Une étude statistique élémentaire de la distribution des caractères et des mots dans *Salammbô*

Pierre Nugues
Université de Lund

24 mars 2015

1 Introduction

Il est banal de le constater, partout dans le monde, la numérisation avance à une vitesse prodigieuse. La littérature et les études littéraires ne restent pas à l'écart de ce mouvement irresistible et des entreprises multiples conduisent, peu à peu, inéluctablement, à la mise à disposition sous forme informatique de tout ce que l'esprit humain a pu produire. La matière intellectuelle cristallisée dans les textes est désormais accessible à des machines qui peuvent en extraire la substance, l'analyser, l'utiliser... Dans le sillage de la société numérique, nous assistons à la naissance de ce que l'on nomme parfois les « humanités numériques ».

Le but de notre article est de présenter quelques analyses statistiques élémentaires portant sur les caractères et les mots d'un texte numérisé, ne serait-ce que pour en contrôler la qualité. À l'origine de tout texte écrit, on trouve, en effet, un code alphabétique et nous décrivons ici comment extraire les symboles de ce code, calculer leur distribution statistique, analyser leur dispersion à l'aide de l'entropie et enfin, appliquer cette entropie à la mesure de la distance entre deux textes. Nous complétons cette présentation par l'exposé d'une méthode pour identifier les associations de mots les plus fréquentes dans un texte.

Notre étude a nécessité l'écriture de trois petits programmes dans le langage Python que l'on pourra, nous l'espérons, réemployer à d'autres œuvres.

2 Le corpus

Cette étude statistique porte sur trois textes principaux. Outre *Salammbô* de 1862, le corpus que nous avons constitué comprend *Notre-Dame de Paris* de Victor Hugo, dans l'édition de 1832, et *Le Conte de deux cités* « A Tale of Two Cities » de Charles Dickens dans sa version anglaise d'origine de 1859. Au-delà de leurs différences, ces textes partagent des traits essentiels : leurs trois auteurs

ont eu une influence immense sur la littérature mondiale ; ces romans comptent parmi leurs œuvres les plus connues et les plus lues et tous les trois portent la marque du XIX^{e} siècle.

Pour effectuer nos calculs, nous nous sommes servis de versions informatiques de ces textes qui proviennent toutes de wikisource.org. Les textes sont des retranscriptions automatiques de livres imprimés, suivis d'une relecture et d'une correction humaines. Cependant, malgré ces corrections, les versions informatiques ne sont pas exemptes d'erreurs.

En annexe, nous fournissons les programmes que nous avons écrits, ainsi que les textes eux-mêmes pour que le lecteur soucieux d'exactitude puisse reproduire les calculs, éventuellement dans des versions plus récentes ou corrigées (ou avec d'autres textes).

3 Lettres et symboles

3.1 Inventaire des caractères

Nous avons d'abord fait l'inventaire de tous les caractères et de tous les symboles qui composent les trois œuvres à l'aide d'un premier programme ; le tableau 1 en donne la liste, qui compte en plus l'espace et le saut de ligne. Une première constatation est que le répertoire de caractères est plus important dans les textes français à cause des lettres accentuées que dans celui en anglais qui utilise 75 caractères différents ; une autre est que *Notre-Dame de Paris* emploie plus de caractères que les deux autres du fait, notamment, de lettres grecques : 133 caractères contre 87 pour *Salammbô*.

Le choix du mode de codage des caractères est essentiel dans la constitution du corpus ; la norme Unicode qui vise à répertorier l'ensemble des symboles écrits de toutes les langues s'est imposée au monde informatique pour représenter les textes et documents. Elle divise les symboles en blocs correspondant aux grandes familles d'alphabets : latin, grec, cyrillique, arabe, chinois, etc. et permet de les faire cohabiter dans un même texte comme c'est la cas dans *Notre-Dame de Paris*.

