Sélection et ordre

Christian Rivière

(dernière révision : 9 avril 2020)

Anecdote :

A l'origine, et dans le contexte de la quantification du tri postal, il s'agissait de savoir si la « quantité de tri » incluait ou non la « quantité d'ordonnancement » (d'une tournée de facteur par exemple). La réponse n'est pas très évidente dans ce contexte car l'action de séquencement, donc de mise en ordre des directions (du moins quand elle est faite manuellement) s'effectue après l'opération de tri. Mais il suffit de décrire le plan de tri une fois pour toutes dans l'ordre, pour s'apercevoir que la mise en ordre fait partie du tri. Il était donc nécessaire que soit vérifiée l'inégalité suivante :

$$\frac{N^N}{\prod_i n_i^{n_i}} > \frac{N!}{\prod_i n_i!} \; si \; N = \sum_i n_i$$

Ce sont bien des considérations d'ordre linguistique *(information = sélection de l'information + mise en ordre de celle-ci), qui m'ont mis sur la voie. En effet, la différence entre les deux expressions (au logarithme près) correspondant à l'information de sélection, celle-ci était de ce fait positive.*

La généralisation s'est faite ensuite pour un échantillon ayant une répartition n_i pas obligatoirement identique à celle de la source (n'_i), ce qui permet d'affirmer que :

$$\frac{N'^N}{\prod_i n'_i{}^{n_i}} > \frac{N!}{\prod_i n_i!} \; si \; N = \sum_i n_i \; et \; N' = \sum_i n'_i$$

Avant-propos

Si un évènement se produit avec une probabilité p (par exemple l'apparition d'un mot lors d'une lecture), la quantité d'information contenue dans cet évènement est $Q_{info} = \log_2 \frac{1}{p}$

Dans une source d'information : N' élément, de répartition n'_i avec $N' = \sum_i n'_i$ le tirage aléatoire d'un élément d'un élément de type i a pour probabilité $p_i = \frac{n'_i}{N'}$

La quantité d'information est donc : $Q_{info\ i} = \log \frac{1}{p_i} = \log \frac{N'}{n'_i}$

Par exemple, la source d'information : A B B A C A D A

N' = 8

Si le tirage est :

un A : $n'_A = 4$	$p_A = 1/2$	$Q_{info\ A} = 1$ bit
un B : $n'_B = 2$	$p_B = 1/4$	$Q_{info\ B} = 2$ bits
un C : $n'_C = 1$	$p_C = 1/8$	$Q_{info\ C} = 3$ bits
un D : $n'_D = 1$	$p_D = 1/8$	$Q_{info\ D} = 3$ bits

Le tirage aléatoire d'un second élément de type i pose le problème suivant : les probabilités p_i des différentes possibilités, sont-elles les mêmes que pour le tirage du premier élément ?

. la réponse est non si la source est finie et que la source (donc la probabilité) est altérée par le tirage du premier élément ;

. la réponse est oui si la source est finie et que la source (donc la probabilité) n'est pas altérée par le tirage du premier élément (restitution à la source du premier élément tiré avant le tirage du second) ;

. la réponse est a priori oui si la source est « infinie » (ce qui n'est jamais le cas), disons si N' est très grand (typiquement : le tri postal) ;

. la réponse est non dans le cas d'un langage (tout particulièrement le langage humain), bien que la source soit « infinie » dans ce cas. Les probabilités p_i dépendent des tirages précédents au fur et à mesure de la construction de la signification (des caractères, des mots, des phrases, … précédents).

Dans le cas d'une source « infinie », les probabilités sont déterminées statistiquement par l'expérience. En particulier pour le langage humain, on utilise un corpus de textes le plus étendu et représentatif possible, de façon à ce que les variations de p_i ne soient dues qu'à l'évolution lente de l'usage de la langue concernée.

Dans ce cas, les valeurs N' et n'_i ne représentent rien de concret, si ce n'est une proportion significative de tel ou tel élément i. D'ailleurs dans ce cas, il ne s'agit pas obligatoirement de valeurs entières.

Dans la suite, la notation suivante est utilisée :

. pour la source : $N' = \sum_i n'_i$ comme ci-avant ;

. pour l'échantillon : $N = \sum_i n_i$ (répartition dans l'échantillon après N tirages).

