Cahiers de l'ILSL, n° 36, 2013, pp. 231–252

L'évaluation (de l'évaluation)$^{+}$ de la diversité lexicale

Aris XANTHOS
Université de Lausanne

Les mesures de diversité lexicale sont systématiquement évaluées du point de vue de leur *robustesse*, mais rarement et peu rigoureusement quant à leur *sensibilité*. Cette contribution propose une méthodologie d'évaluation de la sensibilité basée sur une technique de génération de données textuelles permettant de contrôler artificiellement leur degré de diversité lexicale. La démarche est illustrée par la comparaison entre deux mesures basées sur deux façons différentes – et opposées – d'exploiter le principe de ré-échantillonnage pour l'évaluation de la diversité lexicale.[1]

While measures of lexical diversity are systematically evaluated from the point of view of their *robustness*, they are rarely and rather informally evaluated from the point of view of their *sensitivity*. This paper proposes a method for the evaluation of sensitivity based on a text generation algorithm which makes it possible to control the degree of lexical diversity of the generated data. The method is illustrated by means of a comparison between two measures relying on two different ways of applying a resampling strategy for the evaluation of lexical diversity.

[1] Merci à François Bavaud pour ses commentaires sur une première version de cette contribution.

1. INTRODUCTION

1.1 ROBUSTESSE ET SENSIBILITÉ DES MESURES DE DIVERSITÉ

La mesure de la diversité lexicale, au sens du caractère plus ou moins répétitif de l'usage du vocabulaire, est l'un des thèmes les plus fréquemment et exhaustivement traités dans la littérature portant sur le traitement quantitatif de données langagières et textuelles en particulier. Le point de focalisation de cet intérêt est la dépendance aussi fameuse que fâcheuse entre la diversité observée dans un corpus et sa taille. En effet, toutes les mesures de diversité lexicale dérivent plus ou moins directement de la *variété*, soit le nombre V de mots *distincts* (ou *types*) dans un corpus; or, la variété dépend de façon évidente, du moins pour sa valeur maximale, du nombre N de mots *successifs* (ou *tokens*) qui composent le corpus. A ce titre, elle ne permet pas de comparer directement des corpus de longueur différente – une critique à laquelle n'échappe pas le *rapport types-tokens* (*RTT*) V/N, sans doute l'indice le plus utilisé dans l'histoire de la mesure de la diversité lexicale (voir notamment Malvern, Richards, Chipere & Durán, 2004; Tweedie & Baayen, 1998).

Cet état de fait a conduit au développement de méthodes toujours plus sophistiquées pour tenter non seulement de rendre les mesures de diversité lexicale plus *robustes* vis-à-vis des variations de taille d'échantillon, mais aussi d'évaluer le succès de cette entreprise – deux objectifs souvent approchés par le biais de techniques de ré-échantillonnage visant à contrôler le volume des données. La situation est toutefois moins réjouissante en ce qui concerne les variations auxquelles ces mesures devraient être *sensibles*. En effet, à défaut d'un moyen de contrôler objectivement la diversité lexicale des données, on s'est généralement contenté d'une évaluation impressionniste de la sensibilité des mesures – lorsque la question n'a pas été purement et simplement ignorée.

L'argument le plus fréquemment utilisé consiste à montrer qu'une mesure donnée permet effectivement de discriminer des échantillons pour lesquels une différence de diversité lexicale est attendue. Une variante plus faible du même argument repose sur la démonstration de la similarité entre les résultats obtenus avec la mesure considérée et ceux obtenus avec une autre mesure réputée valide (typiquement en vertu du critère précédent).

Le principal objectif de la présente contribution est de proposer un mode d'évaluation plus rigoureux de la sensibilité des mesures de diversité lexicale, sur la base du processus de génération d'échantillons textuels "approximés" décrit par Shannon (1948). En particulier, on cherchera à contrôler la diversité lexicale du texte généré en appliquant au modèle probabiliste utilisé une opération mathématique analogue au *refroidissement* d'un système physique (Bavaud & Xanthos, 2002). La méthodologie proposée sera illustrée par la comparaison entre deux mesures de diversité lexicale reposant sur une stratégie de ré-échantillonnage et représentant deux façons opposées d'approcher le problème: fixer le volume de données et mesurer la diversité correspondante ou, à l'inverse, fixer la diversité et mesurer le volume de données correspondant.

1.2 L'APPORT DE LA GÉNÉRATION DE TEXTE "APPROXIMÉ"

L'usage de méthodes de ré-échantillonnage est bien établi dans le champ de la mesure de la diversité lexicale. Elles sont notamment appliquées de façon très systématique pour évaluer la robustesse des mesures par rapport aux variations de taille d'échantillon. Dans ce contexte, la pratique la plus courante consiste à extraire d'un corpus donné (de longueur N) B *sous*-échantillons de $S < N$ tokens, appliquer la mesure examinée à chacun de ces sous-échantillons et calculer la moyenne de la

mesure sur les B sous-échantillons. Toute l'opération est alors répétée avec des valeurs différentes de S, afin d'obtenir une estimation de la façon dont la mesure dépend du volume de données: dans l'idéal, les variations observées seront négligeables et attribuables au "bruit" aléatoire introduit par la procédure de ré-échantillonnage.

