Thermodynamique et Statistique Textuelle: concepts et illustrations.

François Bavaud* et Aris Xanthos

Section d'Informatique et de Méthodes Mathématiques et Section de Linguistique - Lettres - Université de Lausanne - CH-1015 Lausanne - Switzerland

Abstract

Statistical Language modelling is currently dominated by Information Theory, based upon Shannon's entropy. Yet, ever since Zipf and Mandelbrot, thermodynamic considerations (energy, temperature) have traditionnally constituted a source of inspiration in Textual Statistics. We briefly recall elements of thermodynamics and statistical physics, which we illustrate on textual problems such as the "heating" of texts, the unsupervised recovering of missing blanks, the estimation of textual temperature, the additive and multiplicative mixture of models, as well as the definition of indices of textual richness.

Keywords: Markov chains, Gibbs distribution, energy, entropy, unsupervised segmentation, temperature.

Résumé

La Théorie de l'Information, basée sur l'entropie de Shannon, s'impose en tant que formalisme dominant en modélisation du Langage. Cependant, les considérations thermodynamiques (énergie, température) ont également joué un rôle essentiel en Statistique textuelle dès les travaux de Zipf et de Mandelbrot. Comme le démontre la mécanique statistique, dont nous rappelons brièvement quelques principes, ces deux formalismes sont essentiellement équivalents. Le propos est illustré par quelques problèmes textuels, tels que le "chauffage" des textes, la détermination non supervisée des espaces manquants, les mélanges additifs et multiplicatifs de textes, et la définition thermodynamique d'indices de richesse textuelle.

Mots-clés : chaînes de Markov, distribution de Gibbs, énergie, entropie, segmentation non supervisée, température.

1. Introduction et concepts

Les concepts d'énergie et de température sont utilisés dans nombre de disciplines extérieures à la physique, parmi lesquelles la statistique textuelle. Les raisons en sont d'ordre heuristiques ou métaphoriques ("principe du moindre effort", "énergie de cohésion d'un texte", "désordre distributionnel", etc.) ainsi que formelles (algorithmes de recuit-simulé, distributions de Gibbs associées au théorème de Hammersley-Clifford ou au principe de maximum d'entropie, etc.).

Ce travail a pour but de rappeler et d'expliciter, dans une perspective historique, les bases essentielles du formalisme thermodynamique dans un contexte de statistique textuelle, de les illustrer, et de discuter des liens avec la Théorie de l'Information, aujourd'hui dominante en modélisation textuelle. Les thèmes formels abordés dans cette contribution sont généralement connus de longue date. Nous souhaitons toutefois que l'on voie un aspect novateur dans leur exposition unifiée et à double entrée (Thermodynamique ↔ Théorie de l'Information), ainsi que dans les

* également en Section de Psychologie de l'Université de Genève.

illustrations proposées ("chauffage de textes", segmentation textuelle non supervisée, estimation de la température d'un texte, mélanges additifs et multiplicatifs de modèles, indices de richesse lexicaux). La problématique parente quoique distincte des algorithmes de recuit-simulé (voir par exemple Rose (1998)) n'est pas discutée ici.

1.1. Rappel de thermodynamique

On considère un système physique pouvant prendre un certain nombre d'états $A \in \mathcal{A}$. Dans le formalisme de mécanique statistique à l'équilibre, le système tend à la fois à minimiser son *énergie* (moyenne) $u[p] := \sum_{A\in\mathcal{A}} P(A)\, U(A)$ (où $P(A)$ est la probabilité d'occuper l'état A et $U(A)$ l'énergie associée) et à maximiser son *entropie* $s[p] := -\sum_{A\in\mathcal{A}} P(A)\ \ln P(A)$. Ces deux tendances, contradictoires, sont arbitrées par la *température* $T > 0$ du système, de façon à ce que le système minimise globalement son *énergie libre* F définie par

$$F := u - Ts = \sum_{A\in\mathcal{A}} P(A)\, U(A) + T \sum_{A\in\mathcal{A}} P(A)\ \ln P(A) \tag{1}$$

dont le minimum (égal à $F_{\min} = -T \ln Z(\beta)$) est atteint par la distribution de Gibbs

$$P(A) = \frac{\exp(-\beta\, U(A))}{Z(\beta)} \qquad \beta := \frac{1}{T} \qquad Z(\beta) := \sum_{A'\in\mathcal{A}} \exp(-\beta\, U(A')) \tag{2}$$

A basse température $\beta >> 1$, l'énergie libre est contrôlée par sa composante énergétique et le système est essentiellement figé dans son *état fondamental* A_0, défini par $\min_{A\in\mathcal{A}} U(A) = U(A_0)$): on a $u \cong U(A_0)$ (minimal), et $s \cong 0$ (minimal). A l'inverse, à haute température $\beta << 1$, l'entropie domine et le système est essentiellement distribué de façon uniforme: $P(A) \cong$ constante, pour lequel $s \cong \ln|\mathcal{A}|$ est maximal.

