 semestre 4

Table des matières

Analyse des facteurs influants

Quel facteur influence la structure de l'arbre de Huffman ?



- La fréquence d'apparition des caractères
- Le nombre de feuilles, i.e la taille de l'alphabet

Génération des fichiers de test

Comment générer des fichiers de test ?

- `--num_chars` : la taille de l'alphabet.
- `--size` : la taille du fichier généré.
- `--output_dir` : le répertoire de sortie.

Cela nous permet de générer des fichiers en controlant la **taille de l'alphabet** et la **taille du fichier**.

Lois de génération

Comment contrôler la fréquence d'apparition des caractères ?

- Loi uniforme
- Loi normale
- Loi de puissance ($P(x) = \frac{1}{x^{\alpha}}$; avec $\alpha = 1.5$)
- Loi aléatoire
- Loi linéaire

L'utilisation de ces lois nous permet de contrôler **la fréquence** d'apparition des caractères dans les fichiers générés.

Expérimentation - Lois de génération

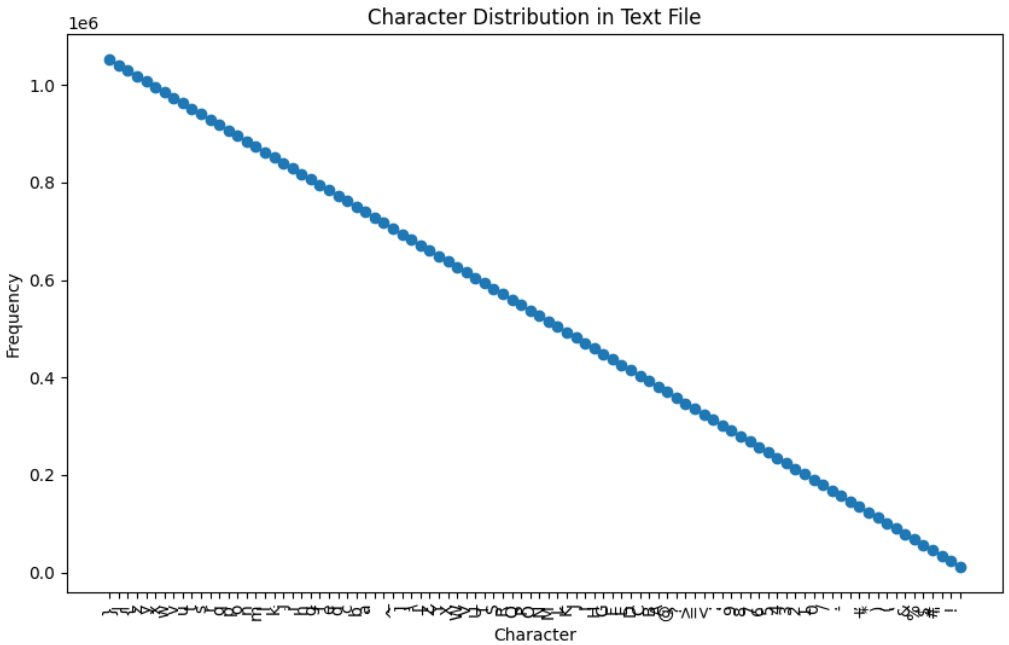


Figure: Loi linéaire

Hypothèse

- Faible différence des profondeurs des feuilles.
- Gain de compression pas très élevé.