Les applications de cette démarche varient du point de vue de la méthode de construction des sous-échantillons. En particulier, on peut distinguer deux approches: la première consiste à découper le corpus en séquences de tokens contigus (voir p.ex. McCarthy & Jarvis, 2010), tandis que la seconde repose sur une sélection aléatoire du nombre d'éléments voulu, sans contrainte spécifique sur leur position dans le texte (voir p.ex. Tweedie & Baayen, 1998). Selon les cas, un même token peut apparaître dans plusieurs sous-échantillons, mais le point commun de toutes les variantes de cette méthodologie d'évaluation est qu'un token donné ne peut apparaître qu'une seule fois dans chaque sous-échantillon. Il s'agit en ce sens de tirage *sans remise*, ce qui fait la simplicité d'application de la procédure mais restreint par ailleurs les possibilités de paramétrage au seul contrôle de la taille des sous-échantillons.

Shannon (1948) introduit une technique de génération de données textuelles dont l'une des versions les plus simples (qu'il désigne par le terme d'*approximation du premier ordre*) peut s'interpréter comme un tirage *avec remise* parmi les tokens d'un corpus. En pratique, il s'agit d'abord de compter le nombre d'occurrences n_i de chaque type de mot i dans le corpus, puis de générer un texte en concaténant des mots sélectionnés aléatoirement avec probabilité $p_i \coloneqq n_i/N$, où $N \coloneqq \sum_i n_i$ dénote le nombre de tokens dans le corpus. Le tirage de chaque mot successif est indépendant des autres, si bien que l'ordre des mots dans les données ainsi produites ne reflète en général pas celui du corpus utilisé; en revanche, à mesure que croît la taille du texte généré, la fréquence relative

de chaque type dans ces données tend vers celle du corpus d'origine.

La procédure de génération de texte approximé peut être paramétrée bien plus finement que l'échantillonnage par tirage sans remise. En effet, le modèle explicite de la probabilité de chaque type de mot sur lequel repose la méthode peut être manipulé préalablement à l'étape de génération proprement dite afin de modifier certaines propriétés des données produites. Ainsi, Bavaud et Xanthos (2002) illustrent la possibilité d'appliquer aux probabilités une transformation mathématique dont l'effet est analogue à la modification de température d'un système physique: "chauffer" le modèle aura pour conséquence de faire tendre la distribution des mots vers l'uniformité; à l'inverse, le "refroidir" aboutira à rendre les mots fréquents encore plus fréquents, jusqu'au point extrême – déterministe – où l'entier de la masse de probabilité sera concentré sur un seul mot.[2] Cette approche permet donc d'augmenter ou de réduire artificiellement la diversité lexicale du texte généré.

Cette possibilité peut être exploitée pour évaluer la sensibilité d'une mesure de diversité. La démarche repose sur l'estimation d'un modèle probabiliste à partir d'un corpus donné et la génération, sur la base de ce modèle, d'échantillons de référence "à température ambiante". Il s'agit ensuite d'abaisser graduellement la température du modèle et générer des échantillons de plus en plus froids – donc de moins en moins

[2] Formellement, on fait l'hypothèse que la température originale du modèle est 1, et l'on fixe sa nouvelle température à T en calculant pour chaque type de mot i la probabilité modifiée $\hat{p}_i \coloneqq (p_i)^\beta / Z$, où $\beta \coloneqq 1/T$ est la température inverse et Z une constante de normalisation assurant que $\sum_i \hat{p}_i = 1$; le modèle est ainsi chauffé si $T > 1$ et refroidi si $T < 1$.

diversifiés lexicalement – afin d'estimer la température à partir de laquelle la mesure considérée permet de détecter une différence statistiquement significative avec les échantillons de référence. La mesure sera ainsi jugée d'autant plus sensible qu'elle répond à de faibles variations de température du modèle.

1.3 RÉ-ÉCHANTILLONNAGE ET MESURES DE DIVERSITÉ

En plus d'être communément exploité pour l'évaluation des mesures de diversité lexicale, le principe de ré-échantillonnage est utilisé depuis longtemps pour tenter de rendre ces mesures plus robustes vis-à-vis des variations de taille d'échantillon. A ma connaissance, Johnson (1944) est le premier à avoir proposé une telle démarche. Son point de départ est le défaut de robustesse notoire du RTT, qui tend à décroître lorsque N croît. Pour y remédier, Johnson propose de découper le corpus en B séquences de $S < N$ tokens contigus, calculer le RTT dans chaque séquence et rapporter finalement la moyenne du RTT dans les B séquences. Notons que dans cette approche, le nombre B de séquences est déterminé par le rapport entre leur longueur S et celle du texte N: en particulier, on a $B = \lfloor N/S \rfloor$, soit le résultat de la division entière de N par S, et le reste de cette division correspond au nombre de tokens à la fin du corpus qui ne pourront pas être exploités pour la mesure.