L'énergie moyenne $u(\beta)$ et la chaleur spécifique $c(\beta)$ (qui est le rapport entre l'augmentation d'énergie et la diminution de température inverse) s'obtiennent comme

$$u(\beta) = \sum_{A\in\mathcal{A}} P(A)\, U(A) = -\frac{\partial \ln Z(\beta)}{\partial\beta} \qquad c(\beta) = -\frac{\partial u(\beta)}{\partial\beta} = \sum_{A\in\mathcal{A}} P(A)\, (U(A) - u(\beta))^2 \tag{3}$$

1.2. Retour aux arguments énergétiques en statistique textuelle

Le concept d'énergie (Clausius 1850) a précédé celui d'entropie (Boltzmann 1890) de quarante ans. Soixante ans plus tard, Shannon (1948, 1951) construisit la Théorie de l'Information, un formalisme entropique purement probabiliste, libre de toute considération énergétique. Cette théorie domine actuellement de nombreuses disciplines, dont le traitement statistique du langage, et, d'un certain point de vue la statistique tout court (Kullback 1959). Un exemple caractéristique en statistique textuelle est fourni par les travaux de Zipf (1949) et de Mandelbrot (1957) sur la Loi de Zipf, basés sur des considérations énergétiques ("principe de moindre effort"), et supplantés aujourd'hui pour l'essentiel par les résultats de Kraft, McMillan et Huffman (voir Cover et Thomas 1991) dans le cadre de la Théorie de l'Information.

Suivant une démarche proche de celle introduisant les modèles log-linéaires en statistique (voir par exemple Christensen (1990)), on définit suivant (2), l'énergie d'un état A de probabilité $P(A)$ par $U(A) := -\ln P(A)$. [1]

[1] l'énergie est en Physique une variable d'intervalle, c'est-à-dire définie à une transformation affine près

1.2.1. L' énergie de cohésion

L'énergie de cohésion $U_{\text{coh}}(A, B)$ entre deux états $A \in \mathcal{A}$ et $B \in \mathcal{B}$ est alors donnée par

$$U_{\text{coh}}(A, B) := U(A)+U(B)-U(A\text{et}B) = -\ln P(A)-\ln P(B)+\ln P(A\text{et}B) = \ln \frac{P(A \text{ et } B)}{P(A)P(B)} \tag{4}$$

La situation d'indépendance $P(A \text{ et } B) = P(A)P(B)$ équivaut donc à $U_{\text{coh}}(A, B) = 0$ (pas d'interaction); $P(A \text{ et } B) > P(A)P(B) \Leftrightarrow U_{\text{coh}}(A, B) > 0$ (attraction) et $P(A \text{ et } B) < P(A)P(B) \Leftrightarrow U_{\text{coh}}(A, B) < 0$ (répulsion) traduisent quant à elles les situations de dépendance. L'énergie de cohésion $U_{\text{coh}}(A, B) = \ln \frac{P(A \text{ et } B)}{P(A)P(B)}$ est appelée *information mutuelle ponctuelle* en Théorie de l'Information (Cover et Thomas, 1991). Lorsque A est un a-gramme et B un b-gramme qui le suit immédiatement, cette quantité s'utilise en traitement du langage naturel, typiquement comme un outil pour détecter les collocations (voir par exemple Church et Hanks 1989; Manning et Schütze 1999, pp. 178-183).

1.2.2. Le théorème de Hammersley-Clifford

Plus généralement, on considère des configurations $A = (a_1, \ldots, a_n)$ déterminées par les valeurs prises sur n variables $i = 1, \ldots, n$. Etant donnée une relation arbitraire de voisinage (symétrique) entre variables, le modèle $P(a_1, \ldots, a_n)$ sera dit markovien si $P(C|A \setminus C) = P(C|(A \setminus C) \cap N(C))$ pour tout $C \subset A$, où $N(C)$ est l'ensemble des variables voisines des variables de C; autrement dit, la probabilité conditionnelle d'une sous-configuration, étant donné son complémentaire, ne dépend que du voisinage de la sous-configuration. Alors, et pour autant qu'il n'y ait pas de configuration impossible, le théorème de Hammersley-Clifford affirme que le modèle est markovien si et seulement si $P(A)$ s'exprime par

$$P(A) = \exp\Big(\sum_{C \subset A \mid C \text{ est une clique}} U(C)\Big) \tag{5}$$

où la somme ne porte que sur les *cliques* de A, formées de groupes de variables toutes voisines deux à deux. Ainsi, le théorème de Hammersley-Clifford réintroduit de façon naturelle le concept de potentiel $U(C)$ et de distribution de Gibbs au-delà des contextes explicitement physiques de mécanique statistique; Besançon et al. (2001) en donnent une application à la désambiguïsation sémantique dans la représentation de textes.