Expérimentation - Lois de génération

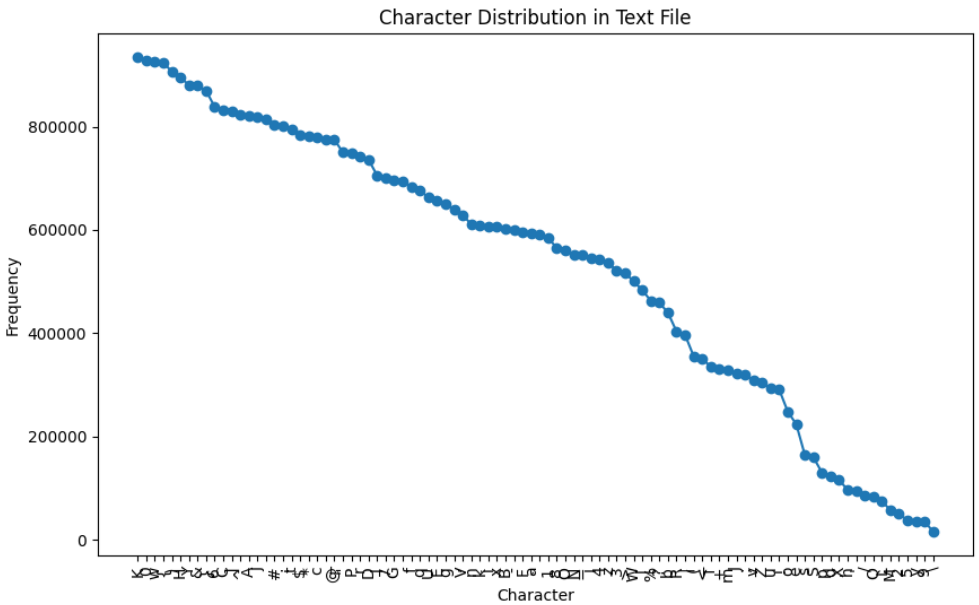


Figure: Loi aléatoire

Hypothèse

- Suit approximativement une loi linéaire.
- Gain de compression similaire à la loi linéaire.

Expérimentation - Lois de génération

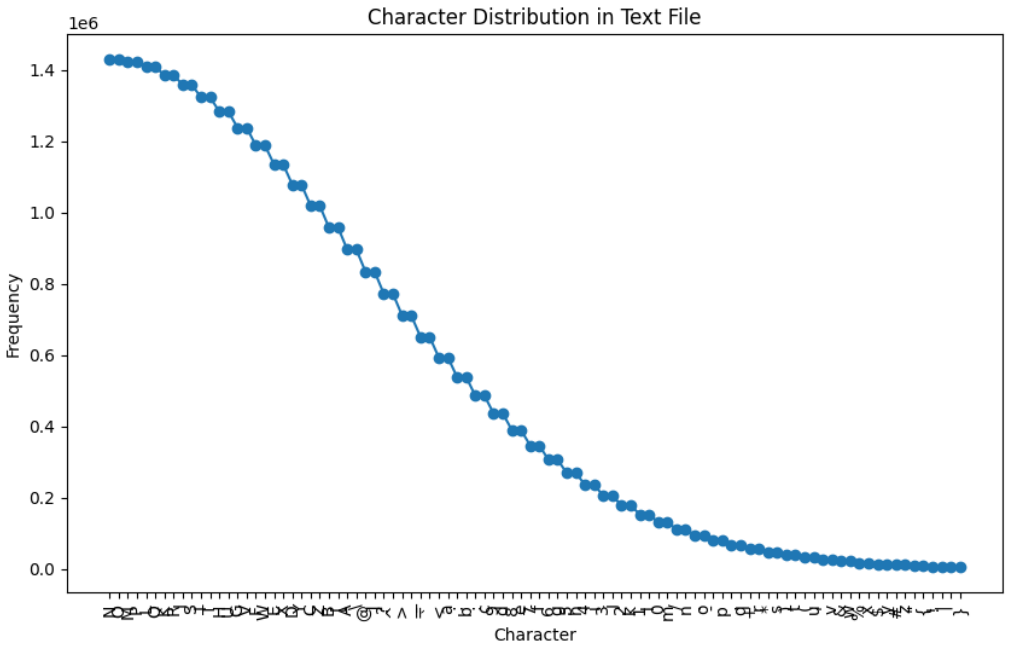


Figure: Loi normale

Hypothèse

- Meilleure compression que la loi linéaire.
- Profondeur des feuilles plus inégale.

