Analyse des différences lexicales entre des corpus : test ou distance du Khi-2 ?

Yves Bestgen[1]

UCL/CECL – Place du cardinal Mercier, 10 – B-1348 Louvain-la-Neuve – Belgique

Abstract

Pearson's Chi-square test is probably the most popular statistical test in corpus linguistics, particularly for studying linguistic variations between corpora. For a certain number of years, its use has been criticize because of the very large number of significant results it produces when it is applied to large corpora. Oakes and Farrow (*Literary and Linguistic Computing*, 2007, 22, 85-99) proposed various adaptations of this test in order to make it more efficient. By means of resampling procedures, the present study shows the seriousness of the problem and the failure of the cures. This negative conclusion contrasts with the benefits brought by correspondence analysis, the most traditional approach in statistical analysis of textual data to study this kind of questions.

Résumé

Le test du Khi-2 de Pearson est probablement le test statistique le plus populaire en linguistique de corpus, tout particulièrement lorsque l'accent est mis sur la mise en évidence de variations linguistiques entre des corpus. Depuis un certain nombre d'années, son utilisation est remise en cause en raison des très nombreux rejets de l'hypothèse nulle qu'il produit lorsqu'il est appliqué à de grands corpus. Oakes et Farrow (*Literary and Linguistic Computing,* 2007, *22*, 85-99) ont proposé différentes adaptations de ce test afin de le rendre plus adéquat. Au moyen de procédures de rééchantillonnage, la présente recherche démontre la gravité du problème et l'insuffisance des remèdes proposés. Cette conclusion négative constraste avec les bénéfices qu'apporte l'analyse des correspondances, l'approche probablement la plus classique en analyse des données textuelles pour traiter ce genre de questions.

Mots-clés: comparaison de corpus, test du Khi-2, taux d'erreur par test, procédure de rééchantillonnage, résidus standardisés.

1. Introduction

Le test du Khi-2 de Pearson est probablement le test statistique le plus populaire en linguistique de corpus, tout particulièrement lorsque l'objectif de la recherche est la mise en évidence de variations linguistiques entre des corpus (Rayson *et al.*, 2004). Hofland et Johansson (1982) y ont eu recours pour identifier les mots dont la fréquence d'emploi différencie l'anglais britannique de l'anglais américain sur la base de la comparaison de deux corpus d'un million de mots. Plus généralement, ce test a été employé pour analyser des corpus qui s'opposent selon leurs sources

1 Chercheur qualifié du F.R.S-FNRS.

(Baker, 2004; Leech et Fallon, 1992), leurs genres (Paquot et Bestgen, 2009; Tribble, 2000) ou encore leur modalité orale ou écrite (Rayson *et al.*, 1997).

Le point de départ d'une telle analyse est une table de contingence construite à partir de la fréquence d'un mot dans les deux corpus à comparer et du nombre total de mots présents dans chaque corpus. L'hypothèse nulle testée est que l'écart entre les fréquences d'emploi dans les deux corpus résulte uniquement de variations aléatoires, les deux échantillons comparés ayant été extraits aléatoirement d'une seule et même population. La statistique employée est :

[1] $$\sum \frac{(O - A)^2}{A}$$

dans laquelle O représente les fréquences observées et A les fréquences attendues sur la base des totaux marginaux et la sommation porte sur les quatre cellules de la table de contingence. Cette statistique est approximativement distribuée comme une Khi-2 à un degré de liberté. On notera que, lors de la comparaison de deux corpus, autant de tables de contingence que de mots à tester sont construites et analysées au moyen de ce test.

Cette approche a été appliquée à la comparaison des fréquences de mots entre un texte et un corpus, un corpus et un autre corpus ou plusieurs corpus entre eux (Bondi et Scott, 2010). La popularité de ce test a sans aucun doute été renforcée par son implémentation dans un logiciel aussi fréquemment employé en linguistique de corpus que *WordSmith Tools* (Scott, 1999) dont la fonction principale est l'identification des «Keywords», c'est-à-dire des mots qui passent avec succès le test du Khi-2 (ou son proche parent, le Khi-2 de vraisemblance maximum) pour un seuil de probabilité de 0,000001.

Malgré cette très grande popularité (ou à cause de celle-ci), l'emploi de ce test en linguistique de corpus a été critiqué en raison des très nombreux rejets de l'hypothèse nulle qu'il produit lorsqu'il est appliqué à des corpus de grandes tailles comme c'est fréquemment le cas (Baker 2004 ; Gries, 2005 ; Groom, 2010 ; Kilgarriff, 1996, 2005). Par exemple, Paquot et Bestgen (2009) ont observé, lors de la comparaison d'un corpus littéraire et d'un corpus académique de 15 millions de mots chacun, que plus de 90% de 10 333 mots testés étaient significativement plus fréquents dans un des deux corpus pour un seuil de probabilité de 0,000001.

Cette objection est au centre de l'étude de Oakes et Farrow (2007) qui propose pour y remédier deux modifications de la procédure de comparaison de corpus au moyen du test du Khi-2. Déterminer précisément la gravité du problème et l'efficacité des remèdes qu'ils proposent sont les deux objectifs principaux de la présente recherche. La section suivante décrit la méthodologie que ces chercheurs ont employée puisque cette étude sert de point de référence pour l'ensemble des analyses rapportées ici. Ensuite, les raisons pour lesquelles le test du Khi-2 signale un très grand nombre de différences significatives lorsqu'il est employé en linguistique de corpus sont discutées ainsi que les modifications proposées pour y remédier. Ces solutions sont évaluées dans la quatrième section au moyen de procédures de rééchantillonnage qui établissent leur insuffisance. Enfin, différentes autres solutions sont considérées dont une approche particulièrement classique en analyse statistique des données textuelles.