Chaque lettre a son numéro Unicode propre ; même si elles sont d'apparence semblables mais de deux alphabets différents, comme le N latin et le N grec (*nu*). Ceci permet à un programme informatique de les distinguer, d'appliquer des transformations automatiques, comme le passage de majuscule en minuscule, ou de les classer dans l'ordre alphabétique. La mise en minuscule d'un N latin ou grec, par exemple, produira respectivement n et ν. Pour obtenir le tableau 1, nous avons dû corriger les premières versions des textes de wikisource qui contenaient des erreurs de codage et utilisaient indifféremment le N latin pour le français et le grec et une lettre cyrillique pour le Γ grec.

Tableau 1 – Les caractères employés par les trois ouvrages

Salammbô, 87 caractères	
Minuscules	a à â æ b c ç d e é è ê ë f g h i î ï j k l m n o ô œ p q r s t u ù û v w x y z
Majuscules	A À Æ B C Ç D E É F G H I J K L M N O Ô Œ P Q R S T U V X Y Z
Ponctuation	- , ; : ! ? . ’ « » ( ) – …
Notre-Dame de Paris, 133 caractères	
Minuscules	a à á â æ b c ç d e é è ê ë f g h i î ï j k l m n ñ o ô œ p q r s t u ù û v w x y z ß
Majuscules	A À Æ B C Ç D E É È Ê F G H I Î J K L M N O Ô Œ P Q R S T U V W X Y Z
Grec	Ἀ Ά Γ Η Κ Ν Ο ϊ α β γ δ ε έ ζ ι ί κ λ μ ν ο ό ς τ φ χ
Chiffres	0 1 2 3 4 5 6 7 8 9
Ponctuation	- , ; : ! ? . ’ « » ( ) – … &
A Tale of Two Cities « Le Conte de deux cités », 75 caractères	
Minuscules	a b c d e é f g h i j k l m n o p q r s t u v w x y z
Majuscules	A B C D E F G H I J K L M N O P Q R S T U V W X Y
Chiffres	1 2 5 6 7 9
Ponctuation	- , ; : ! ? . ‘ ’ “ ” ( ) * —

3.2 Effectifs de caractères

Nous avons ensuite dénombré chacune des lettres et des symboles. Avant d'effectuer ce décompte, nous avons appliqué une fonction de normalisation qui met toutes les lettres en majuscules et réduit les suites d'espaces et de sauts de ligne contiguës à une seule espace.

Le tableau 2 donne la liste de toutes les fréquences absolues de toutes les lettres de *Salammbô*, sans distinction entre les majuscules et les minuscules. Ce roman emploie par ailleurs d'autres caractères : des signes de ponctuations, des guillemets et des espaces. Nous les avons rassemblés dans la catégorie *Autres* sans en donner le détail. L'exécution du programme fourni en annexe permet d'obtenir toutes ces fréquences, y compris celles ne figurant pas dans le tableau.

Il est intéressant de comparer les fréquences relatives, les pourcentages, des caractères des trois œuvres. Le tableau 3 donne ces fréquences pour les dix premières lettres de l'alphabet. Nous avons pris en compte tous les signes dans le calcul de ces pourcentages, y compris les espaces qui comptent pour 17 à 18 % du total dans les trois romans. Nous constatons que *Salammbô* et *Notre-Dame de Paris* ont des distributions à peu près semblables, alors que certaines lettres ont des fréquences très différentes dans *Le Conte de deux cités*, comme le G, le H et le J. Ceci est dû, bien sûr, à la langue, le français pour les deux premières

Tableau 2 – Les fréquences des lettres dans *Salammbô*.

Let.	Fréq.	Let.	Fréq.	Let.	Fréq.	Let.	Fréq.	Let.	Fréq.
A	42297	*I*	33557	*Q*	3950	*Y*	1228	*Ê*	894
B	5746	*J*	1218	*R*	33493	*Z*	411	*Ë*	6
C	14172	*K*	91	*S*	46662	*À*	1894	*Î*	276
D	18867	*L*	30893	*T*	34987	*Â*	603	*Ï*	67
E	70955	*M*	13053	*U*	29196	*Æ*	15	*Ô*	396
F	4974	*N*	32821	*V*	6910	*Ç*	452	*Ù*	179
G	5142	*O*	22570	*W*	1	*È*	2000	*Û*	213
H	5289	*P*	13116	*X*	2209	*É*	7732	*Œ*	120
Autres	126876								

et l'anglais pour la troisième.