Il faut tout d'abord traiter un cas trivial utilisé plus loin. Si la source est identique à l'échantillon reçu, la quantité d'information se résume à la mise en ordre de l'échantillon ($\forall i \quad n_i = n'_i$). Donc dans ce cas :

$$Q_{info} = Q_{ordre} = \sum_i \sum_{k=0}^{n_i - 1} \log \frac{N - \sum_{j<i} n_j - k}{n_i - k} = \log \frac{N!}{\prod_i n_i!}$$

Quantité d'information quand les tirages sont indépendants (source infinie ou finie avec remise)

Nota 1 : Ce sont ces hypothèses qui sont retenues dans le cas général d'une mesure du tri postal. En effet, la source n'est connue qu'au travers de statistiques faites sur des (millions de) plis par jour sur des périodes calendaires assez longues. Mais ces données statistiques ont leurs limites (variations saisonnières, selon les jours de la semaine, évolution sociétale de l'usage du courrier, évènements exceptionnels, …)

Nota 2 : Ce sont ces hypothèses qui sont retenues dans le cas général d'une mesure d'information en langage humain (qu'elle soit écrite ou orale). En effet, la source n'est connue qu'au travers de statistiques faites sur des (milliers de) corpus. Malheureusement, ce qui suit ne fonctionne pas pour les langages (humains ou non d'ailleurs) puisque les probabilités évoluent au fur et à mesure du texte. A noter que ces probabilités n'évoluent pas comme dans le cas d'une source finie, donc la présentation suivante (relative à une source finie) ne fonctionne pas non plus pour les langages.

Nota 3 : Ces hypothèses sont valables dans le cas d'une source finie avec remise. En effet, les probabilités n'évoluent pas au fur et à mesure des tirages.

Quantité d'information reçue quand la source est infinie (répartition n'_i). L'échantillon reçu a la répartition n_i :

$$Q_{info} = \sum_i n_i \log \frac{N'}{n'_i} = \log \frac{N'^N}{\prod_i n'^{n_i}_i} = \log \frac{1}{\prod_i p_i^{n_i}} \qquad \text{avec : } \sum_i n_i = \sum_i n'_i = N < N' \quad \text{et} \quad \frac{n'_i}{N'} = p_i$$

n'_i sont fictifs ou pas, tous non nuls, et pas obligatoirement entiers. Leur somme peuvent être très inférieure à N' si on n'utilise pas tout le thésaurus.

Quantité d'information due à l'action de mise en ordre de l'échantillon :

$$Q_{ordre} = \sum_i \sum_{k=0}^{n_i-1} \log \frac{N - \sum_{j<i} n_j - k}{n_i - k} = \log \frac{N!}{\prod_i n_i!}$$

Quantité d'information due à l'action de sélection de l'échantillon à partir de la source infinie :

$$Q_{sélection} = \log \prod_i \frac{1}{C_{n_i}^{N-\sum_{j<i} n_j} \left(\frac{n'_i}{N' - \sum_{j<i} n'_j}\right)^{n_i} \left(1 - \frac{n'_i}{N' - \sum_{j<i} n'_j}\right)^{N-\sum_{j<i} n_j - n_i}}$$

Nota : Au dénominateur, la loi binomiale traduit le nombre de possibilités de placer n_i éléments i dans l'espace restant dans N (les éléments j *(avec $j<i$)* ayant été traités), multiplié par la probabilité qu'ils soient de type i, multiplié par la probabilité que les autres ne soient pas de type i. *(cf. programme de mathématique de classe de première)*

Maintenant, on simplifie :

$$Q_{sélection} = \log \prod_i \frac{n_i! \, (N - \sum_{j<i} n_j - n_i)! \, (N' - \sum_{j<i} n'_j)^{n_i} \, (N' - \sum_{j<i} n'_j)^{N-\sum_{j<i} n_j - n_i}}{(N - \sum_{j<i} n_j)! \, n'^{n_i}_i \, (N' - \sum_{j<i} n'_j - n'_i)^{N-\sum_{j<i} n_j - n_i}}$$

$$Q_{sélection} = \log \prod_i \frac{n_i! \, (N - \sum_{j<i} n_j - n_i)! \, (N' - \sum_{j<i} n'_j)^{N-\sum_{j<i} n_j}}{n'^{n_i}_i \, (N - \sum_{j<i} n_j)! \, (N' - \sum_{j<i} n'_j - n'_i)^{N-\sum_{j<i} n_j - n_i}}$$

$$Q_{sélection} = \log \frac{\prod_i n_i! N'^N}{N! \prod_i n'_i{}^{n_i}} = \log \frac{N'^N}{\prod_i n'_i{}^{n_i}} - \log \frac{N!}{\prod_i n_i!} = Q_{info} - Q_{ordre}$$