L'idée est simple mais efficace et elle réapparaît quelques décennies plus tard dans les travaux français de statistique lexicale (voir p.ex. Dubrocard, 1988). Dans ce contexte, on pratique plus volontiers un tirage aléatoire sans remise de B sous-échantillons de S tokens qu'un découpage du corpus en séquences de tokens contigus. Cette modification a surtout deux avantages: d'une part, elle fait de B un véritable paramètre, dont l'augmentation se traduit par une réduction de la variance de la mesure; d'autre part, elle permet d'éviter qu'une portion du texte ne soit pas prise en compte dans le cas général

où la longueur N du corpus n'est pas un multiple de la longueur S des sous-échantillons. Par ailleurs, il est plus commun dans les travaux de cette école d'estimer la moyenne de la variété des sous-échantillons (que j'appellerai *variété ré-échantillonnée* dans ce qui suit) que celle de leur RTT, mais la différence est superficielle dans la mesure où la seconde s'obtient en divisant la première par S.

Une contribution importante dans ce paradigme est celle de Serant (1988), qui montre notamment comment la loi hypergéométrique permet de *calculer* de façon exacte la moyenne et la variance de la variété sur tous les sous-échantillons possibles de taille S donnée, sans passer par le tirage effectif d'un seul de de ces sous-échantillons.[3] Cette possibilité semble toutefois être passée largement inaperçue dans le monde anglo-saxon: en effet, le tirage aléatoire et le "bruit" qu'il induit font partie intégrante de la méthodologie désignée par l'abréviation VOCD (pour *vocabulary diversity*, cf. McKee, Malvern & Richards, 2000), devenue le nouveau standard *de facto* pour la mesure de la diversité lexicale – après des décennies de domination du RTT en dépit de ses faiblesses avérées. L'estimation de VOCD implique une combinaison sophistiquée de mécanismes de ré-échantillonnage et d'ajustement de courbe dont l'exposé détaillé dépasserait le cadre de la présente étude. On peut toutefois mentionner que McCarthy et Jarvis (2007) font état d'une corrélation systématiquement très élevée entre VOCD et variété ré-échantillonnée exhaustive[4],

[3] Dans ce qui suit, je parlerai de variété ré-échantillonnée *exhaustive* pour distinguer la quantité calculée selon les indications de Serant (1988) d'une estimation obtenue par le tirage effectif de sous-échantillons.

[4] Voir aussi Xanthos et Gillis (2010) pour des éléments de comparaison supplémentaires dans le domaine de la diversité *flexionnelle*.

au point d'en conclure qu'à un changement d'unité près, la première mesure est véritablement une approximation de la seconde – au même titre que la variété ré-échantillonnée estimée par tirage aléatoire.

Il semble donc que VOCD et variété ré-échantillonnée (exhaustive ou non) soient les incarnations les plus récentes de la stratégie de mesure de diversité initialement proposée par Johnson (1944) et qu'on pourrait résumer informellement comme la mesure du degré de surprise moyen pour un volume de données fixé. Dans une contribution aussi récente qu'originale, McCarthy et Jarvis (2010) proposent d'inverser la perspective et mesurer plutôt le volume de données moyen pour un degré de surprise fixé. Intuitivement, les auteurs définissent la mesure qu'ils introduisent, MTLD (pour *measure of textual lexical diversity*), comme la longueur moyenne de ce qu'ils appellent un "facteur" (*factor*)[5], soit une séquence de tokens contigus qui maintient une valeur de RTT donnée. En pratique, le calcul de MTLD implique de calculer le RTT à chaque position successive du texte jusqu'à ce qu'il tombe en deçà d'un seuil prédéfini (cf. figure 1). Lorsque cela se produit, on incrémente le décompte des facteurs d'une unité et on remet le décompte des types et tokens à zéro. Quand l'entier du texte a été traité de la sorte, la valeur de MTLD s'obtient en divisant la longueur du texte par le nombre de facteurs observés.

Mot	*a*	*b*	*c*	*a*	*a*	*a*	*a*	*a*	*a*
Nb. types	1	2	3	3	3	1	1	1	1
Nb. tokens	1	2	3	4	5	1	2	1	2
RTT	1	1	1	.75	.6	1	.5	1	.5

Figure 1. Exemple de calcul de MTLD (voir détails dans le texte).

[5] Le choix de ce terme n'est pas très heureux, considérant ses acceptions déjà nombreuses dans le contexte des méthodes quantitatives.