Tableau 3 – Distribution fréquentielle de 10 lettres dans les trois œuvres ; tous les caractères sont pris en compte dans le calcul de ces pourcentages, y compris les espaces. S : *Salammbô*, ND : *Notre-Dame de Paris*, TTC : *Le Conte de deux cités*

	A	B	C	D	E	F	G	H	I	J
S	6,87	0,93	2,30	3,07	11,53	0,81	0,84	0,86	5,45	0,20
ND	6,16	0,80	2,47	2,96	11,78	0,87	0,88	0,84	5,73	0,38
TTC	6,27	1,08	1,76	3,66	9,70	1,75	1,61	5,11	5,30	0,08

4 L'entropie

4.1 Vue d'ensemble

Le tableau 3 nous montre que la distribution des lettres n'est pas uniforme ; elle peut varier de moins de 1 % pour la lettre *F* à plus de 10 % pour le *E*. L'entropie de Shannon (1948) est un moyen de caractériser cette dispersion au moyen d'une mesure unique.

Pour aborder cette mesure, imaginons d'abord un alphabet très simple constitué de deux caractères : *a* et *b*. On définit l'entropie H d'un texte X composé de ces deux caractères par la formule suivante :

$$H(X) = -P(a)\log_2 P(a) - P(b)\log_2 P(b),$$

où $P(a)$ et $P(b)$ sont les probabilités des caractères *a* et *b* dans le texte et $\log_2$, le logarithme en base 2. Pour estimer la probabilité d'un caractère, on

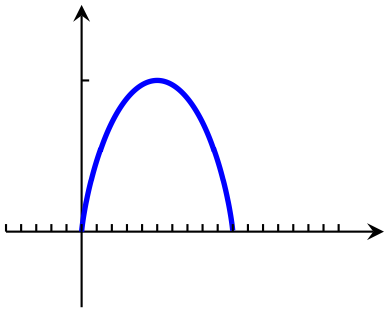

FIGURE 1 – La fonction entropie : $-x \log_2 x - (1-x) \log_2 (1-x)$ où x représente la proportion de a dans le texte variant de 0 à 1

prend simplement sa fréquence relative dans le texte. Ainsi, un texte de 10 000 caractères composé d'un seul a et de 9 999 b aura une entropie de

$$-\frac{1}{10000} \log_2 \frac{1}{10000} - \frac{9999}{10000} \log_2 \frac{9999}{10000} = 0,001473,$$

tandis qu'un texte composé de 5 000 a et de 5 000 b aura une entropie de

$$-\frac{5000}{10000} \log_2 \frac{5000}{10000} - \frac{5000}{10000} \log_2 \frac{5000}{10000} = 1.$$

La figure 1 représente la fonction entropie quand on fait varier la proportion de a dans le texte de 0 à 100 %. Si cette proportion est nulle, l'entropie est nulle ; s'il n'y a que des a, la proportion vaut 1 et l'entropie vaut aussi 0 ; enfin, s'il y a autant de a que de b, l'entropie est maximale et elle vaut 1.

Pour décrire l'entropie de façon plus concrète, on peut la relier à l'incertitude dans une expérience où on tirerait un caractère du texte au hasard. Le résultat du tirage est le plus incertain quand les deux caractères sont en nombres égaux ; il y a alors autant de chance d'avoir un a qu'un b et l'entropie est maximale ; elle est minimale lorsque qu'on n'a que des a ou que des b ; dans ce cas, on est sûr du caractère sur lequel on va « tomber ».

