Cahiers de l'ILSL 56, 2018, 11-30

Combien y a-t-il de variétés distinctes d'anglais?

François BAVAUD
Université de Lausanne

« La réponse … est 42 »

Douglas Adams
The Hitchhiker's Guide to the Galaxy

1. PROLOGUE : LE PLUS BEL ENDROIT SUR TERRE

Davantage que son cadre merveilleux, entre Chamberonne, Sorge et Léman, où conversent en hexamètres dactyliques verts damoiseaux, gris érudits et Roux du Valais, davantage encore que la fulgurance de ses chercheuses, l'éminence de ses enseignantes, la clairvoyance de ses étudiantes, c'est avant tout la *variété de ses études* et la *diversité de ses recherches* qui confèrent à notre Faculté son éclat à nul autre pareil.

J'ai beau m'en moquer gentiment, j'y souscris pleinement : la *liberté* de l'étudiant et du chercheur, l'*égalité* de considération de tous envers tous, la *fraternité* de ses membres unis dans la défense du grec ancien, de l'être de l'étant, de la déconstruction, ou de l'informatique et des méthodes mathématiques pour les sciences humaines[1], tout ceci fait de cet endroit le plus beau lieu du monde, l'Académie rêvée, le Lycée absolu, le Thélème zénithal.

[1] ou pas

La variété des études de propédeutique y est de 13'300 [2]. La diversité de ses recherches est infinie : on peut par exemple vouloir mesurer la diversité[3], ou recenser la diversité des mesures de diversité, ou étudier la diversité des recensions de la diversité des mesures de diversité, etc.

J'adore cette République des Lettres qui permet ce type de questionnement récursif. Et Marianne, sa figure tutélaire. Et aussi les notes de bas de page[4].

2. Notions et mesures de diversité

Au fur et à mesure que se déploie la parole ou que s'écrit le texte, certaines formes se répètent. Le nombre de *types* (formes) distinctes est inférieur à celui des *tokens* (occurrences), et la quantification de ces questions (typiquement : y a-t-il une « famille universelle » de fonctions décrivant la décroissance du rapport types/tokens en fonction du nombre de tokens) a occupé d'éminents linguistes quantitativistes depuis au moins un siècle, produisant de remarquables résultats ci et là. A l'aune des espoirs initiaux et du travail investi, le bilan global actuel reste toutefois questionnable; en particulier, la « famille universelle » se fait toujours attendre…

Les mêmes intérêts, ressources et difficultés se retrouvent en *écologie*, comme a pu le souligner par exemple Jarvis (2013) : au fur et à mesure d'une promenade maraîchère se dévoilent des variétés de salade : romaine, roquette, frisée, scarole, rampon, doucette, mâche, et autres *espèces inconnues* qu'une exploration plus poussée pourra peut-être révéler. A quoi doit s'attendre notre promeneuse? Comment va-t-elle mesurer la biodiversité? Une parcelle contenant

---

[2] Avec vingt branches internes et dix-sept branches externes non cumulables, l'étudiant·e de première année doit confronter $20 \cdot 19 \cdot 18 + 17 \cdot 20 \cdot 19 = 13'300$ choix possibles.

[3] C'est le sujet de ce travail.

[4] J'adorerais aussi, une fois dans mon existence, écrire[a]

[a] des notes de bas de page de notes de bas de page.

nonante plants de roquette et dix de scarole est-elle plus diverse qu'une parcelle contenant vingt plants de roquette, vingt de scarole et vingt de rampon – ou bien pas?

2.1. Fréquences et similarités entre types

Vouloir comparer au moyen d'un seul indicateur deux instances différant à la fois par leur taille (le nombre de tokens) et par la distribution relative des types observés (la proportion de tokens associés à chaque type) est mission presque impossible : deux indicateurs sont requis pour quantifier ces deux aspects distincts[5]. Seule la caractérisation en terme de *profil relatif* est abordée ici, sans tenir compte de l'étendue de l'objet.

Concrètement, il s'agit de mesurer la diversité d'un phénomène possédant q types, dont le profil est donné par les proportions relatives notées $p = (p_1, \ldots, p_q)$, avec $p_a \geq 0$ pour tout a, et $\sum_{a=1}^{q} p_a = 1$. Le nombre de catégories qui apparaissent effectivement, *i.e.* pour lesquelles $p_a > 0$, est une mesure évidente de la variété de l'objet, parfois simplement appelé *variété* tout court, et défini par $v(p) = \sum_{a=1}^{q} I\,(p_a > 0)$, où la fonction indicatrice $I(A)$ prend la valeur 1 si l'événement A est vrai, et 0 sinon. Par construction, $v(p) \leq q$ et la différence $q - v(p)$ constitue le nombre d'espèces encore inobservées ou *inconnues*, un thème fascinant en linguistique ou biologie, qui ne sera pas davantage développé ici. Cela étant, $v(p)$ souffre du défaut majeur de ne pas tenir compte des proportions en jeu : qu'un type de salade apparaisse une fois ou mille fois aura exactement le même effet, à savoir celui d'augmenter la valeur de $v(p)$ d'une unité.