En particulier, si aucune variable n'est voisine d'aucune autre, les cliques seront de la forme $C = \{a_i\}$, d'où indépendance:

$$P(a_1, \ldots, a_n) = P(A) = \exp(\sum_{i=1}^{n} U(a_i)) = \prod_{i=1}^{n} \exp(U(a_i)) = \prod_{i=1}^{n} P(a_i) \tag{6}$$

1.2.3. Le principe du maximum d'entropie.

Les distributions de Gibbs font également leur apparition dans le principe du maximum d'entropie (Jaynes (1978) en présente une synthèse dans une perspective historique), populaire en traitement statistique du langage (Manning et Schütze 1999): pour estimer une distribution $\{p(\omega)\}_{\omega \in \Omega}$

$\tilde{U}(A) = c\,U(A) + d$, où d détermine le zéro de l'énergie et c son unité. On peut alors fixer c et d de telle sorte que, pour le système examiné, considéré comme *système de référence*, l'on ait $\beta_{\text{ref}} = 1$ et $Z(\beta_{\text{ref}}) = 1$, d'où la définition précédente de $U(A)$. Dans cette normalisation, on a $F_{\text{min}} = 0$ et donc $u = s$.

dont la seule information connue est que $E_p(U) := \sum_{\omega\in\Omega} p(\omega)\, U(\omega) = a$, on maximise l'entropie $s[p]$ de la distribution sous la contrainte en question. La solution est alors $p^{\text{ME}}(\omega) = \exp(-\beta\, U(\omega))/Z(\beta)$, où β est fixé de sorte à satisfaire la contrainte $u(\beta) = a$. Cette contrainte est inactive si et seulement si a vaut exactement la moyenne de $U(\omega)$ sous la distribution uniforme p_{unif} (maximisant inconditionnellement l'entropie), i.e. si $u(0) = \frac{1}{|\Omega|}\sum_{\omega\in\Omega} U(\omega) = a$, d'où $\beta = 0$. Pour tous les autres cas, la contrainte est active et $\beta \neq 0$.

Ainsi, β apparaît ici comme mesure de l'*activité de la contrainte* $E_p(U) = a$. La divergence de Kullback-Leibler (entropie relative) entre p^{ME} et p_{unif} est en effet (avec (3))

$$K(p^{\text{ME}}||p_{\text{unif}}) := \sum_{\omega\in\Omega} p^{\text{ME}}(\omega) \ln \frac{p^{\text{ME}}(\omega)}{p_{\text{unif}}(\omega)} = -\beta\, u(\beta) - \ln Z(\beta) = \frac{1}{2}\, c(0)\, \beta^2 + 0(\beta^3) \qquad (7)$$

Lorsque Ω représente les m termes les plus courants d'un lexique indicé par fréquence décroissante (i.e. $p_1 \geq p_2 \geq \ldots p_m$), on retrouve la distribution de Zipf sous la contrainte $\sum_j p_j\, U_j = a = u(\beta)$, où $U_j := \ln j$ (la contrainte équivaut à fixer la moyenne géométrique pondérée des rangs). En effet, on obtient alors

$$p_j^{\text{ME}} = \frac{j^{-\beta}}{Z(\beta)} \qquad\qquad Z(\beta) = \sum_{k=1}^{m} k^{-\beta} \qquad (8)$$

La température inverse β joue ici le rôle d'un exposant de Pareto. En abaissant la température, le rang moyen logarithmique decroît de $u(0) = 1/m \sum_{j=1}^{m} \ln j \cong \ln m - 1$ à $u(\infty) = 0$. Rappelons (Mandelbrot 1957) que les estimations textuelles fournissent des valeurs de β légèrement supérieures à 1.

2. Illustrations

2.1. Chauffer et refroidir des textes

Il est possible, en utilisant (2), de construire des probabilités modifiées $p(A, \beta)$ dépendant de la température $T = 1/\beta$ comme $p(A, \beta) \cong p^{\beta}(A)$ (convenablement normalisé) où $p(A)$ est la distribution de référence à $\beta = 1$.

Concrètement, on considère un modèle textuel markovien d'ordre r défini par $p(\omega|\varphi) \geq 0$ où $\omega \in \Omega$ (l'ensemble des symboles retenus) et $\varphi \in \Omega^r$ (l'ensemble des r-grammes) tel que $\sum_{\omega\in\Omega} p(\omega|\varphi) = 1$. Le modèle de température inverse β d'ordre r est alors défini par les transitions $p_\beta(\omega|\varphi) := p^{\beta}(\omega|\varphi)/\sum_{\omega'\in\Omega} p^{\beta}(\omega'|\varphi)$.