4.2 Entropie d'un texte

L'entropie de Shannon H se généralise à un nombre quelconque de caractères de la manière suivante :

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x),$$

où X est la suite de caractères d'un texte, y compris l'espace et les sauts de lignes ; x, un caractère appartenant à cette suite et $P(x)$, la probabilité du caractère x qu'on estime en calculant sa fréquence relative dans le texte ; par

exemple dans *Salammbô*, il y a 70 955 lettres E (majuscules et minuscules) et 615 531 caractères au total en éliminant les espaces doubles. Nous obtenons :

$$P(E) = \frac{70955}{615531} = 0,1153,$$

qui nous permet de calculer le terme :

$$-P(E)\log_2 P(E) = -\frac{70955}{615531}\log_2 \frac{70955}{615531} = 0,3593.$$

L'entropie est la somme de ces termes appliqués à toutes les lettres. Pour *Salammbô*, *Notre-Dame de Paris* et *Le Conte de deux cités*, les entropies sont respectivement de 4,40, 4,46 et 4,43 en respectant la distinction majuscule-minuscule.

Pour chacun de ces textes, l'entropie théorique maximale est fonction du nombre de leurs caractères et elle serait atteinte si tous les caractères étaient équiprobables ; un caractère pris au hasard pourrait alors avoir n'importe quelle valeur avec autant de chance ; l'entropie serait égale dans ce cas à $\log_2 N$ où N est le nombre de caractères. Pour *Salammbô* et ses 87 caractères différents, l'entropie maximale serait donc de $\log_2 87 = 6,44$.

Comme dans les trois textes les caractères ne sont pas équiprobables, l'entropie est beaucoup plus basse et elle reflète le fait que les chances d'un caractère pris au hasard sont inégales ; un *E* par exemple est beaucoup plus probable qu'un *W*. L'information du texte, en terme de symboles, est alors moins dense. L'entropie donne ainsi une idée de la diversité de leur distribution. Il est cependant délicat d'aller beaucoup plus loin dans l'interprétation de ces chiffres.

5 L'entropie relative

Si on ne peut conclure aisément du sens de l'entropie d'un texte isolé, il est possible de la modifier légèrement pour comparer deux distributions. L'examen rapide des trois histogrammes du tableau 3 nous donne à penser que *Salammbô* et *Notre-Dame de Paris* sont plus proches entre elles que *Le Conte de deux cités*, mais il ne nous donne pas une quantification de cette proximité. L'entropie relative ou divergence $D_{KL}(P||Q)$ de Q par rapport à P de Kullback et Leibler (1951), où P et Q sont deux distributions, fournit un modèle mathématique pour ceci.

Pour définir cette divergence, on introduit d'abord un terme, l'entropie croisée, de la façon suivante :

$$H(P,Q) = -\sum_{x\in X} P(x)\log_2 Q(x).$$

où il est facile de voir que $H(P,P)$ est l'entropie de Shannon. La divergence de Q par rapport à P correspond à la différence de ces deux entropies :

$$\begin{aligned} D_{KL}(P||Q) &= H(P,Q) - H(P,P), \\ &= \sum_{x\in X} P(x)\log_2 P(x) - \sum_{x\in X} P(x)\log_2 Q(x). \end{aligned}$$

Cette divergence, toujours positive, permet de quantifier la similitude de deux distributions de probabilités où P représente, de façon typique, une distribution à tester et Q, une distribution qui lui sert de modèle. On peut considérer la divergence comme une sorte de distance. Elle est asymétrique, cependant, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, et n'en suit donc pas la définition mathématique rigoureuse.

Nous avons calculé la divergence entre les distributions de caractères de *Salammbô* et celles des deux autres textes. Du fait de l'asymétrie entre les termes P et Q, nous l'avons prise à partir de *Salammbô* et à partir de l'œuvre considérée. Enfin, nous avons ajouté la nouvelle *Un Cœur simple* à notre corpus pour mesurer le divergence entre deux œuvres de Flaubert.

Le tableau 4 présente le détail des résultats que nous avons obtenus et qui sont conformes à ce que nous attendions. La divergence entre *Salammbô* et *Un Cœur simple* est très faible, quel que soit le sens du calcul ; elle est double entre *Salammbô* et *Notre-Dame de Paris* et beaucoup plus élevée avec *Le Conte de deux cités* en anglais.