Par exemple, sur la figure 1 (où les mots sont représentés par des lettres), le RTT vaut 1 pour les trois premières positions, puis décroît jusqu'à 0.6 après cinq tokens. Comme cette valeur est inférieure au seuil de 0.72 que McCarthy et Jarvis (2010) proposent d'adopter, on pose une frontière de facteur à cette position (ligne traitillée) et les comptes sont réinitialisés. Après avoir examiné les 9 tokens du texte, on dénombre un total de 3 facteurs, d'où une valeur de MTLD de 9/3 = 3.[6]

Tel que défini par McCarthy et Jarvis (2010), MTLD est remarquablement proche du RTT ré-échantillonné de Johnson (1944): la première mesure évalue la longueur moyenne de séquences de tokens dont le RTT est fixé, tandis que la seconde estime le RTT moyen de séquences dont la longueur est donnée. Les deux méthodes semblent donc toutes désignées pour servir de base à une comparaison en termes de robustesse et de sensibilité. Toutefois, comme elles reposent sur un mode *séquentiel* de constitution des sous-échantillons, elles n'offrent aucun contrôle sur le nombre de sous-échantillons impliqués dans le calcul de moyenne – nombre dont dépend en partie la variance de chaque mesure.

Pour remédier à cet inconvénient, on peut aisément concevoir une variante de MTLD basée sur le tirage aléatoire de sous-échantillons plutôt que sur le découpage du corpus en séquences de tokens contigus, de façon analogue aux dévelop-

[6] Cette présentation succincte laisse dans l'ombre certains détails de la méthode décrite par McCarthy et Jarvis (2010), notamment en ce qui concerne le traitement des facteurs "partiels", soit le cas où la dernière séquence de tokens du texte n'atteint pas le seuil fixé: en bref, les auteurs proposent d'estimer par interpolation linéaire la longueur totale du facteur partiel, et de réduire en proportion le poids de cette observation dans le calcul de la longueur moyenne.

pements récents que la proposition de Johnson a connus. Ainsi modifiée, la mesure ne tient plus compte de l'ordre original des tokens et perd donc la seule spécificité que McCarthy et Jarvis (2010) aient choisi de mettre en avant dans la dénomination "measure of *textual* lexical diversity". En contrepartie, cette nouvelle définition fait du nombre B de sous-échantillons un paramètre arbitraire, comme c'est le cas dans le calcul de la variété ré-échantillonnée tel que pratiqué par Dubrocard (1988) notamment (cf. *supra*). Ce sont en définitive ces deux mesures qu'on cherchera à comparer dans la suite de cette contribution pour illustrer la méthodologie d'évaluation proposée.

2. MÉTHODE

2.1 DONNÉES

Le corpus utilisé dans cette étude est la version électronique de *Notre-Dame de Paris* de Victor Hugo (1832) distribuée sur le site du projet Gutenberg. Le texte contient 186'101 mots (16'647 types) définis comme des séquences de caractères alphanumériques.

2.2 MESURES ÉVALUÉES

Comme discuté en section 1 *supra*, les deux mesures comparées dans les expériences suivantes ont été choisies pour représenter deux façons différentes d'appliquer le principe de ré-échantillonnage à l'évaluation de la diversité lexicale. La première mesure est la variété ré-échantillonnée estimée à partir de $B = 100$ sous-échantillons de $S = 50$ tokens tirés aléatoirement sans remise. La seconde est la variante de MTLD calculée sur la base de $B = 100$ sous-échantillons (de longueur variable) tirés aléatoirement sans remise, avec la valeur de seuil de RTT de 0.72 (voir section 1.3 *supra*). Dans

la suite, afin d'éviter d'alourdir excessivement la rédaction, les deux mesures seront simplement désignées par les termes *variété* et *MTLD*.

2.3 EXPÉRIENCES

Evaluation de la robustesse

L'évaluation de la robustesse des mesures de diversité vis-à-vis des variations de taille d'échantillon est conduite de façon tellement systématique dans la littérature qu'il paraît impensable de s'y soustraire dans le cadre d'une étude thématisant spécifiquement la méthodologie d'évaluation. Dans cette perspective, on a appliqué au texte de *Notre-Dame de Paris* la technique de génération de texte approximé de Shannon (introduite en section 1.2 *supra*[7]) afin de produire 100 échantillons de 100 tokens, puis 100 échantillons de 200 tokens, 400 tokens et ainsi de suite (800, 1'600, 3'200, 6'400 et 12'800 tokens).[8] Pour chaque échantillon de chacune des 8 tailles retenues, on calcule variété et MTLD tels que définis en section 2.2 *supra*. Pour chacune des deux mesures, on effectue finalement une analyse de variance (ANOVA) séparée afin de tester l'hypothèse que la moyenne de la mesure ne dépend pas de la taille d'échantillon, en espérant pouvoir l'accepter.

---

[7] Notons qu'aucune modification de température du modèle n'est appliquée dans cette première expérience.

[8] Pour être précis, les échantillons de 12'800 tokens ont d'abord été générés selon la procédure de Shannon, puis les échantillons de moindre taille ont été obtenus en prélevant les 100, 200, ..., 6'400 premiers tokens de chaque échantillon de 12'800.