Considérons l'échantillon textuel donné par le roman complet *Emma* d'Austen (1816). Sans distinction majuscule / minuscule, sans séparateurs et sans ponctuation (à l'exception du blanc, du trait d'union et de l'apostrophe), on obtient un texte de $n = 868'945$ symboles contenant $|\Omega| = 29$ types différents. En estimant les probabilités de transition par leur fréquence empirique[2], on obtient un texte d'ordre 3 simulé (à température inverse de référence $\beta_{\text{ref}} = 1$) de la forme:

```
feeliciousnest miss abbon hear jane is arer that isapple did
ther by the withour our the subject relevery that amile sament
is laugh in ' emma rement on the come februptings he some thed
```

[2]les calculs ont été effectués par le freeware Entropizer (Xanthos 2000), disponible à `http://www.unil.ch/ling/`

Pour $\beta = 0.1$, i.e. pour une température 10 fois plus grande que celle du texte de référence:

```
torables - hantly elterdays doin said just don't check comedina
inglas ratefusandinite his happerall bet had had habiticents'
oh young most brothey lostled wife favoicel let you cology
```

Dans la limite des hautes températures, le processus devient maximalement aléatoire: $p_0(\omega|\varphi)$ produit uniformément n'importe quel symbole $\omega \in \Omega$ ayant suivi φ au moins une fois dans l'échantillon d'apprentissage. Comme la grande majorité des transitions possibles d'ordre 3 n'est pas observée sur le corpus de référence, l'origine anglaise de ce dernier est clairement reconnaissable, malgré la température élevée du texte simulé ($\beta = 0.01$):

```
et-chaist-temseliving dwelf-ash eignansgranquick-gatefullied
georgo namissedeed fessnee th thusestnessful-timencurves - him
duraguesdaird vulgentroneousedatied yelaps isagacity in quainf
```

Refroidir le texte rend les transitions fréquentes encore plus fréquentes, et les transitions rares encore plus rares. En conséquence, le texte qui en résulte, deux fois plus froid ($\beta = 2$) que le texte de référence, devient passablement prédictible:

```
's good of his compassure is a miss she was she come to the of
his and as it it was so look of it i do not you with her that i
am superior the in ther which of that the half - and the man the
```

Dans la limite des basses températures $\beta \to \infty$, le processus devient déterministe: $p_\infty(\omega|\varphi) = 0$ pour tout $\omega \in \Omega$ à l'exception de $\omega_0(\varphi)$, défini comme le successeur le plus probable de φ. Le texte résultant (ici obtenu à $\beta = 4$) est périodique, proprement *cristallin* d'un point de vue physique:

```
ll the was the was the was the was the was the was the was the
was the was the was the was the was the was the was the was the
was the was the was the was the was the was the was the was the
```

2.2. L'énergie de cohésion comme critère de segmentation

On cherche ici à retrouver les frontières de mots d'un texte, préalablement effacées, sans disposer d'un lexique[3]. L'une des premières méthodes proposées à cet effet est celle du nombre de successeurs (Harris 1955, 1967), qui consiste à segmenter les énoncés aux points où la variété conditionnelle $V(\varphi)$ (voir 2.5) des phonèmes susceptibles de suivre le phonème φ est maximale. Cette approche s'est avérée particulièrement fertile en développements divers, en particulier dans son extension à un formalisme markovien, en substituant au nombre de successeurs $V(\varphi)$ l'entropie conditionnelle d'ordre r, définie par $H^r(\varphi) := -\sum_{\psi\in\Omega^r} P(\psi|\varphi)\ \ln P(\psi|\varphi)$ (Gammon 1969, Hutchens et Alder 1998).

Plutôt que de fonder la segmentation sur un indicateur de diversité (voir 2.5), on peut envisager d'utiliser un critère énergétique inspiré de l'énergie de cohésion, c'est-à-dire faire l'hypothèse

[3]La seule observation des marques typographiques (espaces, apostrophes, tirets, ponctuation, etc.) est notoirement insuffisante, sans parler des langues n'utilisant pas de séparateurs explicites (Chinois et Japonais notamment) ou du cas des transcriptions phonétiques.