Parmi les nombreuses mesures de variété proposées qui prennent en considération les proportions p, l'*entropie de Shannon*

$$\mathcal{H}(p) = -\sum_{a=1}^{q} p_a \ln p_a \qquad (1)$$

[5] même si de récents et notables progrès ont permis de mieux comparer les variétés entre échantillons de taille différentes (voir par exemple Xanthos & Gillis 2010, ou Xanthos dans ce volume).

est incontestablement la meilleure d'un point de vue mathématico-formel (justifier cela demanderait plusieurs pages d'exposition en Théorie de l'Information, dont (1) constitue la première pierre). D'intéressantes considérations additionnelles ont pu suggérer que, au sein de problématiques bien spécifiques (« loi » de Zipf, allométrie, « power laws », systèmes « non-extensifs », fractales) il pourrait être avantageux de généraliser (1) à des familles permettant de sur- ou de sous-pondérer les proportions grandes ou petites, comme dans les familles d'entropie de *Rényi* ou de *Tsallis*, dont $\mathcal{H}(p)$ et $v(p)$ sont des cas particuliers.

Cette contribution concerne une autre généralisation de (1), appelée « entropie effective », et *motivée par le fait que les q catégories en jeu ne sont en général pas entièrement distinctes* : scarole et frisée sont proches, parfois confondues. Rampon, doucette et mâche sont identiques - la dénomination varie simplement selon les aires linguistiques. Ces *similarités* doivent, à l'évidence, diminuer la valeur de la mesure canonique (1) de variété – mais de combien? Voilà tout l'enjeu.

2.2. Un peu de formalisme

La nouvelle mesure de variété proposée, l'*entropie effective* E définie en (2) (p. 13), peut s'appliquer à toute situation impliquant q types, de poids relatifs $p_a > 0$ avec $\sum_{a=1}^{q} p_a = 1$, mesurant la fréquence ou l'importance relative de chaque type. Il faut également disposer d'une quantification des différences entre types sous la forme d'une matrice $q \times q$ de dissimilarité $D = (d_{ab})$ supposée *propre*, c'est-à-dire satisfaisant à :

$$d_{ab} = d_{ba} \qquad d_{aa} = 0 \qquad d_{ab} > 0 \quad \text{si } a \neq b.$$

L'idée qui sous-tend la définition de l'entropie effective est que, si les types ou catégories a et b sont suffisamment proches ou peu

fréquents, ils pourront être confondus[6], tandis qu'ils seront distingués s'ils sont suffisamment distincts ou fréquents, contribuant ainsi à accroître la diversité du phénomène étudié. Plus précisément, on peut concevoir les q types comme autant de *stimuli*, lesquels seront identifiés, classifiés sous la forme de *percepts* étiquetés selon les mêmes q types, selon un schéma d'attribution décrit par la *matrice* $q \times q$ *de confusion normalisée* $Z = (z_{ab})$, décrivant la probabilité que le stimulus a soit perçu comme b. Par construction, $z_{ab} \geq 0$, et $\sum_{b=1}^{q} z_{ab} = 1$. Aussi, la probabilité ou poids relatif du percept b s'obtient comme $r_b = \sum_a p_a \, z_{ab}$, et satisfait à $r_b \geq 0$ et $\sum_b r_b = 1$.

L'ensemble de tous les schémas d'attribution possibles sera dénoté par $\mathcal{Z}$, au sein duquel figurera l'*identification parfaite* ($z_{ab} = 1$ si $a = b$, et $z_{ab} = 0$ si $a \neq b$, qui est la matrice identité) décrivant l'identification sans faute de chaque stimulus au bon percept, ainsi que l'*identification aléatoire* qui détermine le percept indépendamment du stimulus ($z_{ab} = r_b \geq 0$ avec $\sum_b r_b = 1$).

L'identification parfaite se trouve minimiser la *distorsion* donnée par $U[Z] = \sum_{a,b} p_a \, z_{ab} d_{ab}$, qui mesure l'erreur moyenne totale dans la reconnaissance des stimuli.

L'autre ingrédient de base est l'*information mutuelle* $K[Z] = \sum_{a,b} p_a \, z_{ab} \ln (z_{ab}/r_b)$ de Shannon. L'information mutuelle $K[Z]$ (dont le *chi-carré* des cours de statistique constitue l'approximation quadratique) mesure la dépendance entre le stimulus et le percept, et son minimum de zéro est précisément atteint pour l'identification parfaite décrite ci-dessus.

Un compromis entre la situation de perception parfaite (minimisation de $U[Z]$) et la situation de bruit intégral ruinant toute possibilité de discrimination entre stimuli (minimisation de $K[Z]$) est fourni par la minimisation de l'*énergie libre* $F[Z]$, dont le minimum E constitue justement cette nouvelle mesure de variété, l'*entropie effective*. Formellement,

---

[6] Un exemple entre mille : Gynura Procumbens, rare dans nos contrées, peut être facilement pris pour un épinard, auquel il ressemble.

$$F[Z] = \beta U[Z] + K[Z] \qquad E = \min_{Z \in \mathcal{Z}} F[Z] \qquad (2)$$

où $\beta > 0$ est un paramètre libre, bien connu des thermodynamiciens et physiciens statisticiens sous le nom de *température inverse*, lequel joue le rôle d'arbitre du compromis en question. La température $1/\beta$ contrôle ici l'imperfection du système perceptif : plus elle est grande, plus importante sera la distorsion.