Tableau 4 – Divergence entre les œuvres du corpus et *Salammbô*, $D_{KL}(P||Q)$. La divergence de Kullback-Leibler n'est pas symétrique et nous l'avons calculée à partir de *Salammbô* et à partir de l'œuvre considérée. UCS : *Un Cœur simple*, ND : *Notre-Dame de Paris*, TTC : *Le Conte de deux cités*

P	Q	Entropie de P	Entr. croisée	Diff.
Salammbô	*Salammbô*	4,4010	4,4010	0
Salammbô	UCS	4,4010	4,4092	0,0082
Salammbô	ND	4,4010	4,4187	0,0177
Salammbô	TTC	4,4010	4,5501	0,1490
UCS	*Salammbô*	4,4284	4,4349	0,0065
ND	*Salammbô*	4,4624	4,4747	0,0123
TTC	*Salammbô*	4,4294	4,8077	0,3783

6 Des lettres aux mots

6.1 Découpage d'un texte en mots

Après l'analyse statistique des caractères, on peut facilement passer à celle des mots. Pour ceci, on doit d'abord découper le texte. Cette étape, qui paraît simple à première vue, peut se révéler délicate dans certains cas : comment traiter l'apostrophe, par exemple ? Si on considère qu'elle sépare deux mots dans l'expression *l'avenue*, doit-on appliquer la même règle à *aujourd'hui* ?

Nous avons écrit un programme de découpage très simple qui se fonde sur l'analyse des caractères du tableau 1 et où nous avons posé la règle suivante :

toute espace et toute ponctuation délimitent un mot ; nous réduisons par ailleurs les suites contiguës de plusieurs de ces « délimiteurs » à un seul délimiteur.

Avec cette procédure de découpage et en mettant tous les mots en minuscules, nous obtenons un total de 105 129 mots pour *Salammbô* dont 11 144 mots différents, respectivement 183 244 et 16 519 mots pour *Notre-Dame de Paris* et 138 193 et 9 721 pour *Un Conte de deux villes*. Le tableau 5 donne la liste des 20 mots les plus fréquents des trois œuvres où nous voyons que *Salammbô* se distingue par l'emploi plus fréquent du pronom personnel réfléchi *se*, tandis que *Notre-Dame de Paris* utilise plus de conjonctions ou de pronoms *que* ainsi que de verbes *être*.

Tableau 5 – Fréquences absolues des 20 premiers mots. Pour *Salammbô*, les mots en gras sont ceux qui n'apparaissent pas dans les vingt mots les plus fréquents de *Notre-Dame de Paris* et réciproquement

Salammbô				*Notre-Dame de Paris*				*Le Conte de deux cités*			
1-10			11-20	1-10			11-20	1-10			11-20
les	4230	un	1407	de	8262	en	2210	the	8021	he	1854
de	3966	en	1321	la	5397	une	2047	and	4999	was	1774
des	3097	**se**	1239	et	4583	**que**	2029	of	4008	you	1427
et	2717	**s**	1100	le	4122	qui	1978	to	3570	with	1311
la	2662	dans	1097	à	3498	**du**	1594	a	2947	had	1305
le	1881	une	1060	l	3478	**est**	1592	in	2599	as	1163
il	1767	**ils**	988	il	2980	**qu**	1475	it	2066	her	1044
d	1724	**sur**	926	un	2603	des	1457	his	2011	at	1033
à	1692	**on**	755	les	2456	**était**	1426	i	1987	him	976
l	1527	qui	749	d	2222	dans	1388	that	1941	for	960

6.2 Associations de mots

Au-delà des mots isolés, on peut chercher à identifier les associations de mots propre à une œuvre ou à un auteur. L'information mutuelle $I(m_i, m_j)$ est une mesure mathématique de ces associations que l'on définit par la formule suivante (Fano, 1961) :

$$I(m_i, m_j) = \log_2 \frac{P(m_i, m_j)}{P(m_i)P(m_j)},$$

où m_i et m_j sont deux mots contigus dans un texte. L'information mutuelle de deux mots sera positive s'ils apparaissent plus souvent ensemble que séparément, nulle s'ils sont indépendants et négative s'ils sont plus souvent séparés qu'ensemble.