que l'énergie de cohésion moyenne

$$U^r_{\text{coh}}(\varphi) := \sum_{\psi \in \Omega^r} P(\psi|\varphi)\, U_{\text{int}}(\varphi, \psi) = \sum_{\psi \in \Omega^r} P(\psi|\varphi)\, \ln \frac{P(\psi|\varphi)}{P(\psi)} \tag{9}$$

entre la séquence φ et ses successeurs possibles de taille r est généralement plus faible lorsque φ est une fin d'unité typique. En pratique, on fixe un seuil s (dont la valeur optimale dépend des propriétés distributionnelles du texte en question) en deçà duquel $U^r_{\text{coh}}(\varphi)$ est tenu pour significativement faible, auquel cas l'on insère un séparateur. Par exemple, pour le corpus sans blancs du début des *Métamorphoses* d'Ovide (comprenant $n = 352'419$ symboles à $|\Omega| = 24$ types), en prenant $\varphi \in \Omega^3$ et $r = 1$ (modèle de Markov d'ordre 3) et $s = .95$, on obtient la segmentation suivante:

```
(in n)ova fer_t=anim_us muta_t_a_s dice_re forma_s corpora di co-
eptis nam vos muta_s_tis et illa_s a_d_spira_t_e=me_is prima_que
ab=or_igine m_undi=a_d=me_a=per_petu_um=de_ducit_e=t_empora ca_rmen
a_nt_e=mare et ter_ra_s et quod tegit omnia=c_a_elum unus erat
```

où les blancs dénotent des espaces correctement inférés, et symboles '=' et '_' dénotent respectivement les espaces manqués et les fausses alarmes[4]. La proportion de manqués et celle de fausses alarmes valent toutes deux 23.6%, et ces scores peuvent être encore réduits en utilisant une moyenne pondérée des énergies de cohésion "vers la droite" et "vers la gauche", i.e. en prenant la variable de décision $\delta\, U^r_{\text{coh}}(\varphi) + (1 - \delta)\, \check{U}^r_{\text{coh}}(\varphi')$, avec $0 \leq \delta \leq 1$ et où $\check{U}^r_{\text{coh}}(\varphi')$ se définit comme $U^r_{\text{coh}}(\varphi)$ *après avoir inversé l'ordre du texte*. Dans les mêmes conditions que ci-dessus, et avec $\delta = .66$, la proportion de manqués tombe à 20.4% et celle de fausses alarmes à 20.9%:

```
(in n)ova fer_t anim_us muta_t_a_s dic_e_re forma_s corpora di
coeptis nam vos muta_s_ti_s et ill_a_s ad_spira_t_e=me_is
prima_que=ab origine m_undi=ad me_a per_p_etu_um de_ducit_e=temp_ora
car_men ant_e m_are et ter_ra_s et quod tegit omnia caelum=unus
```

Les courbes ROC ci-dessous permettent de comparer les performances de l'énergie de cohésion $U^r_{\text{coh}}(\varphi)$ avec celles de l'entropie conditionnelle $H^r(\varphi)$ pour $r = 1$, pour $\delta = 0.5$ et pour φ de taille 2, 3 et 4. Elles montrent que l'énergie est un meilleur critère que l'entropie pour les ordres faibles, tout en n'étant guère moins efficace pour les ordres élevés.

2.3. Estimer la température d'un texte

Soit un modèle markovien d'ordre r dont les transitions $p_{D_0}(\omega|\varphi)$ sont estimées sur un texte de référence D_0. Soit D_1 un nouveau texte dont les transitions empiriques $p_{D_1}(\omega|\varphi)$ sont conçues comme produites par $p_\beta(\omega|\varphi) \sim p^\beta_{D_0}(\omega|\varphi)$, i.e. par le modèle de référence à une température relative $T = 1/\beta$ éventuellement différente. Un tel texte pourrait être produit par un auteur sur-représentant ($\beta > 1$) les catégories les plus fréquentes dans le corpus de référence D_0 ou au contraire les sous-représentant ($\beta < 1$).

L'estimation de β par maximum de vraisemblance est passablement intriquée quoique possible. Sacrifiant le réalisme à la simplicité, on obtient pour un modèle d'ordre $r = 0$ (indépendance)

[4] Le premier type d'erreur est toujours induit par l'existence d'homonymes admettant une segmentation différente; le second est fréquemment explicable par la possible décomposition de certaines unités en morphèmes.