Evaluation de la sensibilité

Pour évaluer la sensibilité des mesures aux variations de diversité lexicale, la procédure de génération de texte approximé décrite au paragraphe précédent (100 échantillons de 100, 200, ..., 12'800 tokens) a été répétée à l'identique après avoir préalablement refroidi la distribution des mots estimée dans *Notre-Dame de Paris* à la température $T = 0.999$ selon les indications données en section 1.2 *supra* (cf. note 1), puis $T = 0.998, 0.997, ..., 0.991, 0.99$. Pour chacune des deux mesures et pour chaque taille d'échantillon, on teste alors successivement l'hypothèse que la moyenne de la mesure est la même pour les échantillons refroidis à chaque température que pour les échantillons (dits *de référence*) produits "à température ambiante" dans le cadre de la première expérience – cette fois-ci avec l'espoir de *rejeter* l'hypothèse (pour la plus large gamme de tailles d'échantillon).

3. RÉSULTATS

3.1 EVALUATION DE LA ROBUSTESSE

Les figures 2 et 3 (page suivante) représentent la moyenne de la variété et de MTLD respectivement (au sens défini en section 2.2 *supra*) en fonction de la taille d'échantillon. L'inspection de ces diagrammes suggère que la variété est plus robuste que MTLD vis-à-vis des variations de taille d'échantillon. En effet, elle semble relativement stable sur tout l'intervalle de taille considéré, tandis que MTLD présente tendance une légèrement croissante, avec en particulier une valeur sensiblement inférieure aux autres pour la taille 100.

Les résultats de l'ANOVA confirment l'impression visuelle: l'hypothèse que la variété ne dépend pas de la taille d'échantillon est loin d'être rejetée ($F[7,792] = 0.336$, $p = 0.938$), contrairement à l'hypothèse correspondante pour MTLD ($F[7,792] = 21.19$, $p < 0.001$). L'application des tests post-hoc

de Tukey permet de conclure que c'est spécifiquement la moyenne de MTLD observée pour la taille 100 qui s'écarte de celle obtenue pour les autres tailles ($p < 0.001$), celles-ci n'étant pas significativement différentes entre elles prises deux à deux ($0.425 \leq p < 1$).

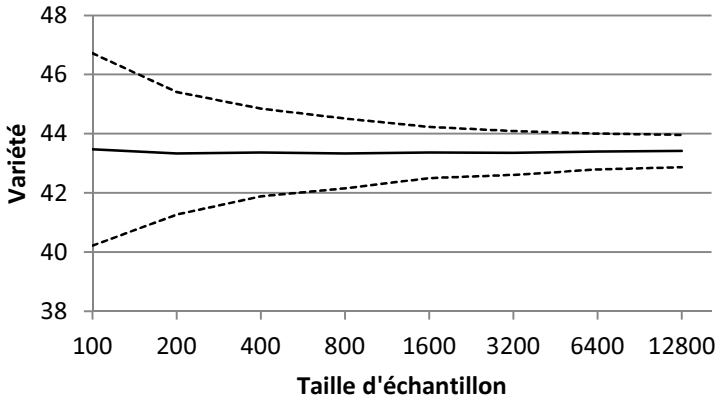

Figure 2. Variété moyenne en fonction de la taille d'échantillon (les lignes traitillées représentent ± 2 écarts-type).

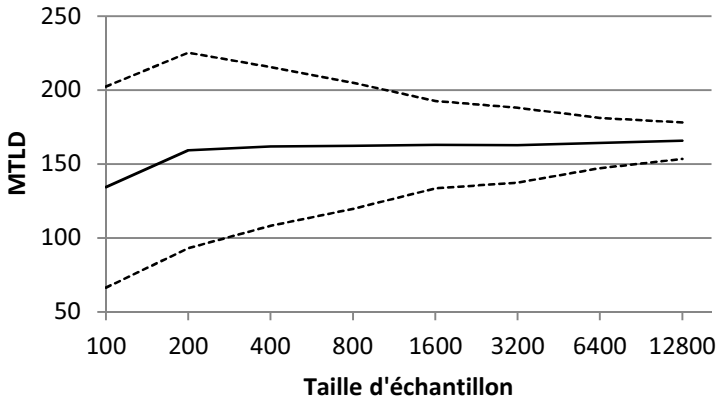

Figure 3. MTLD moyen en fonction de la taille d'échantillon (les lignes traitillées représentent ± 2 écarts-type).

Les résultats de l'ANOVA confirment l'impression visuelle: l'hypothèse que la variété ne dépend pas de la taille d'échantillon est loin d'être rejetée ($F[7,792] = 0.336$, $p = 0.938$), contrairement à l'hypothèse correspondante pour MTLD ($F[7,792] = 21.19$, $p < 0.001$). L'application des tests post-hoc de Tukey permet de conclure que c'est spécifiquement la moyenne de MTLD observée pour la taille 100 qui s'écarte de celle obtenue pour les autres tailles ($p < 0.001$), celles-ci n'étant pas significativement différentes entre elles prises deux à deux ($0.425 \leq p < 1$).