Pour mesurer l'information mutuelle, on doit d'abord identifier toutes les suites de deux mots adjacents, les bigrammes, et dénombrer chacun de ces bigrammes dans chacun des textes. On procède de même avec les mots. On calcule ensuite l'information mutuelle de chaque bigramme trouvé par une simple division des fréquences, ainsi par exemple, le bigramme *les barbares* apparaît 157 fois dans *Salammbô* ; sa fréquence relative est de $157/105128 = 0{,}00149$; la fréquence relative de *les* est de $4230/105129 = 0{,}04024$; *barbares* apparaît 237 fois ; sa fréquence relative est de $237/105129 = 0{,}00225$. L'information mutuelle du bigramme *les barbares* est donc :

$$I(\text{les}, \text{barbares}) = \log_2\left(\frac{157}{4230 \cdot 237} \cdot \frac{105129^2}{105128}\right) = 4,0412.$$

Une fois l'information mutuelle calculée, on classe les bigrammes par valeurs croissantes pour déterminer les associations les plus étroites. Les valeurs maximales correspondent à des mots et des bigrammes qui n'apparaissent qu'une seule fois dans un texte, comme l'expression *vagabondage perpétuel* où *vagabondage*, ainsi que *perpétuel* sont des emplois uniques dans *Salammbô*. Il est difficile de tirer des conclusions pour de telles expressions et nous avons posé un seuil de fréquence de 25. Le tableau 6 présente les 20 associations les plus fortes dans *Salammbô* et *Notre-Dame de Paris* pour des expressions apparaissant au moins 25 fois.

Sans prétendre à une analyse littéraire, deux traits propres à chacune des œuvres ressortent de ces chiffres : la prédominance des lieux, de l'espace et de sa description dans *Salammbô* avec les expressions *au milieu*, *au bord*, *au fond*, *par dessus*, *l'acropole* et *l'horizon* ; le contraste avec *Notre-Dame de Paris* est saisissant où ce sont les personnes et les expressions verbales qui dominent dans le tableau : *Louis XI*, *dom Claude*, *jeune fille*, *maître Jacques*, *nous avons*, *eut été*.

Tableau 6 – Les 20 plus hautes informations mutuelles des bigrammes apparaissant au moins 25 fois

Salammbô				*Notre-Dame de Paris*			
	1-10		11-20		1-10		11-20
peut être	10,48	c était	6,89	aujourd hui	11,51	quelque chose	8,60
narr havas	10,40	au bord	6,70	quinzième siècle	11,26	j ai	8,35
quelques uns	9,69	au fond	6,67	louis xi	10,87	cria t	8,34
quelque chose	9,51	s écria	6,58	dom claude	10,10	maître jacques	8,25
j ai	9,43	ceux qui	6,28	claude frollo	9,66	nous avons	8,18
sans doute	9,30	du côté	6,23	notre dame	9,34	ses lèvres	7,59
c est	7,76	se trouvait	6,16	peut être	8,92	eût été	7,50
ou bien	7,43	par dessus	6,12	sans doute	8,80	je vais	7,45
au milieu	7,13	l acropole	6,11	jeune fille	8,70	s approcha	7,30
tandis que	7,07	l horizon	6,01	eh bien	8,65	s écria	7,30

7 Conclusion

Dans cette petite analyse, nous avons présenté les notions d'entropie et d'entropie relative et nous les avons appliquées à trois textes pour les comparer à *Salammbô*. Dans chacun des exemples, nous avons pu voir que les résultats numériques correspondaient bien à l'idée que nous pouvions avoir de la proximité entre ces textes. Nous avons aussi montré comment les mots les plus fréquents ou les associations de mots récurrentes pouvaient mettre en lumière certaines marques distinctives d'une œuvre.

Le lecteur intéressé pourra reprendre les programmes que nous avons créés, les modifier, les améliorer ou leur adjoindre d'autres fonctions. Nous espérons au bout du compte que notre petite étude pourra servir les études littéraires et donner lieu à des analyses encore plus précises de l'œuvre de Flaubert.