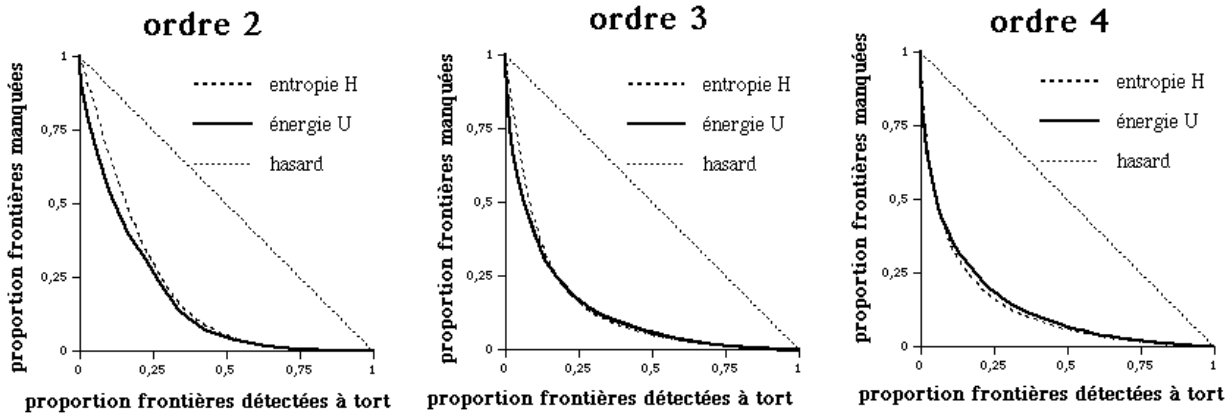

Figure 1: *Performances comparées de l'entropie conditionnelle $H^r(\varphi)$ et de l'énergie de cohésion $U^r_{coh}(\varphi)$ pour une tâche de segmentation*

l'approximation linéaire suivante:

$$\beta(D_1|D_0) \cong 1 + \frac{1}{c_{D_0}} \sum_{\omega\in\Omega} [p_{D_1}(\omega) - p_{D_0}(\omega)] \ln p_{D_0}(\omega) \tag{10}$$

avec

$$c_{D_0} = \text{Var}_{p_{D_0}}(U_{D_0}) \qquad\qquad U_{D_0}(\omega) = -\ln p_{D_0}(\omega) \tag{11}$$

Exemple: on considère les $n = 725'001$ premiers symboles des textes *Emma* d'Austen, déjà rencontré, ainsi que de *La bête humaine* de Zola, tous deux codés à 29 symboles. Leurs entropies et chaleurs spécifiques d'ordre 0 sont $s_{\text{Austen}} = 2.848$, $s_{\text{Zola}} = 2.787$, $c_{\text{Austen}} = 0.700$ et $c_{\text{Zola}} = 0.729$. (10) donne alors

$$\beta(\text{Zola}|\text{Austen}) \cong 1 + \frac{-0.080}{0.700} = 0.886 \qquad\qquad \beta(\text{Austen}|\text{Zola}) \cong 1 + \frac{-0.307}{0.729} = 0.579$$

Dans les deux cas, le nouveau texte est jugé *plus chaud* que le texte de référence ($1/0.886 = 1.13$, respectivement $1/0.579 = 1.73$ fois plus chaud): la répartition empirique des symboles dans Zola étant peu probable à l'aune du modèle estimé sur le corpus d'Austen (et vice-versa), il faut alors chauffer le texte de référence pour permettre à des événements rares d'apparaître plus souvent. Cette tendance à l'élévation systématique de température trahit ici la grande dissimilarité des distributions $p_{D_0}(\omega)$ et $p_{D_1}(\omega)$.

Une situation plus adaptée d'estimation de la température est celle d'un texte D_1 corrompu de façon uniforme, i.e. $p_{D_1}(\omega) = (1-\epsilon)\, p_{D_0}(\omega) + \epsilon\, p_{\text{unif}}(\omega)$, où $\epsilon \in [0,1]$ est une mesure de l'altération de D_1 et $p_{\text{unif}}(\omega) = 1/|\Omega|$. Comme il se doit, on a $T(\epsilon) \geq 1$ avec $T(0) = 1$:

$$\beta(D_1|D_0) \cong 1 + \frac{\epsilon}{c_{D_0}} \sum_{\omega\in\Omega} [p_{\text{unif}}(\omega) - p_{D_0}(\omega)] \ln p_{D_0}(\omega) \leq 1$$

2.4. Mélanges additifs et multiplicatifs de deux modèles

Etant donnés deux modèles de textes H_0 et H_1 d'ordre r, on peut définir un nouveau modèle H_λ de *mélange additif* ainsi qu'un modèle H_μ de *mélange multiplicatif* comme

$$p_\lambda(\omega|\varphi) := (1-\lambda)p_0(\omega|\varphi) + \lambda p_1(\omega|\varphi) \qquad p_\mu(\omega|\varphi) := \frac{p_0^{(1-\mu)}(\omega|\varphi)\, p_1^{\mu}(\omega|\varphi)}{\sum_{\omega'\in\Omega} p_0^{(1-\mu)}(\omega'|\varphi)\, p_1^{\mu}(\omega'|\varphi)} \tag{12}$$

avec $0 < \lambda, \mu < 1$: les probabilités sont moyennisées dans le mélange additif, tandis que ce sont les *énergies* qui le sont pour le mélange multiplicatif. En conséquence, il suffit qu'une transition soit possible dans l'un des deux modèles H_0 *ou* H_1 pour qu'elle le soit dans H_λ; en revanche, une transition possible sous H_μ doit l'être sous H_0 *et* H_1.