Donc, on a bien : $\boxed{Q_{info} = Q_{sélection} + Q_{ordre}}$

Quantité d'information quand les tirages sont dépendants (source finie sans remise)

Quand la source (n'_i) n'est pas infinie et qu'il n'y a pas de remise (ex : jeu de scrabble) :

$$Q_{info} = \log \frac{N'! \prod_i (n'_i - n_i)!}{(N'-N)! \prod_i n'_i !} = \log \frac{A^N_{N'}}{\prod_i A^{n_i}_{n'_i}} \qquad \text{avec : } \sum_i n_i = N \le \sum_i n'_i \le N' \text{ et } \forall i,\ n_i \le n'_i$$

Nota : Les n'_i sont ici obligatoirement entiers.

Quantité d'information due à l'action de mise en ordre de l'échantillon :

$$Q_{ordre} = \log \frac{N!}{\prod_i n_i!}$$

Quantité d'information due à l'action de sélection de l'échantillon à partir de la source :

$$Q_{sélection} = \log \prod_i \frac{1}{C^{n_i}_{N-\sum_{j<i} n_j} \dfrac{(n'_i !...(n'_i - n_i + 1)!)}{((N'-\sum_{j<i} n_j)!...(N'-\sum_{j<i} n_j - n_i + 1)!)} \dfrac{((N'-\sum_{j<i} n_j)!...(N'-\sum_{j<i} n_j - n_i + 1)!}{((N'-\sum_{j<i} n_j - n_i)!...(N'-N+1)!)}}$$

Nota : Au dénominateur, la loi binomiale a une allure différente, puisque la composition de la source (celle-ci étant non infinie) change à chaque sélection.

$$Q_{sélection} = \log \prod_i \frac{(n'_i - n_i)!(N'-\sum_{j<i} n_j)!(N'-\sum_{j<i} n_j - n_i)!(N'-\sum_{j<i} n_j - n_i)!}{C^{n_i}_{N-\sum_{j<i} n_j} n'_i !(N'-\sum_{j<i} n_j - n_i)!(N'-\sum_{j<i} n_j)!(N'-N)!}$$

$$Q_{sélection} = \log \prod_i \frac{(n'_i - n_i)!(N'-\sum_{j<i} n_j - n_i)!}{C^{n_i}_{N-\sum_{j<i} n_j} n'_i !(N'-N)!}$$

$$Q_{sélection} = \log \prod_i \frac{n_i!(N-\sum_{j<i} n_j - n_i)!(n'_i - n_i)!(N'-\sum_{j<i} n_j - n_i)!}{(N-\sum_{j<i} n_j)! n'_i !(N'-N)!}$$

$$Q_{sélection} = \log \prod_i \frac{n_i!(n'_i - n_i)!(N-\sum_{j<i} n_j - n_i)!(N'-\sum_{j<i} n_j - n_i)!}{n'_i !(N'-N)!(N-\sum_{j<i} n_j)!}$$

$$Q_{sélection} = \log \frac{\prod_i (N'-\sum_{j<i} n_j - n_i)! n_i!(n'_i - n_i)!}{N! \prod_i (N'-N)! n'_i !}$$

$$Q_{sélection} = \log \frac{N'! \prod_i n_i!(n'_i - n_i)!}{(N'-N)! N! \prod_i n'_i !} = \log \frac{C^N_{N'}}{\prod_i C^{n_i}_{n'_i}} \text{ (généralisation de la loi hypergéométrique)}$$

Or, en reprenant les expressions ci-dessus, on a aussi :

$$Q_{sélection} = Q_{info} - Q_{ordre} = \log \frac{N'! \prod_i (n'_i - n_i)! n_i!}{N!\,(N'-N)! \prod_i n'_i!} = \log \frac{C^N_{N'}}{\prod_i C^{n_i}_{n'_i}}$$