Remerciements

Je voudrais remercier chaleureusement Monsieur François Lapelerie qui m'a invité à publier ce travail pour le site d'études sur Flaubert de l'université de Rouen et qui m'a ainsi permis de renouveler et d'approfondir une étude précédemment esquissée dans mon livre sur le traitement des langues (Nugues, 2006, 2014).

Annexe

Pour reproduire les analyses de cet article, nous avons mis à disposition le corpus des quatre textes provenant de Wikisource et les trois programmes en Python :

1. Le programme `symboles.py` calcule les fréquences relatives d'un texte. Il s'utilise de la manière suivante :

   ```
   python3 symboles.py -[normalise] fichier
   ```

   où l'option facultative `normalise` transforme toutes les lettres en majuscules et `fichier` contient le texte à analyser.

   On extrait l'ensemble des caractères de *Salammbô* avec leurs fréquences par la commande suivante :

   ```
   python3 symboles.py corpus/salammbo.txt
   ```

2. Le programme `divergence.py` calcule la divergence entre deux œuvres :

   ```
   python3 divergence.py fichier_p fichier_q
   ```

   Par exemple :

   ```
   python3 divergence.py corpus/coeursimple.txt corpus/salammbo.txt
   ```

Nota bene : Le programme de calcul de $D_{KL}(P||Q)$ est assez simple à écrire, si ce n'est que l'on doit traiter le problème pratique suivant : la divergence de Kullback et Leibler n'est pas définie lorsqu'un caractère a une fréquence nulle dans la distribution Q ; on a alors une division par zéro. Ce cas se présente concrètement dans notre corpus, avec, par exemple, le caractère ñ qui est absent aussi bien de *Salammbô* que du *Conte de deux cités*, mais qu'on trouve dans *Notre-Dame de Paris* dans la phrase :

> Señor caballero, para comprar un pedaso de pan !

Pour le résoudre, nous avons fait l'approximation suivante : lors du calcul de $D_{KL}(P||Q)$, nous n'avons considéré que les caractères communs aux deux distributions P et Q ; pour ñ, ceci signifie que nous l'avons tout simplement ignoré. Si cette façon de faire n'est pas complètement satisfaisante d'un point de théorique, elle simplifie beaucoup les calculs. Une amélioration possible du programme serait d'estimer les probabilités des fréquences nulles par une technique de lissage, Laplace (1820) ou autre.

3. Le programme `mots.py` calcule les fréquences des mots, des bigrammes et l'information mutuelle. Il s'utilise de la manière suivante :

   ```
   python3 mots.py fichier option
   ```

   où `option` peut être soit `mots`, `bigrammes`, ou `im` suivi d'une valeur de seuil. Par exemple
   - `python3 mots.py corpus/salammbo.txt mots`
     calcule les fréquences des mots dans *Salammbô*.
   - `python3 mots.py corpus/notredame.txt bigrammes`
     calcule les fréquences des bigrammes dans *Notre-Dame de Paris*.
   - `python3 mots.py corpus/notredame.txt im 25`
     calcule l'information mutuelle des bigrammes dans *Notre-Dame de Paris* et affiche les valeurs les plus hautes. Le seuil de bigrammes est de 25.

Références

Fano, R. M. 1961, *Transmission of Information : A Statistical Theory of Communications*, MIT Press, New York.

Kullback, S. et R. A. Leibler. 1951, «On information and sufficiency», *Annals of Mathematical Statistics*, vol. 22, n° 1, p. 79–86.

Laplace, P. 1820, *Théorie analytique des probabilités*, 3e éd., Coursier, Paris.

Nugues, P. M. 2006, *An Introduction to Language Processing with Perl and Prolog. An Outline of Theories, Implementation, and Application with Special Consideration of English, French, and German*, Springer Verlag, Berlin Heidelberg New York.

Nugues, P. M. 2014, *Language Processing with Perl and Prolog. An Outline of Theories, Implementation, and Application*, 2e éd., Springer Verlag, Berlin Heidelberg New York.

Shannon, C. E. 1948, «A mathematical theory of communication», *Bell System Technical Journal*, vol. 27, p. 398–403 and 623–656.