Pour les deux mesures, on constate par ailleurs que l'écart-type est strictement décroissant en fonction de la taille d'échantillon. Cette observation traduit le fait qu'indépendamment du caractère plus ou moins stable de la moyenne de la mesure, il y a naturellement quelque chose à gagner à disposer d'un plus large échantillon. C'est aussi ce qui justifie d'évaluer la sensibilité de la mesure pour plusieurs tailles d'échantillon comme on le fera dans la section suivante. En effet, en supposant que le résultat d'une mesure donnée ne dépende pas de la taille d'échantillon, la différence de moyenne observée entre échantillons de référence et refroidis à une température donnée sera d'autant plus significative que l'écart-type correspondant sera faible. Pour l'évaluation de la sensibilité, on peut donc s'attendre à des résultats différents en fonction de la taille d'échantillon – résultats qui favoriseront en particulier la mesure pour laquelle la décroissance de l'écart-type est la plus rapide.

3.2 Evaluation de la sensibilité

Les figures 4 à 6 (pages suivantes) représentent la significativité de la différence entre variété et MTLD moyens dans les échantillons de référence et dans les échantillons refroidis à température $T = 0.999, 0.998, ..., 0.99$, et ce pour des tailles d'échantillon de 100, 800 et 6'400 tokens. Rappelons que (i) plus la significativité (ou valeur p) est faible, plus on rejette

fortement l'hypothèse que la moyenne de la mesure dans les échantillons refroidis est la même que dans les échantillons de référence; (ii) une température plus basse correspond à une diversité lexicale plus réduite. En théorie, les courbes de significativité associées aux deux mesures devraient donc être strictement décroissantes et les écarts à ce principe s'expliquent par le "bruit" introduit par le processus de génération aléatoire des échantillons.

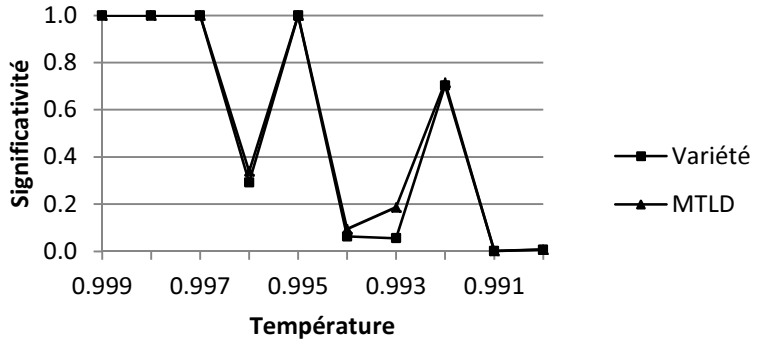

Figure 4. Evaluation de la sensibilité (échantillons de 100 tokens).

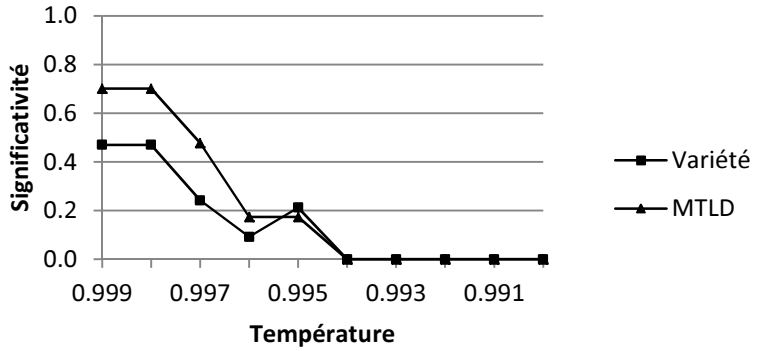

Figure 5. Evaluation de la sensibilité (échantillons de 800 tokens).

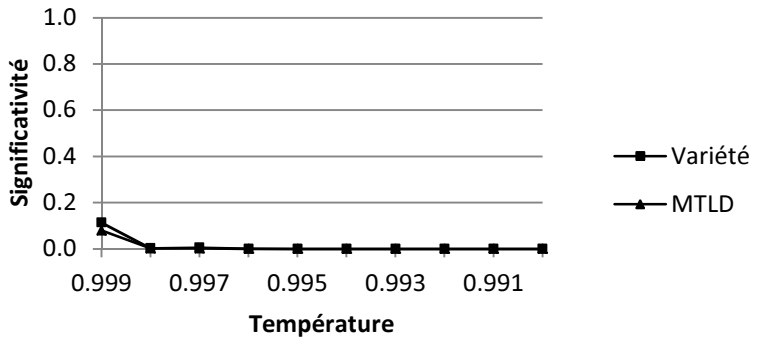

Figure 6. Evaluation de la sensibilité (échantillons de 6'400 tokens).

Les trois diagrammes montrent une grande similarité de comportement entre les deux mesures. Globalement, elles s'avèrent relativement peu sensibles pour des échantillons de petite taille (100 tokens): ce n'est qu'à partir d'une température de 0.991 qu'elles atteignent définitivement le seuil de significativité standard de 0.01. Avec des échantillons de taille plus conséquente (6'400 tokens), le même seuil est atteint déjà à une température de 0.998 pour les deux mesures. Les différences les plus importantes sont observées avec les échantillons de taille intermédiaire (800 tokens): la variété semble alors légèrement plus sensible que MTLD en général – encore que ce soit à la même température (0.994) que les deux mesures franchissent définitivement le seuil de 0.01.