Donc, on a toujours : $\boxed{Q_{info} = Q_{sélection} + Q_{ordre}}$

Ce qui peut s'exprimer par :

$$\boxed{\log \frac{A^N_{N'}}{\prod_i A^{n_i}_{n'_i}} = \log \frac{C^N_{N'}}{\prod_i C^{n_i}_{n'_i}} + \log \frac{N!}{\prod_i n_i!}} \text{ avec : } \sum_i n_i = N \le \sum_i n'_i = N' \text{ et } \forall i,\ n_i \le n'_i$$

Si $N' \to \infty$ donc si la source devient infinie (ou presque), à partir de la formule précédente avec : $\sum_i n_i = N << \sum_i n'_i = N'$ et $\forall i,\ n_i << n'_i$ on obtient à la limite :

$$\boxed{\log \frac{N'^N}{\prod_i n'^{n_i}_i} = \log \frac{N'^N \prod_i n_i!}{N! \prod_i n'^{n_i}_i} + \log \frac{N!}{\prod_i n_i!}} \text{ (formule vue plus haut)}$$

Etant donné que $\forall i$ N'/n'_i représente l'inverse de la probabilité de l'élément *i* on peut revenir à : $\sum_i n_i = \sum_i n'_i = N \le N'$ (mais les n'_i ne sont plus obligatoirement entiers)

Interprétation

La quantité d'information totale est la stricte somme de l'information de sélection (acquisition de l'ensemble) et de l'information de mise en ordre (production d'une liste ordonnée). Cette dernière est indépendante de la répartition statistique de la source.

Dualité entre la sélection et l'ordre :

Cette dualité ne se retrouve pas que dans la mesure de l'information à proprement parler. D'autres phénomènes ou domaines d'activité sont concernés. Citons en trois :

. l'activité de tri (tri postal)

$Q_{tri} = Q_{identification} + Q_{ordonnancement}$ respectivement reconnaissance et interprétation du pli et mise en ordre (ordre de la tournée du facteur par exemple). La mise en ordre est une partie identifiable de la quantité de tri.

. l'activité du jeu (jeu de cartes)

$Q_{tirage} = Q_{main} + Q_{ordre}$ Dans la plupart des jeux de cartes, la valeur du jeu (d'un joueur) ne dépend pas de l'ordre du tirage. C'est sans doute ce qui laisse au joueur une marge d'initiative, sans quoi le jeu n'aurait aucun intérêt.

. l'activité de mémorisation (liste de mots par exemple)

$Q_{mémorisation} = Q_{acquisition} + Q_{attribution}$

Par exemple, pour la mémorisation des capitales des pays suivants, il est possible d'arriver à la situation où les mots sont connus, sans qu'on soit capable de les associer correctement aux pays.

Ouzbékistan	Tachkent
Tadjikistan	Douchanbe
Kazakhstan	Astana
Kirghizistan	Bichkek
Turkménistan	Achgabat
Azerbaïdjan	Bakou

Nota : Il semble que les méthodes de mémorisation ne prennent pas toutes en compte ce point. C'est dommage.

Terminologie

L'ensemble des différentes valeurs de *i* (caractérisant la source) s'appelle un **alphabet** au sens très large du terme. Bien entendu, si les entités sélectionnées ne sont pas des caractères, il est possible d'utiliser différents termes : système de numération, lexique, dictionnaire, annuaire, base d'adresse (pour le tri postal), répertoire, **thésaurus**, … Le nombre d'éléments de l'alphabet *Card(alphabet)=max i*

Une valeur indicée par *i* est un élément de l'alphabet, soit un caractère, une lettre, un signe, une possibilité, un chiffre, une adresse (tri postal), …

L'ensemble des *N* éléments issus de la source forment un message, une phrase, un mot (si les éléments sont des lettres), un **corpus**, une main (jeu de carte), un résultat (fruit du hasard), un numéro, un identifiant, un code … dont nous nous proposons de déterminer la quantité d'information.

n'_i / N' exprime la probabilité de tirage de l'élément *i* de l'alphabet.

n_i s'appelle l'occurence de l'élément *i* (de l'alphabet) dans le message.

Unité

Le choix le plus simple et le plus international est le bit. Pour ce faire, tous les logarithmes mentionnés sont en base 2.