La minimisation de (2) débouche sur une équation non-linéaire pour l'identification optimale Z, à savoir :

$$z_{ab} = \frac{r_b[Z]\, s_{ab}}{\sum_{c=1}^{q} r_c[Z]\, s_{ac}} \qquad (3)$$

où $s_{ab} = \exp(-\beta\, d_{ab})$ et $r_b[Z] = \sum_a p_a\, z_{ab}$ est le poids du percept. L'équation (3) possède une solution unique, qui peut être obtenue de façon itérative, en définissant le nouveau Z comme membre de gauche de la première identité, dont le membre de droite est calculé à partir de l'ancien Z, jusqu'à convergence, à partir d'une identification Z initiale quelconque. La matrice $q \times q$ de composantes $S = (s_{ab})$ définies plus haut est une matrice de *similarité entre types*, satisfaisant à :

$$s_{ab} = s_{ba} \qquad s_{aa} = 1 \qquad 0 \leq s_{ab} < 1 \quad \text{si } a \neq b.$$

L'analyse du problème (2) et de l'algorithme itératif issu de (3) permet de démontrer l'existence de deux températures inverses critiques β_H et β_L, telles que $0 < \beta_H < \beta_L < \infty$, pour lesquelles :

— tous les stimuli en jeu sont reconnus, du moins partiellement (*i.e.* $r_b > 0$ pour tout b), dans le *régime de basse température* $\beta > \beta_L$

— tous les stimuli sont attribués au même *percept prototypique* $b^\star$, solution (qu'on admettra unique; c'est le cas générique) de $b^\star = \arg\min_b \sum_a p_a\, d_{ab}$, dans le *régime de haute température* $\beta < \beta_H$; en particulier, $b^\star$ est le type le plus proche du centre de gravité de la configuration (p, D)

lorsque D est le carré d'une distance euclidienne, comme en Analyse classique de Données

— le *régime tempéré* $\beta_H \leq \beta \leq \beta_L$ contient, à mesure que la température augmente (et que la température inverse β diminue) une série de $q-1$ *transitions de phase*, passant de q types identifiés pour $\beta = \beta_L$ à un seul type identifié pour $\beta = \beta_H$, avec extinction cumulée des types, à commencer par les « moins résistants » – parce que trop rares, trop excentrés dans la configuration (p, D), ou trop proches d'un type plus résistant; la recherche d'un critère analytique unique est en cours

— naturellement, l'entropie effective $E(\beta)$ croît avec β, avec $\lim_{\beta\to 0} E(\beta) = 0$ (un seul type perçu : « la nuit, tous les chats sont gris »), et $\lim_{\beta\to\infty} E(\beta) = \mathcal{H}$: on retrouve l'entropie classique de Shannon dans le cas particulier de types entièrement distincts. De façon moins attendue,

$$E(\beta) \leq R(\beta)$$

où $R(\beta) = -\sum_{a=1}^{q} p_a \ln\ b_a\,(\beta)$ et $b_a(\beta) = \sum_{b=1}^{q} s_{ab}\ p_b$.

La *diversité réduite* R, plus facile à calculer que E (pas besoin d'itérations), est moins satisfaisante d'un point de vue formel (indépendance et décomposition « intra+inter », sans pouvoir développer davantage ici). La quantité $b_a(\beta)$, qui mesure la similarité attendue entre un individu de type a et un autre individu tiré au hasard (Leinster & Cobbold 2012) est appelée « banalité » dans la littérature écologique (Marcon 2015). Elle va de $\lim_{\beta\to 0} b_a(\beta) = 1$ (types confondus dans un seul prototype) à $\lim_{\beta\to\infty} b_a(\beta) = p_a$ (types entièrement distincts).

2.3. Isomorphisme avec le transport optimal

La lectrice versée en recherche opérationnelle n'aura pas manqué de relever l'identité entre la distorsion $U[Z]$ et *l'énergie de transport*, mesurant la distance physique totale de déplacement des q types,

depuis leur origine a vers des destinations b fixes mais librement choisies ici. Lorsque les capacités des destinations $r_b = \sum_a p_a z_{ab}$ sont fixées, le problème de la minimisation de $U[Z]$, dit du *transport optimal*, et auquel les noms de Monge, Kantorovitch, Villani et bien d'autres sont associés, devient nettement plus difficile. Le problème dit *relaxé* (2), incorporant l'entropie mutuelle et favorisant les trajets entre origines et destinations distinctes, est paradoxalement plus simple à résoudre, et permet de modéliser les *flux* décrivant les mouvements des unités (marchandises, personnes, information), tels que les transports internationaux, les migrations humaines ou l'attribution des écoliers lausannois à leur établissement scolaire (Emmanouilidis, Guex & Bavaud 2016).

3. VARIÉTÉS MORPHO-SYNTACTIQUES DE L'ANGLAIS

3.1. Les données du projet eWAVE

Le projet *eWAVE* (*The Electronic World Atlas of Varieties of English*) (Kortmann & Lunkenheimer 2013) contient une base de données de $p = 235$ caractéristiques morpho-syntaxiques[7] de $q = 76$ variétés d'anglais, décrites dans Schneider et Kortmann (2004). Chaque caractéristique y est catégorisée comme « A = obligatoire ou omniprésente », « B = ni très rare, ni omniprésente », « C = très rare », « D = absence attestée », « X = caractéristique non pertinente pour des raisons structurelles » ou enfin comme « ? = pas d'information ». Ces catégories sont ensuite converties en scores d'apparition selon la proposition (au caractère heuristique et un peu arbitraire assumé) du projet eWAVE, à savoir comme $A \to 1$, $B \to 0{,}6$, $C \to 0{,}3$ et $D \to 0$. Quant à X et ?, ils ont été remplacés ici, pour chaque caractéristique, par la moyenne des valeurs A, B, C et D apparaissant dans ladite caractéristique.

[7] telles que « *No gender distinction in third person singular* », « Ain't *as the negated form of* have », « *Deletion of* to *before infinitives* », etc.