La figure 7 (page suivante) résume les résultats observés sur toute la plage de tailles d'échantillon considérée. Sur ce diagramme, les courbes représentent la taille d'échantillon minimale requise pour passer définitivement en deçà du seuil de significativité de 0.01, en fonction de la température. Ainsi, avec des échantillons de 100 et 200 tokens, les deux mesures ne détectent une différence significative que pour un fort refroidissement (T = 0.991); avec des échantillons de 400 tokens, elles répondent à un abaissement de température légèrement moindre (T = 0.994), et ainsi de suite. A l'aune de cette

représentation synthétique, les deux mesures présentent exactement le même comportement, si bien que leurs courbes sont confondues.[9]

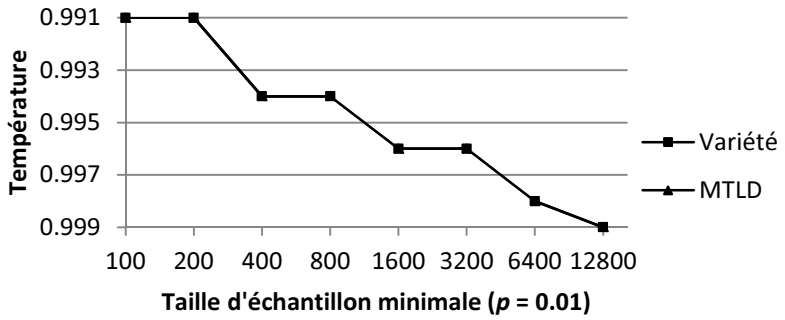

Figure 7. Evaluation de la sensibilité (vue d'ensemble)

DISCUSSION

Les résultats de l'évaluation de la robustesse (cf. section 3.1 *supra*) montrent que la moyenne de la variété est stable sur toute la plage de tailles d'échantillon considérée, tandis que celle de MTLD est significativement différente entre les échantillons de 100 tokens et ceux de taille plus élevée: environ 134 contre 163 respectivement. Cette dernière valeur permet de comprendre le phénomène. Elle indique en effet qu'il faut en moyenne tirer 163 tokens du corpus utilisé dans cette étude pour atteindre un RTT de 0.72 et ainsi observer un facteur complet (au sens de McCarthy & Jarvis, 2010). En

[9] Par ailleurs, le diagramme montre bien l'effet positif du phénomène de décroissance de l'écart-type avec la taille d'échantillon discuté au dernier paragraphe de la section 3.1 *supra*.

conséquence, la plupart des échantillons de 100 tokens ne contiendront qu'un facteur partiel et leur valeur de MTLD sera donc estimée sur la base d'une interpolation linéaire (cf. section 1.3 *supra*, note 5). Or cette estimation est peu satisfaisante, comme en témoigne la différence observée entre les échantillons de 100 tokens et ceux de taille supérieure.

Cette situation tient vraisemblablement au choix d'utiliser ici une variante de MTLD basée sur un tirage aléatoire des sous-échantillons plutôt que sur le découpage du corpus en séquences de tokens contigus. Toutes choses étant égales par ailleurs, on peut s'attendre à ce que la diversité lexicale de sous-échantillons aléatoires soit en moyenne plus élevée que celle de séquences de tokens, dans la mesure où l'unité thématique de ces dernières rend plus probable l'occurrence de mots répétés. Si cette conjecture est correcte, il est possible que l'utilisation d'un seuil de RTT plus élevé que celui proposé par McCarthy et Jarvis (2010) pour la version originale de MTLD permette de corriger le défaut de robustesse observé.

Sur le plan de la sensibilité (cf. section 3.2 *supra*), les résultats obtenus montrent que les deux mesures se comportent de façon remarquablement similaire. Lorsque ce n'est pas le cas, c'est le plus souvent que la variété s'avère légèrement plus sensible que MTLD – une différence qui disparaît complètement quand on ne s'intéresse qu'au point où la significativité atteint le seuil standard de 0.01.

Dans l'ensemble, les similarités l'emportent sur les différences dans les résultats de cette comparaison. Il semble ainsi que les aspects méthodologiques que variété et MTLD ont en commun – fondamentalement, leur intégration du principe de ré-échantillonnage – conditionnent leurs performances plus fortement que les traits qui les distinguent, soit le fait de fixer le volume de données (taille des sous-échantillons) ou le degré de diversité (seuil de RTT). On peut se demander dans quelle mesure les valeurs standard affectées à ces deux paramètres dans le cadre de cette étude ont déterminé le (faible)

avantage observé pour la variété sur les deux critères. Pour connaître la réponse à cette question, il faudra que soient conduites des expériences supplémentaires permettant d'examiner de quelle façon le comportement des deux mesures est affecté par les variations de leurs paramètres.