Néanmoins, d'autres unités peuvent se rencontrer : octet en informatique, Ashley en linguistique (largement oublié), pixel en traitement d'image, digit en numérique, …

Suite de la réflexion

Dans un premier temps, il s'agit de comprendre pourquoi il y a moins d'information dans « le chat mange la souris » que dans « la souris mange le chat » (alors que les mots sont les mêmes). De même : « étrange » et « gérante », ou encore « La crise économique » et « Le scénario comique » alors que les caractères sont les mêmes. C'est obligatoirement le dernier terme Q_{ordre} qui est en cause.

Prenons un exemple plus simple :

Tirages pile ou face (l'alphabet a deux caractères, le résultat a huit caractères)
PFFPFPPF $Q_{info}=log_2 256=8$ $Q_{ordre}=log_2 70=6,13$ donc $Q_{sélection}=log_2 3,657=1.87$ ($n_1=4$ $n_2=4$)
PFFPFPFF $Q_{info}=log_2 256=8$ $Q_{ordre}=log_2 56=5,81$ donc $Q_{sélection}=log_2 4,653=2,19$ ($n_1=3$ $n_2=5$)

Commentaire :
Quand la source fournit des éléments équiprobables, Q_{info} est constant.

Imaginons que « pile » soit trois fois plus probable que « face » ($n'_1=6$ $n'_2=2$) :
PPPFPFPP $Q_{info}=log_2 89,90=6,49$ $Q_{ordre}=log_2 28=4,81$ donc $Q_{sélection}=log_2 3,20=1.68$ ($n_1=6$ $n_2=2$)
PFFPFPPF $Q_{info}=log_2 809,08=9,66$ $Q_{ordre}=log_2 70=6,13$ donc $Q_{sélection}=log_2 11,55=3.53$ ($n_1=4$ $n_2=4$)
PFFPFPFF $Q_{info}=log_2 2427,26=11,25$ $Q_{ordre}=log_2 56=5,81$ donc $Q_{sélection}=log_2 43,41=5.44$ ($n_1=3$ $n_2=5$)
FPPFPFPP $Q_{info}=log_2 269,70=8,08$ $Q_{ordre}=log_2 56=5,81$ donc $Q_{sélection}=log_2 4,82=2.27$ ($n_1=5$ $n_2=3$)

Commentaires :
Q_{info} est maximale quand la répartition du résultat est celle de la source. Ce maximum est minimal à l'équi-probabilité de la source (*).
$Q_{sélection}$ est minimale quand la répartition du résultat est celle de la source (idem pour Q_{info}). En effet, plus elle s'éloigne de cette situation, moins le tirage est probable (loi binomiale) et plus $Q_{sélection}$ est grande. Ce minimum est maximal à l'équi-répartition.
Q_{ordre} est maximale quand les éléments du résultat sont équi-répartis.

(*) le lemme de Gibbs permet de prouver que l'entropie est maximale pour une répartition uniforme des probabilités.

Enoncé :

Soient $\mathrm{P} = (p_1, p_2, ..., p_n)$ et $Q = (q_1, q_2, ..., q_n)$, deux vecteurs de probabilités ($p_i, q_i \geq 0$ et $\sum_i p_i = \sum_i q_i = 1$).

Alors, $\sum_i^n p_i \log \frac{1}{p_i} \leq \sum_i^n p_i \log \frac{1}{q_i}$ avec égalité si et seulement si $\forall i,\ p_i = q_i$.

Démonstration du lemme de Gibbs

$$\forall x > 0,\ Ln\, x \leq x - 1 \Rightarrow Ln \frac{q_i}{p_i} \leq \frac{q_i}{p_i} - 1 \Rightarrow \log \frac{q_i}{p_i} \leq \frac{1}{Ln\, 2}\left(\frac{q_i}{p_i} - 1\right)$$

$$\Rightarrow \sum_i^n p_i \log \frac{q_i}{p_i} \leq \frac{1}{Ln(2)} \sum_i^n p_i \left(\frac{q_i}{p_i} - 1\right) = \frac{1}{Ln(2)}\left(\sum_i^n q_i - \sum_i^n p_i\right) = 0$$

$$\Rightarrow \sum_i^n p_i \log \frac{1}{p_i} \leq \sum_i^n p_i \log \frac{1}{q_i}$$