Expérimentation - Lois de génération

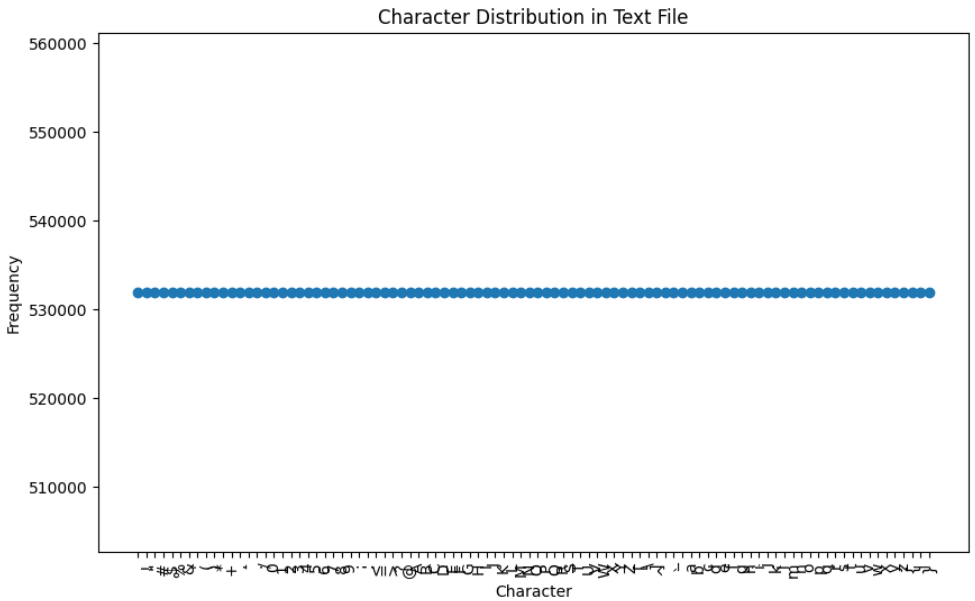


Figure: Loi uniforme

Hypothèse

- Fréquence d'apparition des caractères égale.
- Faible gain de compression par rapport aux autres lois.

Expérimentation - Lois de génération

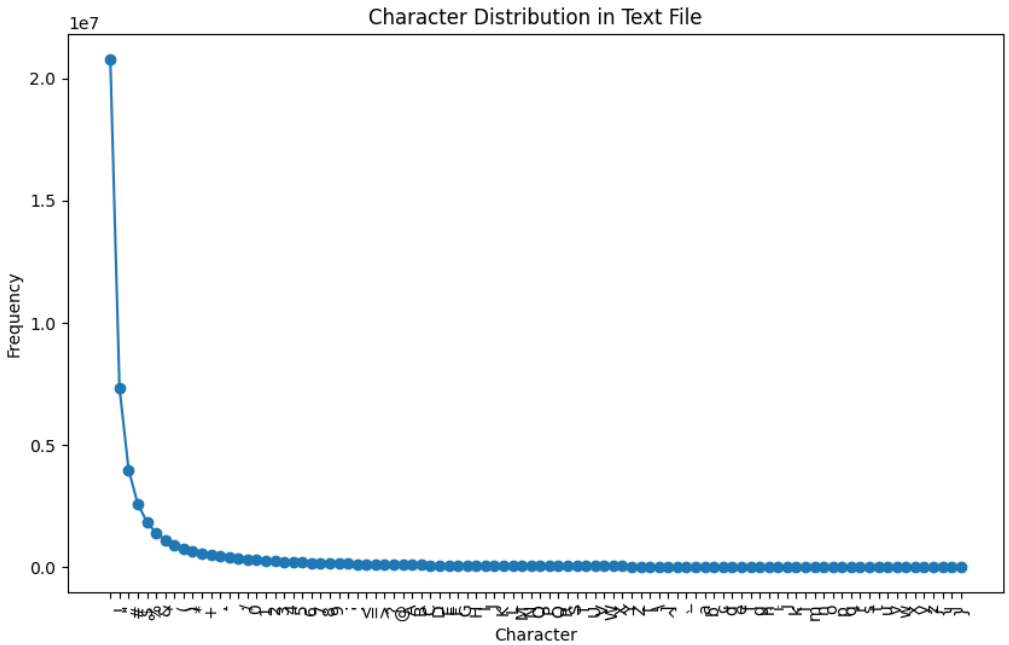


Figure: Loi de puissance

Hypothèse

- Meilleure résultat de compression attendu.
- Fréquence d'apparition des caractères très inégale.
- Gain de compression élevé.