CONCLUSION

Le point de départ de cette contribution est le constat que les méthodes d'évaluation des mesures de diversité lexicale sont bien développées en ce qui concerne le critère de robustesse, mais encore rudimentaires pour ce qui est du critère de sensibilité. On s'est donné pour objectif de proposer un traitement rigoureux de l'évaluation de la sensibilité, sur la base de la technique de génération de texte approximé de Shannon (1948), modifiée selon les indications de Bavaud et Xanthos (2002) pour contrôler le degré de diversité lexicale des données générées par le biais d'un paramètre de température.

En guise d'illustration, la méthodologie d'évaluation proposée a été appliquée à la comparaison de deux mesures de diversité reposant sur le principe de ré-échantillonnage des données et choisies pour représenter deux approches opposées du problème: la variété ré-échantillonnée (voir par exemple Dubrocard, 1988) et une variante de MTLD (McCarthy & Jarvis, 2010) basée sur le tirage aléatoire de tokens plutôt que sur le découpage du corpus en séquences de tokens contigus. Sur la base du texte de *Notre-Dame de Paris* (Hugo, 1832), les expériences conduites ont permis de conclure que les deux méthodes obtiennent des résultats très similaires en termes de robustesse et de sensibilité. La seule différence notable est un problème de sous-évaluation de la diversité pour de très petits échantillons (100 tokens) avec MTLD – problème qu'on peut sans doute imputer aux modifications apportées ici à l'algo-

rithme de McCarthy et Jarvis (2010) sans avoir par ailleurs ajusté ses paramètres en conséquence.

Considérant que les deux mesures de diversité évaluées dans cette étude constituent deux développements radicalement différents de la proposition séminale de Johnson (1944), il est remarquable de parvenir à la conclusion qu'elles sont à peu près interchangeables du point de vue de leurs performances: grosso modo, elles présentent la même réaction (ou absence de réaction) aux variations de taille et de diversité des données. Cette interprétation des résultats repose naturellement sur le présupposé que la méthodologie d'évaluation introduite dans cette contribution est valide. Une autre interprétation possible est que la méthodologie proposée n'est elle-même pas suffisamment sensible pour pouvoir discriminer correctement les mesures de diversité lexicale. Afin d'exclure cette possibilité, il faudrait être en mesure de conduire objectivement une évaluation de l'évaluation de l'évaluation de la diversité lexicale, et ainsi de suite – au risque, on le voit, de ne plus guère avoir de diversité à évaluer en fin de compte.

Dans une perspective épistémologique qui dépasse le contexte particulier de la mesure de la diversité lexicale pour toucher celui, plus général, de la quantification appliquée aux problématiques de sciences humaines, on peut alors se demander si, et le cas échéant, dans quelles conditions, il est justifié d'interrompre l'engrenage consistant à reporter systématiquement le défaut d'objectivité à un niveau supérieur. Je ne serais pas surpris que ce questionnement ait déjà fait l'objet d'une réflexion de la part de Remi Jolivet, à qui cet ouvrage est dédié et qui, le premier, a suscité chez moi l'intérêt qui deviendrait passion pour le raisonnement quantitatif appliqué aux faits de langue – un exercice auquel il s'adonnait déjà (cf. Jolivet, 1982) tandis que je découvrais à peine les joies de l'arithmétique.

RÉFÉRENCES

Bavaud F. et Xanthos, A. (2002). Thermodynamique et statistique textuelle: concepts et illustrations. In *Actes des 6es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2002)*, pp. 101–111.

Dubrocard M. (1988). Evaluation de l'étendue du lexique, quelques essais de simulation. In D. Labbé, P. Thoiron et D. Serant (éd.), *Etudes sur la richesse et la structure lexicale.* Paris-Genève: Slatkine–Champion, pp. 43–66.

Hugo V. (1832). *Notre-Dame de Paris [en ligne]*. Téléchargé depuis http://www.gutenberg.org/ebooks/19657 (août 2012).

Johnson, W. (1944). Studies in language behaviour: I. A program approach. *Psychological Monographs, 56*: 1–15.

Jolivet R. (1982). *Descriptions quantifiées en syntaxe du français. Approche fonctionnelle*. Genève-Paris: Slatkine.

Malvern D. D., Richards B. J., Chipere N. et Durán, P. (2004). *Lexical diversity and language development: quantification and assessment.* Basingstoke: Palgrave Macmillan.

McCarthy P. M. et Jarvis S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing, 24 (4)*: 459–488.

McCarthy P. M. et Jarvis S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity measurement. *Behavior Research Methods, 42(2)*: 381–392.

McKee G., Malvern D. et Richards B. (2000). Measuring Vocabulary Diversity Using Dedicated Software. *Literary and Linguistic Computing*, 15(3): 323–337.

Serant D. (1988). A propos des modèles de raccourcissement de textes. In D. Labbé, P. Thoiron et D. Serant (éd.), *Etudes sur la richesse et la structure lexicale*. Paris-Genève: Slatkine–Champion, pp. 77–91.

Shannon C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*: 379–423; 623–656.

Tweedie F. J. et Baayen R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*: 323–352.

Xanthos A. & Gillis S. (2010). Quantifying the development of inflectional diversity. *First Language, 30*: 175-198.