En appliquant le lemme précédent avec $q_1 = q_2 = ... = q_n = \frac{1}{n}$, nous obtenons :

$$\sum_i^n p_i \log \frac{1}{p_i} \leq \sum_i^n p_i \log n \Rightarrow \sum_i^n p_i \log \frac{1}{p_i} \leq \log n . \sum_i^n p_i \Rightarrow \sum_i^n p_i \log \frac{1}{p_i} \leq \log n$$

On a donc toujours $\sum_i p_i \log \frac{1}{p_i} \leq \log n$, avec égalité si et seulement si $p_i = \frac{1}{n}$, donc l'entropie est maximale pour une répartition uniforme des probabilités.

Cas des langues naturelles

Quand les différents ordres sont équiprobables : $Q_{ordre} = \log \frac{N!}{\prod_i n_i!} = \log \frac{1}{\prod_i p_i}$

Si tous les ordres étaient possibles et équiprobables (pour les mots « chat », « la », « le », « mange », « souris »), la mesure de Q_{ordre} serait $\log 5 ! (\approx 7)$

Mais dans le cas des langues naturelles, tous les ordres possibles ne sont pas équiprobables, certains sont même impossibles car interdits par les règles de syntaxe. Résultat :

$Q_{ordre} \neq \log \frac{N!}{\prod_i n_i!}$ et la valeur moyenne $\overline{Q}_{ordre} << \log \frac{N!}{\prod_i n_i!}$

Quelques exemples :

Notons que « le chat regarde la souris » et « la souris regarde le chat » ont quasiment la même quantité d'information Q_{info}. Or, $Q_{selection}$ étant identique, Q_{ordre} a aussi la même valeur approximative.

« le chat regarde la souris » (ou le contraire) : $Q_{ordre} = \log 2 = 1$

« le chat mange la souris » (et pas le contraire) : $Q_{ordre} \approx \log 1 = 0$ (cette phrase ne contient pratiquement que de l'information de sélection)

En fait, le propre des langages humains est de permettre d'exprimer tout, y compris (intentionnellement) ce qui est faux, absurde, stupide, illogique, mensonger, … Ainsi la phrase « la souris mange le chat » est hautement improbable (mais a un sens), donc contient de ce fait énormément d'information. Comme la valeur de $Q_{sélection}$ est identique pour les deux phrases, c'est donc bien la valeur de Q_{ordre} qui est très grande dans ce cas.

Les différents ordres possibles d'un groupe de mots composant une phrase ne sont pas équiprobables pour une langue naturelle.

Si les cinq mots de la phrase pouvaient se retrouver dans n'importe quel ordre, on aurait :

$$Q_{ordre} = \log \frac{1}{\frac{1}{5}.\frac{1}{4}.\frac{1}{3}.\frac{1}{2}.\frac{1}{1}} = \log 5! = \log 120 \approx 7$$

« La souris regarde le chat » ou « Le chat regarde la souris » (le premier mot peut être « Le » ou « La », puis les autres s'imposent) :

$$Q_{ordre} = \log \frac{1}{\frac{1}{2}.\frac{1}{1}.\frac{1}{1}.\frac{1}{1}.\frac{1}{1}} = \log 2 = 1$$

« Le chat mange la souris » (les mots sont tous à leur place la plus probable) :

$$Q_{ordre} = \log \frac{1}{\frac{(1-\varepsilon)}{1}.\frac{1}{1}.\frac{1}{1}.\frac{1}{1}.\frac{1}{1}} \approx \varepsilon$$

« La souris mange le chat » (le mot « La » en premier est très improbable, les mots suivants s'imposent) :

$Q_{ordre} = \log \frac{1}{\frac{\varepsilon}{1}.\frac{1}{1}.\frac{1}{1}.\frac{1}{1}.\frac{1}{1}}$ soit une valeur très élevée (qui est explicitée ci-après)

D'après des études statistiques faites sur la langue anglaise il y a plusieurs décennies, la quantité d'information d'une phrase simple, courante comme « La souris regarde le chat » ou « Le chat mange la souris » est vraisemblablement d'un dizaine de bits environ. Par souci de simplicité, nous conserverons la valeur de 10 bits pour $Q_{selection}$ dans la suite.