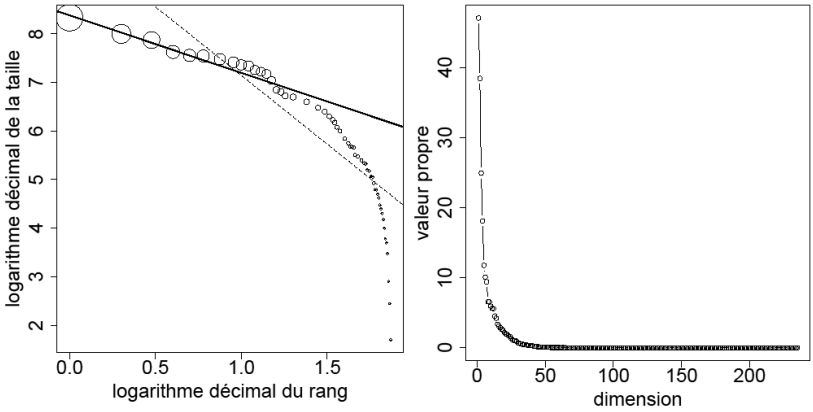

Figure 1 : Gauche : relation logarithmique rang-taille ou « loi » de Zipf; la droite en pointillé est la droite de régression uniforme du logarithme de la taille par celui du rang (pente = –2,8), et la droite en continu celle de la même régression, mais pondérée par la taille (pente = –1,2). Droite : « graphe des éboulis » donnant les valeurs propres de la matrice des corrélations $r_{kl} = corr(x_k, x_l)$ *entre les 235 caractéristiques : 4 facteurs expriment 55 % de la variabilité de ces dernières; voir aussi la figure 3 (p. 22, droite, haut).*

En résumé, on obtient une matrice $X = (x_{ak})$, de taille 76×235, donnant pour chaque variété a le score d'apparition $x_{ak} \in [0,1]$ de chaque caractéristique k, à partir de laquelle les dissimilarités d_{ab} seront construites dans la section suivante.

Une estimation du nombre de locuteurs de chacune des variétés, permettant d'en calculer les fréquences ou proportions p_a, m'a été communiquée durant l'été 2017 par Bernd Kortmann, de l'Englisches Seminar de l'Université de Fribourg-en-Brisgau, ce dont je lui suis très reconnaissant. Ni la fréquence du *Earlier African-American Vernacular English* (disparu de nos jours) ni celle du *Butler English* (son existence en tant que variété suffisamment homogène, distincte d'un simple anglais rudimentaire, est sujette à caution), n'ont pu être estimées. La relation « rang-taille » des effectifs des $q = 74$ variétés restantes, totalisant 780 millions de locuteurs, est représentée en

figure 1. Les variétés les plus fréquentes sont le *Colloquial American English*, suivi de l'*Indian English* et du *Nigerian Pidgin*.

3.2. La construction des dissimilarités entre variétés

La matrice $q \times p = 74 \times 235$ des profils numériques $X = (x_{ak})$ des variétés permet de calculer des mesures de dissimilarité entre ces dernières – une opération à la fois standard et cruciale en Analyse de Données. Trois mesures sont considérées, à savoir :

— dissimilarité brute ou de covariance

$$d_{ab}^{\mathrm{cov}} = \sum_{k=1}^{p} (x_{ak} - x_{bk})^2$$

— dissimilarité standardisée ou de corrélation

$$d_{ab}^{\mathrm{corr}} = \sum_{k=1}^{p} \frac{(x_{ak} - x_{bk})^2}{\sigma_{kk}}$$

— dissimilarité de Mahalanobis

$$d_{ab}^{\mathrm{Maha}} = \sum_{k,l=1}^{p} (x_{ak} - x_{bk}) \sigma^{kl} (x_{al} - x_{bl})$$

où $\Sigma = (\sigma_{kl})$, définie par $\Sigma = (X^c)' \Pi X^c$ [8] est la matrice de covariance pondérée entre les caractéristiques des variétés. Aussi, $\Sigma^+ = \left(\sigma^{kl}\right)$ est l'*inverse généralisé* de Moore-Penrose de la matrice Σ, laquelle n'est pas inversible ici, au vu de $q < p$.

On considère de plus une quatrième dissimilarité, la *dissimilarité du maximum* $D^{\max} = (d_{ab}^{\max})$ apparaissant dans la classification hiérarchique ascendante avec saut maximum (fig. 2), à partir des dissimilarités standardisées D^{corr}. Par définition, $d_{ab}^{\max}$ est la hauteur,

[8] $\Pi = \mathrm{diag}(f)$ est la matrice diagonale des poids des variétés, $X^c = HX$ la matrice des scores centrés et $H = I - \mathbf{1}f'$ la matrice de centration.

telle que lue sur le dendrogramme, à laquelle les variétés a et b sont agrégées[9].