Analyse des résultats

Quelques chiffres

- 80 fichiers générés
- 5 lois de génération
- 10 tailles de fichier (de 1000 octets à 50 000 000 octets)
- 7 tailles d'alphabet (de 2 à 94 puissance de 2)

Définitions

Quelques définitions

Taille originale : taille du fichier avant compression.

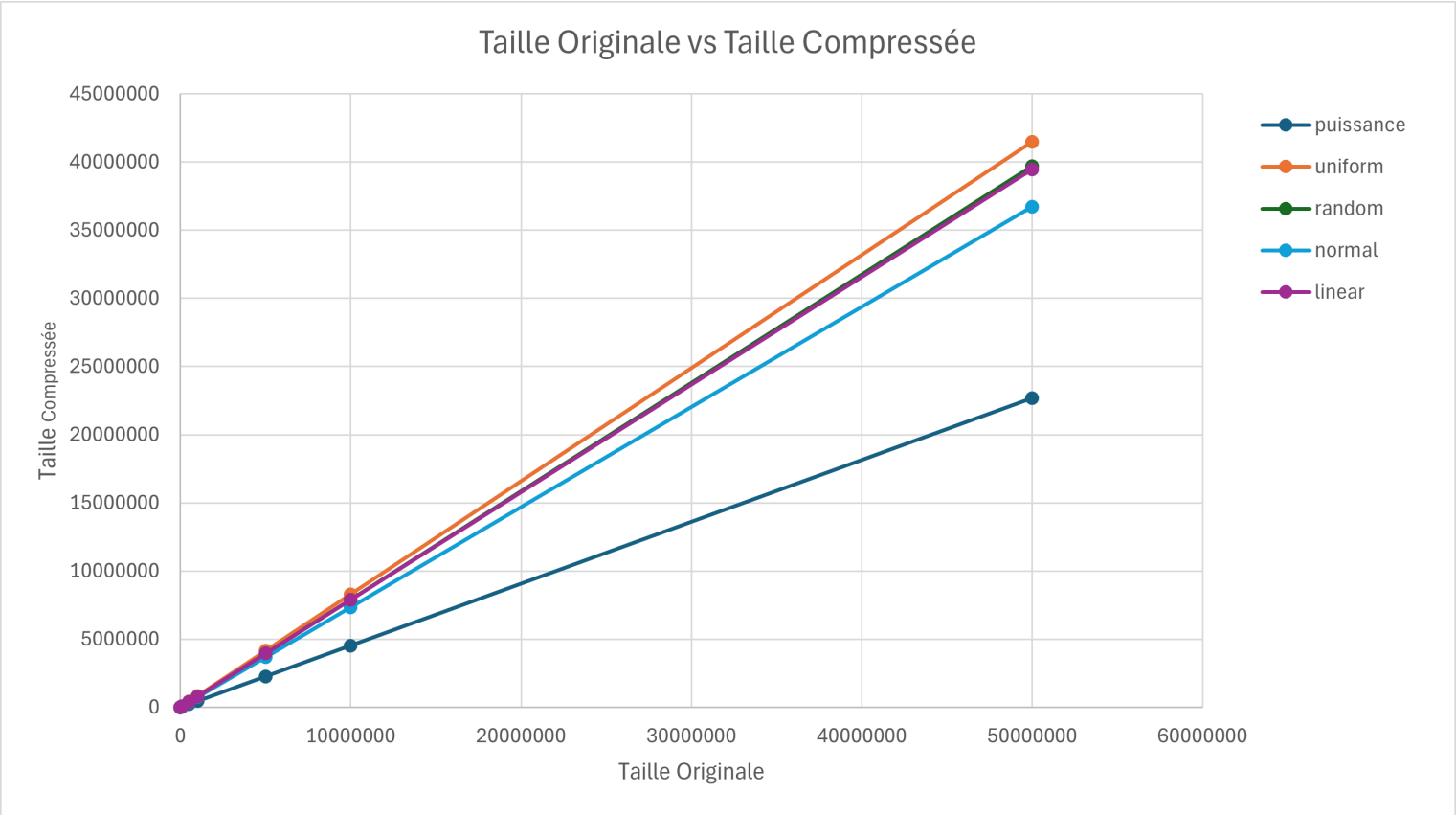
Taille compressée : taille du fichier après compression.