Contrairement au jet de dés ou d'autres sources d'information simples, l'écart entre Q_{info} et Q_{ordre} est très important, ce qui veut dire qu'une fois les mots sélectionnés, la mise en ordre des mots apporte relativement peu d'information dans ce type de phrase. Cela vient de deux causes :

. $Q_{selection}$ est très grand, car bien entendu aucun texte n'utilise la totalité du lexique (ni du corpus ni du thésaurus), ce qui se traduit par : $\sum_i n'_i << N'$

. Q_{ordre} est très petit car les règles de syntaxe sont drastiques (la redondance syntaxique est très importante).

Si la phrase « La souris mange le chat » est 1000 fois (disons 1024 pour avoir une puissance de 2) moins fréquente que « Le chat mange la souris », c'est que sa quantité d'information Q_{info} vaut 20 bits (10+10, étant donné que log 1024=10), et que l'information Q_{ordre} due à la mise en ordre vaut 10 bits, ce qui est effectivement énorme.

Au lieu de considérer que les phrases ci-dessus ont chacune cinq éléments :
la, le, chat, regarde, souris → $N=5$ $n_{la}=1$ $n_{le}=1$ $n_{chat}=1$ $n_{regarde}=1$ $n_{souris}=1$

il est possible de voir le problème différemment :
la souris, le chat, regarde, → $N=3$ $n_{la\ souris}=1$ $n_{le\ chat}=1$ $n_{regarde}=1$

ou bien :
a, c, d, e, g, h, i, l, o, r, s, t, u, espace → $N=25$ $n_a=2$ $n_c=1$ $n_d=1$ $n_e=3$ $n_g=1$ $n_h=1$ … $n_{« »}=4$

Bien sûr, cela ne modifie pas les valeurs de Q_{info} mais les valeurs de $Q_{selection}$ et Q_{ordre} seraient sans doute modifiées, et donc leurs parts respectives aussi. Dans le dernier cas (description du problème à partir des caractères), $Q_{selection}$ serait plus faible et Q_{ordre} plus grande.

-=-=-=-=-=-=-=-=

Que se passe-t-il dans le cas d'une langue où « l'ordre des éléments est libre », typiquement en espéranto ?
Pour mieux comparer (entre les deux langues), il est plus simple de prendre le modèle N=3, soient les trois éléments : « la souris » « le chat » « mange », ce qui correspond en espéranto aux trois éléments, respectivement « la muson » « la kato » « manĝas ».
Posons comme postulat que la valeur de Q_{info} est la même pour la phrase « Le chat mange la souris » que pour la phrase « La kato manĝas la muson ».

En espéranto, la même phrase peut s'exprimer de six façons différentes :

« La kato manĝas la muson »
« La kato la muson manĝas »
« La muson manĝas la kato »
« La muson la kato manĝas »
« Manĝas la kato la muson »
« Manĝas la muson la kato »

Nota : en espéranto, c'est le « n » situé à la fin de « muso » qui indique que c'est la souris qui est mangée, et non le chat.

Il s'ensuit que l'information de sélection $Q_{selection}$ est déjà suffisante pour comprendre (donc égale à Q_{info} pour la phrase française). Or, Q_{ordre} n'est pas nulle !! Sa valeur moyenne sur les six phrases (si elles étaient équiprobables, ce qui n'est certainement pas le cas) est $\log_2 6$, ce qui fait environ 2,5 bits.

On en déduit que : $Q_{info\ espéranto} > Q_{info\ français}$!!!

Comment expliquer ce paradoxe ?

En fait, il s'agit là de deux points de vue différents dans l'analyse du problème. Cette dualité se retrouve dans l'évaluation de $Q_{selection}$ dans le cas des synonymes. Il est possible de présenter le problème de la façon suivante :

. point de vue n°1 : le sens des deux expressions ou phrases est strictement identique. Le sens ne dépend pas du choix de l'une ou l'autre forme, donc de l'ordre des mots dans le cas qui nous occupe, les six formes sont « synonymes » en espéranto, Q_{ordre} ne doit pas être prise en compte dans le calcul de Q_{info}. On a $Q_{ordre} = 0$ et $Q_{info\ espéranto} = Q_{info\ français}$ Dans cette optique, l'espéranto, langue internationale, propose un choix de possibilités strictement équivalentes, c'est-à-dire qu'elles ont le même sens. Le locuteur choisit telle ou telle forme en fonction de ses habitudes, par exemple sous l'influence de sa langue maternelle.