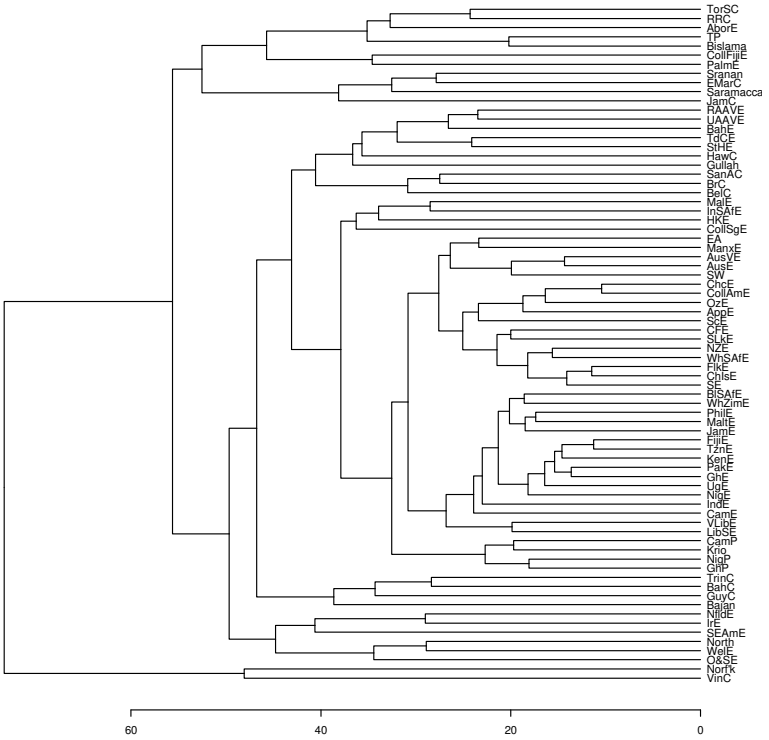

Figure 2 : Dendrogramme résultant de la classification hiérarchique ascendante des 72 variétés d'anglais[10], avec saut maximum, à partir des dissimilarités standardisées.

On a que $\max_{ab} d_{ab}^{\max} = \max_{ab} d_{ab}^{\mathrm{corr}} = 73{,}4$ comme il se doit. Les premières variétés à être fusionnées dans le dendrogramme sont le *Colloquial American English* et le *Chicano English*. Les deux dernières variétés à rejoindre le reste de l'amas sont le *Vincentian Creole*

9 Elle s'obtient à l'aide des fonctions `hclust()` et `cophenetic()` de R.

10 La liste des abréviations se trouve sur http://ewave-atlas.org.

(dans les Caraïbes) et le *Norfolk Island* ou *Pitcairn English* (un pidgin du Pacifique), en bas du dendrogramme.

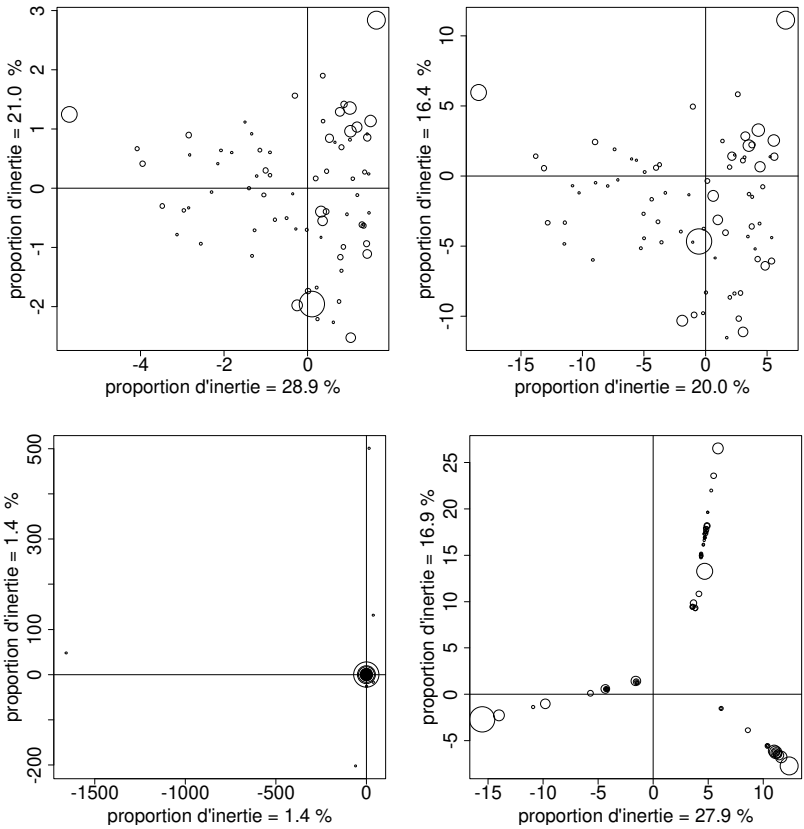

Figure 3 : Configuration (p, D) sur les deux premières dimensions déterminées par MDS, et proportion d'inertie Δ expliquée. Les représentations de D^{cov} (gauche, haut) et D^{corr} (droite, haut) sont usuelles, et les autres le sont beaucoup moins : D^{Maha} (gauche, bas) génère une configuration sphérique des données, où chaque dimension exprime une même proportion de $1/(q-1) = 1{,}4\ \%$ d'inertie (configuration incompressible). Les alignements observés dans le MDS de $D^{\max}$ (droite, bas) sont une surprise, et pourraient être dus à la nature ultramétrique de $D^{\max}$ – une conjecture excitante à vérifier.

Chacune de ces quatre dissimilarités D est le carré d'une distance euclidienne, pour laquelle s'applique le multidimensional scaling (MDS) classique pondéré (p. ex. Bavaud 2011), consistant à représenter au mieux l'*inertie* $\Delta = 1/2 \sum_{ab} p_a \, p_b d_{ab}$ de la configuration (p, D) dans un espace de basse dimension (fig. 3).

3.3. Transitions de phase et entropie effective

Comme décrit dans la section 2.2, au fur et à mesure que l'on augmente β (*i.e.* que l'on *refroidit* un système perceptif dont on *augmente* ainsi le pouvoir de discrimination), on passe d'une seule variété d'anglais identifiée dans le régime de haute température à q variétés d'anglais identifiées dans le régime de basse température, en passant par une série de transitions de phases *discontinues* consistant chacune en l'émergence d'une nouvelle variété, grâce à l'amélioration de la finesse perceptive (figures 4, p. 24, et 5, p. 25).