En prenant pour H_0 de l'anglais à 29 caractères estimé par le début du roman d'Austen, et pour H_1 du français estimé par le début du roman de Zola, on trouve pour les modèles additifs d'ordre 3, pour $\lambda = 0.17$, $\lambda = 0.5$ et $\lambda = 0.83$ respectivement:

```
ll thin not alarly but alabouthould only to comethey had be the
sepant a was que lify you i bed at it see othe to had state cetter
but of i she done a la veil la preckone forma feel inute and it
```

```
daband shous ne findissouservait de sais comment do be certant she
cette l'ideed se point le fair somethen l'autres jeune suit onze
muchait satite a ponded was si je lui love toura la
```

```
les appelleur voice the toodhould son as or que aprennel un
revincontait en at on du semblait juge yeux plait etait resoinsit-
tairl on in and my she comme elle ecreta-t-il avait autes foiser
```

Comme on s'y attendait, la ressemblance avec le français augmente avec λ. Le même phénomène se produit à μ croissant pour les mélanges multiplicatifs, à la différence que la simulation d'un texte produit par H_μ se bloque (ce qui est indiqué par "***") dès qu'apparaît un trigramme φ n'ayant pas de continuation commune possible en anglais *et* en français. On obtient respectivement pour $\mu = 0.17$, $\mu = 0.5$ et $\mu = 0.83$:

```
licatellence a promine agement ano ton becol car emm*** ever an-
s touche-***i harriager gonistain ans tole elegards intellan enour
bellion genea***he succept wa***n instand instilliaristinutes
```

```
n neignit innerable quit tole ballassure cause on an une grite
chambe ner martient infine disable prisages creat mellesselles
dut***grange accour les norance trop mise une les emm*** mand
```

```
es terine fille son mainternistonsidenter ing sile celles tout
a pard elevant poingerent une graver dant lesses jam***core son
luxu***que eles visagemensation lame cendance materroga***e
```

On observe que les mélanges multiplicatifs produisent un certain nombre de formes à consonnances latines, qui constituent justement une portion considérable de ce que les lexiques français et anglais ont en commun.

Le mélange multiplicatif jouit d'une vertu inférentielle particulière: dans le test de maximum de vraisemblance de H_0 contre H_1, l'erreur de première espèce (respectivement de seconde espèce)

décroît exponentiellement avec un exposant qui n'est autre que l'entropie relative entre p_μ et p_0 (respectivement entre p_μ et p_1), où la valeur de μ est fixée par le seuil de décision adopté (Cover et Thomas, 1991, pp. 312-314). De ce point de vue, p_μ constitue le modèle intermédiaire entre p_0 et p_1 permettant une discrimination optimale de ces derniers.

2.5. Température, indices de richesse du vocabulaire, et entropie de Rényi

La recherche d'une bonne mesure de richesse lexicale est un thème récurrent en statistique textuelle. On peut y distinguer des indices "qualitatifs" (comptant le nombre de formes distinctes) des indices "quantitatifs" (tenant également compte des fréquences de ces dernières); en travaillant au niveau des r-grammes $\varphi \in \Omega^r$, on peut citer la *variété* $V := |\{\varphi \in \Omega^r \,|\, P(\varphi) > 0\}|$ et l'entropie de Shannon $s := -\sum_{\varphi\in\Omega^r} P(\varphi)\,\ln P(\varphi)$ comme exemples typiques. Ces deux indices sont des cas particuliers de la famille des *entropies de Rényi*

$$R_\beta := \frac{1}{1-\beta}\ln \sum_{\varphi\in\Omega^r} P^\beta(\varphi) = \frac{1}{1-\beta}\ln \sum_{\varphi\in\Omega^r} \exp(-\beta\, U(\varphi)) = \frac{1}{1-\beta}\ln Z(\beta) \qquad (13)$$

pour laquelle on a les limites

$$\lim_{\beta\to 0} R_\beta = \ln V \qquad \lim_{\beta\to 1} R_\beta = s \qquad \lim_{\beta\to\infty} R_\beta = -\ln P(\varphi_0) = U(\varphi_0) \qquad (14)$$

où φ_0 est le r-gramme le plus fréquent, i.e. l'état fondamental du système dans un formalisme énergétique. On peut montrer que R_β est décroissant en β (la richesse d'un système augmente avec sa température). L'indice S de diversité de Simpson s'obtient comme $S = \exp(-R_a(2)) - 1/n$, et la caractéristique K de Yule (Yule, 1944) comme $K = 10'000\,\frac{n-1}{n}S$.