. point de vue n°2 : quand deux expressions ou phrases ne sont pas identiques, elles ont obligatoirement des sens différents, et a priori des valeurs de Q_{info} aussi. La phrase « La kato manĝas la muson » étant plus fréquente que « La muson la kato manĝas » ou « Manĝas la kato la muson », Q_{info} est obligatoirement moindre. Quelle est donc alors la nature de l'information supplémentaire dans les deux dernières phrases ? Peut-être la précision qu'il s'agit bien de la souris, ou de l'action de manger, ce que l'on traduit à l'oral par une accentuation de la voix, ou par les formes pas toujours correctes syntaxiquement « C'est la souris que le chat mange », « C'est la souris qu'il mange, le chat », « Il la mange le chat la souris ». Dans cette optique, la stricte égalité $Q_{info\ espéranto} = Q_{info\ français}$ est pratiquement impossible, donc implicitement la fidélité d'une traduction est illusoire.

. point de vue n°3 : l'espéranto est une langue textuellement moins redondante, donc la part d'information « utile » dans la phrase est plus importante.

. point de vue n°4 : la redondance contextuelle n'est bien sûr pas présente dans l'information textuelle. Il n'en reste pas moins que la probabilité (donc la quantité d'information) d'une phrase dépend indiscutablement du contexte. Il n'y a donc aucune raison que les quantités d'information aient la même valeur dans les deux langues. D'ailleurs, cette remarque vaut aussi pour la même phrase (dans la même langue !) selon qu'elle est exprimée dans un pays où pullulent ces deux animaux par rapport à un autre pays où l'un des deux n'existe pas.

Le débat n'est pas tranché, d'autant moins que chaque point de vue s'appuie sur de solides arguments.

Cas des langues naturelles (Quantité d'information due à la sélection)

La présentation probabiliste de $Q_{selection}$ bien qu'étant d'intérêt théorique évident, ne nous aide pas vraiment dans le cas des langues naturelles (très compliqué en espéranto, inextricable en français). *Umberto Eco lui-même est arrivé à la conclusion que c'était impossible.*

Le nombre de problèmes à affronter est impressionnant (ne parlons bien sûr que de la **langue écrite**, à moins d'être complètement fou). Bien que ce ne soit pas l'objet de ces quelques pages, citons parmi ce qui peut influer sur l'information et donc sur la mesure de $Q_{selection}$:

. le choix du thésaurus le plus approprié (les mots, les éléments de la phrase, les phonèmes, les caractères y compris accentués ou avec un appendice, les majuscules). En espéranto, on peut ajouter les racines élémentaires à cette liste

. selon le choix du thésaurus, faut-il ajouter l'espace, certains signes de ponctuations, trait d'union, apostrophe, trait de dialogue, parenthèses, …

. que faire des noms propres qui pullulent dans les corpus mais sont inconnus des thésaurus

. comment traiter les homonymes (exemple le mot « ferme » en français qui peut être un nom, un adjectif, un verbe ou un adverbe)

. comment traiter les mots à plusieurs orthographes (« dièse » / « dièze » ; « gaîté » / « gaieté » ; « clé » / « clef », …)

. comment traiter les synonymes, les relations de généricité, les niveaux d'abstraction …

. la conjugaison (pluriel, temps des verbes, …) les différentes formes (par exemple du même verbe) sont-elles considérées comme un même mot malgré les différences morphologiques

. comment prendre en compte un nouveau mot (dans aucun corpus mais déjà à la mode)

. comment traiter les marques particulières (souligné, gras, italique, …)

Mais le plus grave, c'est que les probabilités des éléments d'un texte ne sont pas indépendantes. Elles dépendent toutes les unes des autres (notion de contexte textuel). Enfin, elles dépendent énormément du contexte lui-même.

Nota : entre les cas extrêmes des jets de dés (*) ou simples codes d'une part, et les langues naturelles d'autre part, il existe des sujets d'étude intermédiaires plus accessibles, par exemple les langages informatiques qui se situent à mi-chemin. Grâce à leur relative simplicité, l'étude ces langages est très pertinente et très instructive.

(*) Les mots « dé » et « donnée » ont exactement la même étymologie !! Eh oui, le dé donne un résultat.