Dans le régime de haute température, un seul percept subsiste, lequel peut à ce titre prétendre au statut de « prototype de variété d'anglais ». Géométriquement, c'est simplement le type le plus proche du centre de gravité de la configuration (p, D), et il se trouve être le *Chicano English* pour les dissimilarités brutes et standardisées, et le *Colloquial American English* pour les dissimilarités de Mahalanobis et du maximum.

A l'autre extrémité, la transition de basse température consiste en la disparition du premier percept, une possibilité particulièrement menaçante pour les variétés rares, excentrées ou proches d'une variété plus fréquente. On s'attend à ce que les poids perceptifs $r_b(\beta)$ des variétés (prototype excepté) croissent en général en fonction de β, mais il est déjà apparent que cette dynamique n'est pas triviale au vu de son comportement non monotone (fig. 5, p. 25, gauche, bas). Cela étant, et comme le veut la théorie, l'entropie effective $E(\beta)$ est continue et croissante (fig. 6, p. 26). Elle semble aussi fort bien approximée ici par la diversité réduite R. Est-ce toujours, nécessairement, le cas? La question reste ouverte, de même que bien d'autres à propos de ce modèle perceptif (section 2.2), qui, à l'instar d'un

Sudoku bien adapté au niveau du joueur, est suffisamment simple pour espérer progresser, et suffisamment difficile pour nous occuper un bon moment.

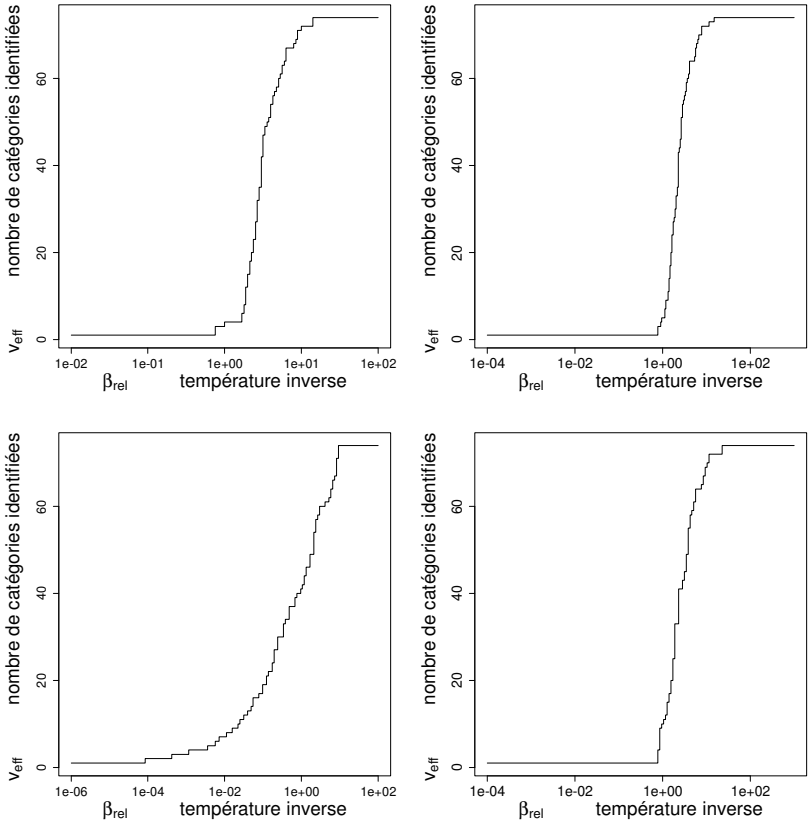

Figure 4 : Augmentation du nombre $v_{\text{eff}} = \sum_{b=1}^{q} I\,(r_b > 0)$ *de variétés d'anglais identifiées, en fonction de l'augmentation de la température inverse relative* $\beta_{\text{rel}} = \beta\Delta$*, où* Δ *est l'inertie de la configuration, pour les quatre dissimilarités correspondant à la figure 3 (p. 22). La définition précédente rend* β_{rel} *invariant par rapport à l'unité de mesure de* D*.*