Taille de l'alphabet : nombre de caractères différents.

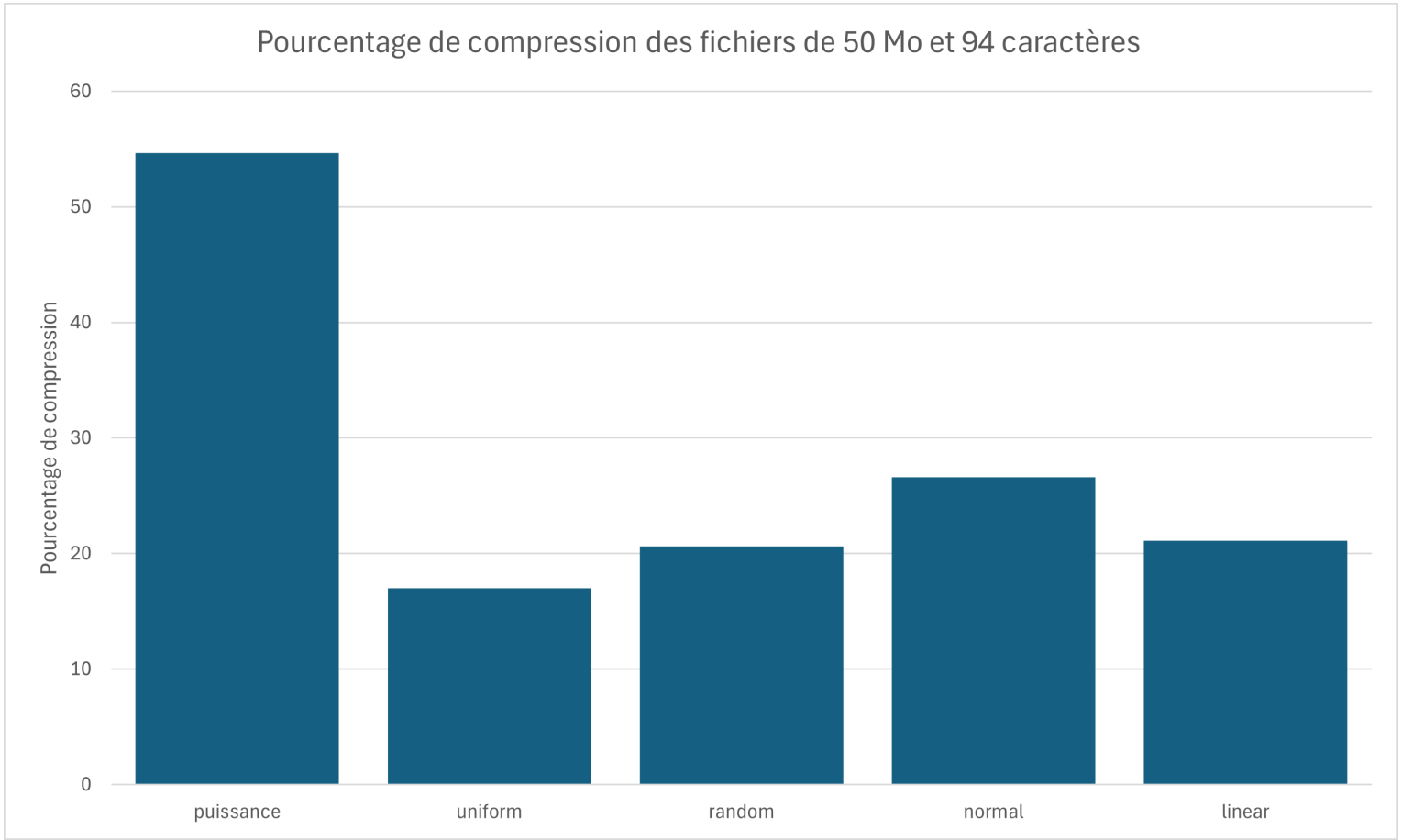
Taux de compression : $\frac{taille\_originale}{taille\_compressee}$