L'ensemble des symboles distincts Ω (on inclut ici la possibilité que lesdits symboles soient eux-même constitués de r-grammes relativement à des sous-symboles "élémentaires") peut être agrandi à un ensemble plus étendu Ω', en distinguant dans ce dernier des symboles jusqu'alors identifiés dans Ω. On dit que Ω est plus grossier que Ω' (ou que Ω' est plus fin que Ω), ce que l'on note $\Omega \leq \Omega'$. En procédant par induction et en considérant l'agrégation (= l'identification) de deux symboles, on montre que

$$\Omega \leq \Omega' \qquad \Rightarrow \qquad R_\beta(\Omega) \leq R_\beta(\Omega') \qquad (15)$$

Plus fine est la partition choisie, plus grande est la valeur de R_β, ainsi qu'il convient à un indice de richesse textuelle. Dans la limite de la partition triviale (i.e. identifiant tous les symboles) on a $R_\beta \equiv 0$ quel que soit $\beta \geq 0$.

3. Conclusion

Le non-spécialiste peut éprouver quelques difficultés initiales face à l'abstraction formelle de la Thermodynamique et de la Théorie de l'Information; il est clair, cependant, que ce même non-spécialiste possède une compréhension très intuitive des mécanismes que la première approche permet de décrire fort efficacement. En modélisant la dépendance entre des symboles successifs en termes de cohésion plutôt que d'information, ou en liant la diversité des transitions observables dans un texte au concept de température plutôt qu'à celui d'entropie, nous espérons avoir montré que les objets de la statistique textuelle peuvent bénéficier d'un éclairage pertinent lorsqu'on les examine à la lumière de phénomènes dont chacun peut faire quotidiennement l'expérience. De plus, l'équivalence des deux formalismes assure que tout développement issu

d'une approche thermodynamique trouvera son expression en Théorie de l'Information; une voie possible et de portée générale en modélisation textuelle pourrait ainsi se formuler comme suit: "intuition de base -> thermodynamique intuitive -> thermodynamique formelle -> Théorie de l'Information".

Références

Besag, J. (1974). "Spatial interaction and the statistical analysis of lattice systems", *Journal of the Royal Statistics Society* 36 pp. 192-236.

Besançon, R., Rozenknop, A. Chappelier, J.-C. et Rajman, M. (2001). "Intégration probabiliste de sens dans la représentation de textes", *Proceedings of TALN 2001*.

Christensen, R. (1990). *Log-Linear Models*. Springer, New York.

Church, K.W. and Hanks, P. (1989). *Word association norms, mutual information and lexicography*, *ACL* 27 pp. 76-83.

Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory*. Wiley, New York.

Gammon, E. (1969). "Quantitative approximations to the word", in *Papers presented to the International Conference on Computational Linguistics COLING-69.*

Harris, Z.S. (1955). "From phoneme to morpheme", *Language* 31, pp. 190-222, réimprimé dans Harris, Z.S. (1970), *Papers in Structural and Transformational Linguistics*, Dordrecht, D.Reidel, pp. 32-67.

Harris, Z.S. (1967). "Morpheme Boundaries within Words: Report on a Computer Test", *Transformations and Discourse Analysis Papers* 31, réimprimé dans Harris Z.S. (1970), *Papers in Structural and Transformational Linguistics*, Dordrecht, D.Reidel, pp. 68-87.

Hutchens, J.L. et Alder, M.D. (1998). "Finding Structure via Compression", *Proceedings of the International Conference on Computational Natural Language Learning*.

Jaynes, E.T. (1978). *Where do we stand on maximum entropy ?*, presented at the Maximum Entropy Formalism Conference, MIT, Cambridge.

Kullback, S. (1959). *Information Theory and Statistics*, Wiley, New York.

Mandelbrot, B. (1957). "Linguistique Statistique Macroscopique". In Apostel, L., Mandelbrot, B. et Morf, A. *Logique, Langage et Théorie de l'Information*, pp. 1-78. Presses Universitaires de France, Paris.

Manning, C.D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT-Press, Cambridge.

Rose, K. (1998). "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems", *Proceedings of the IEEE* 86, pp. 2210-2239.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Tech. Journal* 27, pp. 379-423; 623-656.

Shannon, C.E. (1951). Prediction and entropy of printed English. *Bell Sys.Tech. Journal* 30, pp. 50-64.

Xanthos, A. (2000). Entropizer 1.1: un outil informatique pour l'analyse séquentielle. *Proceedings of the 5th International Conference on the Statistical Analysis of Textual Data (JADT 2000).*

Yule, G.U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Hafner Publishing Company, New York.