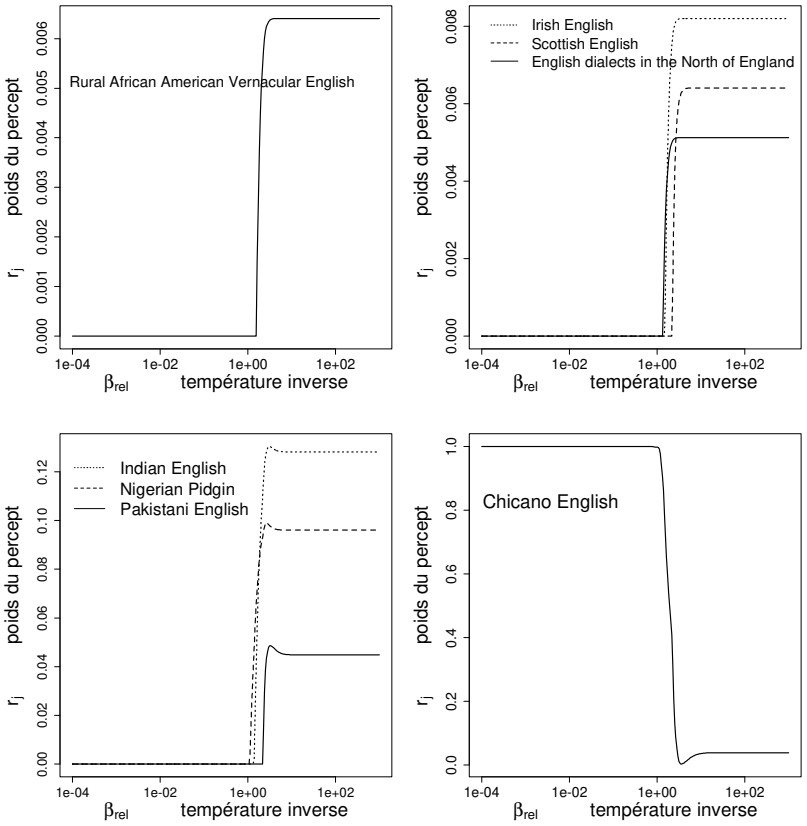

Figure 5 : Evolution du poids des percepts pour quelques variétés d'anglais, pour la dissimilarité standardisée D^{corr}. La transition de basse température β_L caractérise la première disparition, celle de la variété la plus « fragile » dans la configuration (p, D^{corr}), à savoir le Rural African American Vernacular English *(gauche, haut). En augmentant la température (*i.e. *en diminuant la température inverse), d'autres variétés disparaissent à leur tour. Le poids des percepts p_b, égal à leur poids réel f_b à très basse température, décroît en gros avec la température (droite, haut), mais pas toujours de façon monotone (gauche, bas) où un rebond peut être observé. A haute température, une seule variété subsiste, à savoir le* Chicano English *(droite, bas).*

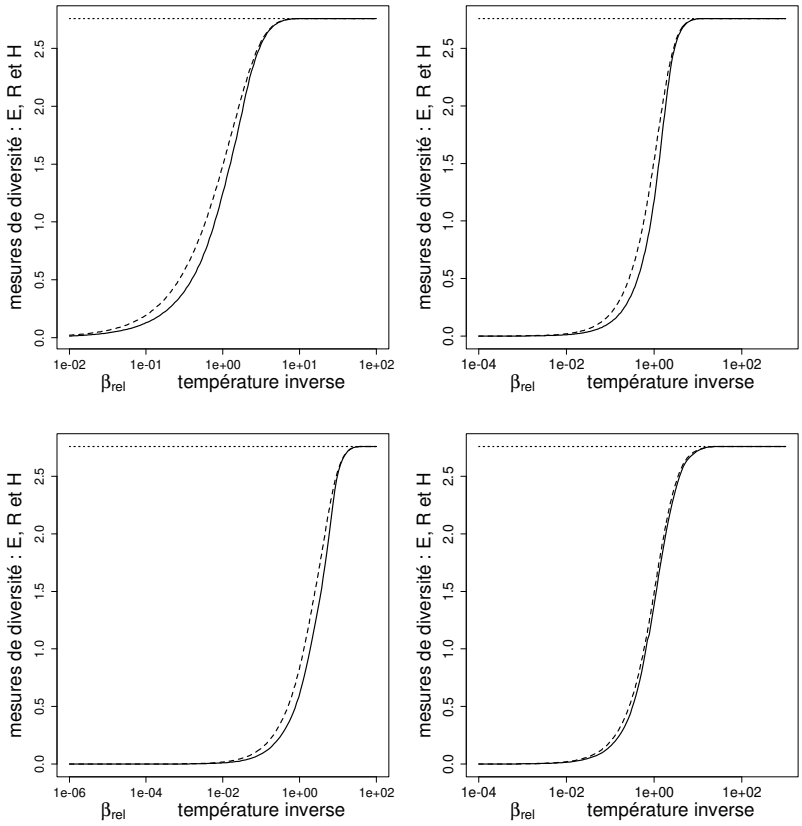

Figure 6 : Entropie effective E (trait continu), diversité réduite R (en traitillé) et entropie $\mathcal{H}$ (en pointillé), pour les quatre dissimilarités correspondant à la figure 3.

4. Discussion

Alors, combien y a-t-il de variétés distinctes d'anglais? Tout dépend de la faculté de discernement (au sens premier) du sujet-chercheur, ou de la sensibilité de l'appareil-machine si la question est déléguée. En utilisant des dissimilarités standardisées, et avec $\beta_{\mathrm{rel}} = 2{,}3$, la

réponse sera 42 comme dans le roman de Douglas Adams, mais elle sera 3 pour $\beta_{\mathrm{rel}} = 0{,}8$ ou encore 74 pour $\beta_{\mathrm{rel}} \geq 15{,}3$.

Certes, dans la meilleure (et la pire, ajouteront les chagrins) tradition académique, on a transformé une question simple (combien de variétés? combien de groupes? combien de dimensions dans une analyse factorielle? etc.) en une nouvelle question plus compliquée mais plus générale aussi, à savoir celle de la détermination du compromis optimal entre les idéaux opposés que sont la conservation de toute l'information et le souhait de la simplifier. Et dont le point d'équilibre dépend des usages et du public visés : on se contentera peut-être, pour le grand public, des trois ou quatre variétés les plus prototypiques, en réservant la description complète aux seuls spécialistes[11].