Pourcentage de compression :
$= \frac{taille\_originale - taille\_compressee}{taille\_originale} \times 100$

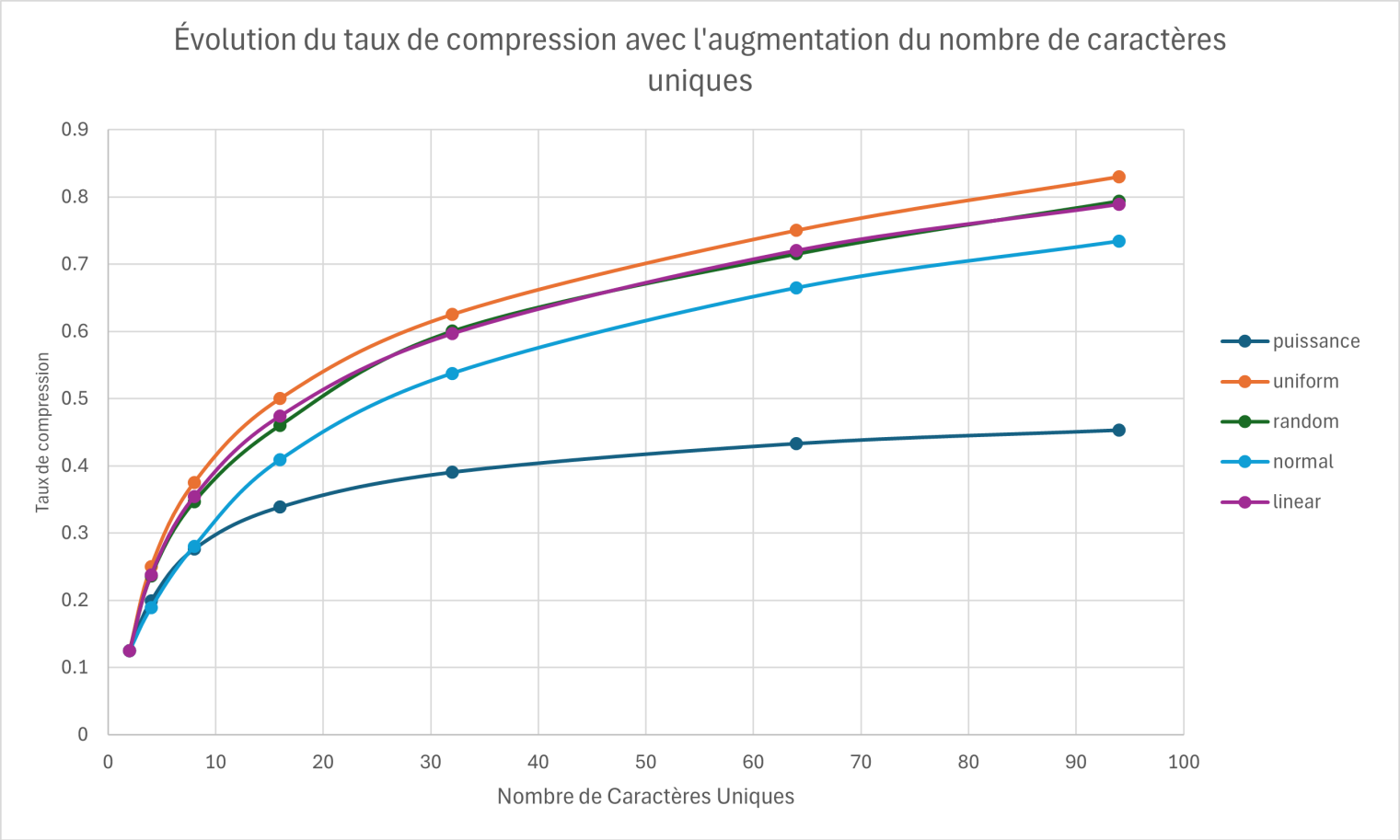
Taille des fichiers compressés selon la taille d'origine