La problématique, vieille comme la science, mais dont la Théorie de l'Information a semble-t-il véritablement permis l'essor conceptuel et computationnel, est celle de la *compression avec perte*. Pour revenir au sujet, la transformation a permis :

— de caractériser le compromis au moyen d'un seul paramètre (la température ou son inverse) et de l'ancrer dans un formalisme familier : mécanique statistique, recherche opérationnelle, matrices de confusion

— de faire évoluer un problème immédiat (comment quantifier la réduction de diversité consécutive à la similarité possible entre types?) en la modélisation d'un possible mécanisme perceptif, formellement cohérent

— de tester, dans de futurs travaux, de l'adéquation (ou non) des prédictions de ce formalisme (par ex. quels percepts

11 Dans la présente étude, c'est la qualité du système perceptif, ou, ce qui revient au même, le souhait de donner une description plus ou moins simplifiée qui détermine la valeur de β_{rel}, et, partant, de toutes les mesures de diversité associées. D'autres applications possibles sont basées directement sur la matrice des similarités S, dont la donnée fixe *de facto* la valeur de β_{rel}.

sont attendus, avec quelles fréquences?) avec les données, telles qu'exprimées par les matrices de confusion, dont le caractère *quasi-symétrique*, tel que noté par Takane (1987), est également vérifié dans le formalisme proposé ici

— de revisiter et questionner la notion même de catégorisation « intuitive », omniprésente au début de toute enquête et récolte de données (linguistique ou non), consistant à regrouper d'emblée, par exemple, tous les locuteurs du *Scottish English* en un seul groupe, sur la base de leur *similarité* attestée ou postulée.

Naturellement, un travail *vraiment* fondamental devrait, dans une visée idéale, et totalement hors de portée, porter sur les $q = 780$ millions de locuteurs anglophones et leurs similarités mutuelles, plutôt que sur les $q = 74$ variétés étudiées ici.

Réalisons néanmoins cette *Erwärmung Gendankenexperiment* : à basse température, chacun des locuteurs, de même poids, et aux caractéristiques supposément uniques, est perçu individuellement. Une fois atteinte la température critique basse, les locuteurs disparaissent progressivement par assimilation à leurs voisins, jusqu'à ne compter plus qu'un représentant à partir de la température critique haute. Ce représentant unique est le locuteur anglophone *prototypique*, dont les caractéristiques se rapprochent le plus des caractéristiques moyennes de l'ensemble des locuteurs, pour une dissimilarité donnée.

Le recours à la notion de *prototype* est fréquent dans le raisonnement intuitif, et il est en soi intéressant de disposer ici d'un formalisme permettant de quantifier la dynamique de ce concept, pratiquement absent de l'Analyse des Données classique. Cette dernière lui préfère nettement la notion de *groupe* (cluster), formé par l'agrégation successive des individus, jusqu'à ne former qu'un seul groupe comme dans le sommet du dendrogramme de la figure 2 (p. 21). Le plus illustre des algorithmes de constitution de groupes est certainement le *soft K-means* ou *modèle de mélange*, où les données

que constituent la configuration (p, D) sont modélisées par un mélange de m gaussiennes isotropes, aux moyennes ajustables, et aux variances identiques, précisément données par la température $1/\beta$. Au fur et à mesure que l'on augmente la température, la dispersion des groupes augmente et ces derniers fusionnent, à la manière de gouttes d'eau, jusqu'à ne former plus qu'un seul groupe, si accommodant par son étendue qu'il finit par « expliquer » entièrement les données observées.

A cette approche classique correspond également une mesure de diversité (le négatif du « logarithme de vraisemblance »), quantifiant la difficulté à « expliquer » la configuration par un mélange gaussien (compression avec perte, de nouveau), décroissante avec la température, et à laquelle l'entropie effective E peut être directement comparée; mais ceci est une autre histoire et matière à nouvelle recherche, on en a déjà trop dit ici.

Ou presque : la différence entre l'approche développée ici et l'approche classique en Analyse des Données tient finalement aux contraintes de localisation des percepts, tenus de *coïncider avec les positions des stimuli* dans le premier cas, et *libres* dans le second. C'est en cours d'achèvement de cet article que cette question, rétrospectivement centrale, a progressivement émergé et a été finalement résolue, grâce aux bonnes conditions matérielles, intellectuelles et humaines dans lesquelles ce travail a pu être mené; merci à beaucoup de monde, et en premier lieu à Aris et à Marianne, pour cette belle opportunité.

RÉFÉRENCES

Bavaud François (2011), On the Schoenberg transformations in data analysis : Theory and illustrations, *Journal of Classification, 28(3)*, 297-314.

Emmanouilidis Théophile, Guex Guillaume & Bavaud François (2016), The transformed optimal transportation problem : sensitivity and segregation of the children-to-school constrained assignment in Lausanne, In *Proceedings of Spatial Accuracy 2016,* 333-340.

Jarvis Scott (2013), Capturing the diversity in lexical diversity, *Language Learning, 63(s1)*, 87-106.

Kortmann Bernd & Lunkenheimer Kerstin (2013), *The electronic world atlas of varieties of English*, Leipzig, Max Planck Institute for Evolutionary Anthropology.

Leinster Tom & Cobbold Cristina A. (2012), Measuring diversity : the importance of species similarity, *Ecology, 93(3)*, 477-489.

Marcon Eric (2015), *Mesures de la biodiversité*, thèse de doctorat, AgroParisTech.

Schneider Edgar W. & Kortmann Bernd (2004), *A handbook of varieties of English : A multimedia reference tool (vol. 1)*, Berlin, Mouton de Gruyter.

Takane Yoshio (1987), Analysis of contingency tables by ideal point discriminant analysis, *Psychometrika, 52(4)*, 493-513.

Xanthos Aris & Gillis Steven. (2010), Quantifying the development of inflectional diversity, *First Language, 30(2)*, 175-198.

Xanthos Aris (2018), Sur le rôle de la diversité lexicale dans la mesure de la diversité flexionnelle, *Cahiers de l'ILSL 56*, *311-331*.