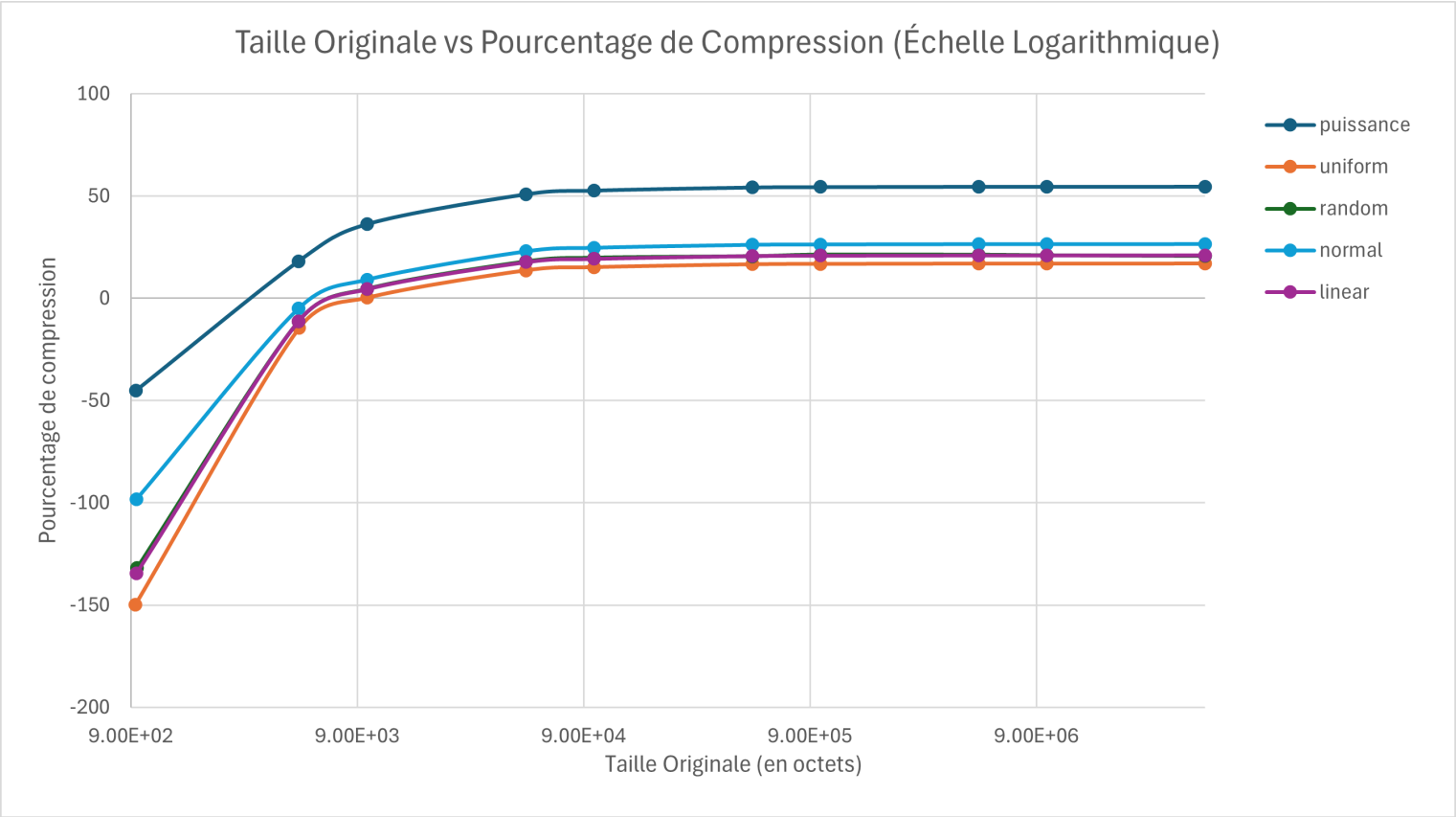
Pourcentage de compression selon la loi de génération



Evolution du taux de compression selon la taille de l'alphabet



Analyse et interprétation



Conclusion

- Une grande disparité des fréquences d'apparition des caractères favorise la compression
- La taille de l'alphabet influence le taux de compression
- Les taux de compression semblent converger vers une valeur limite pour des tailles de fichiers très grandes

Demonstration

- Demonstration