2. Méthodologie de l'étude de Oakes et Farrow (2007)

L'objectif de Oakes et Farrow (2007) était de mettre en évidence les différences de vocabulaire entre des variétés d'anglais employées dans différents pays. Pour ce faire, ils ont, dans leur première expérience, comparé cinq corpus contenant des extraits de textes anglophones originaires de cinq pays : Australie (AU), États-Unis (US), Grande-Bretagne (UK), Inde (IN) et Nouvelle-Zélande (NZ). Ces corpus sont composés de 500 extraits de textes de 2000 mots chacun, soit un million de mots, sauf le corpus australien, qui pour des questions de droits de reproduction ne contient que 373 extraits. Ils ont été construits de manière à être les plus comparables possible quant à la proportion de textes de différents genres (œuvres littéraires, articles de journaux, textes scientifiques...) qu'ils incluent.

Afin de pouvoir comparer ces corpus, Oakes et Farrow ont construit une table de contingence avec en colonne les 5 corpus et en ligne tous les mots différents dont la fréquence attendue dans chaque corpus est au moins égale à 5. Le test du Khi-2 appliqué à cette table permettant de rejeter l'hypothèse nulle d'une absence de différence entre les 5 corpus quant à la distribution des mots, ils ont calculé la participation de chaque cellule à ce Khi-2 global (voir formule 1) et en ont dérivé le résidu standardisé (Haberman, 1973) dont la formule est

[2] $$(O-A)/\sqrt{A}$$

Ce résidu est positif lorsqu'un mot est trop fréquent dans un corpus par rapport à la fréquence attendue et négatif dans le cas contraire. Considérant que la distribution de ce résidu est approximativement normale, Oakes et Farrow déterminent son degré de signification en les comparant à cette distribution. Il faut toutefois noter qu'à la suite d'Agresti (1990), García-pérez et Núñez-antón (2003) ont montré que les résidus standardisés calculés par cette formule sont asymptotiquement distribués normalement avec une moyenne de 0, mais une variance égale à $(I-1)(J-1)/(IJ)$ où I et J représentent le nombre de lignes et de colonnes dans la table. Pour que ces résidus soient réellement standardisés, il est donc nécessaire de les diviser par $\sqrt{(I-1)(J-1)/(IJ)}$. Cette correction a pour effet d'accroître la valeur du résidu et donc de le rendre plus facilement significatif. On notera qu'un logiciel de statistique aussi courant que SPSS (version 18) n'applique pas cette correction aux résidus «standardisés « ; il fournit heureusement en plus les résidus ajustés qui sont normalement distribués (Bibby, 1978 ; García-pérez et Núñez-antón, 2003).

La procédure qui vient d'être décrite est une simple extension du test du Khi-2, décrit dans l'introduction, à une table de contingence contenant plus de deux lignes et plus de deux colonnes.

3. Pourquoi tant de rejets de l'hypothèse nulle ?

Kilgarriff (2005) et Gries (2005) sont probablement les deux études qui ont investigué le plus directement l'origine du très grand nombre de résultats significatifs qui résultent de l'application du test du Khi-2 en linguistique de corpus. Les différentes explications considérées sont discutées dans la suite de cette section.

3.1. Le nombre de tests effectués

L'explication la plus évidente réside dans le très grand nombre de tests effectués (Gries, 2005). Chaque mot différent, dont la fréquence attendue dans les corpus comparés est au moins de 5, fait l'objet d'un test de signification statistique. Au total, cela peut représenter plusieurs dizaines de milliers de tests. L'analyste fait alors face à «l'écueil des comparaisons multiples, bien connu des statisticiens» (Lebart *et al.*, 2003, p. 203). Le seuil de rejet (α) de l'hypothèse nulle (classiquement 0,05) est valable pour un test. Il correspond à ce qu'on appelle *le taux d'erreur par test*, c'est-à-dire à la probabilité de commettre une erreur de type I (rejeter une hypothèse nulle étant donné qu'elle est vraie) dans le cas d'un seul test (Howell, 2008, pp. 354-355). Si deux tests *indépendants* sont effectués au seuil de 0.05, le taux d'erreur par test reste à 0,05, mais le *taux d'erreur de l'ensemble* (*familywise error rate*), c'est-à-dire la probabilité de rejeter à tord l'hypothèse nulle dans au moins un des deux tests, est beaucoup plus élevée puisqu'il est égale à $1-(1-0{,}05)^2$, soit 0,0975. Pour trois tests, cette probabilité vaut 0,1426. Pour contrecarrer cet accroissement de la probabilité d'au moins une erreur de type 1 sur un ensemble de tests, on procède classiquement en choisissant un seuil α' tel que $\alpha = 1-(1-\alpha')^n$ avec *n* égal au nombre de tests effectués. Pour 3 tests et un α souhaité de 0,05, cela donne un α' de 0,017. Ce qui précède est valable lorsque les tests sont indépendants les uns des autres, ce qui n'est évidemment pas le cas lorsque l'analyse porte sur la fréquence de nombreux mots dans les mêmes corpus. En situation de non-indépendance, on emploie fréquemment l'inégalité de Bonferroni qui stipule que *la probabilité d'occurrence d'un ou plusieurs événements ne peut jamais dépasser la somme de leurs probabilités individuelles* (Howell, 2008, p. 368) et donc on obtient α' en divisant le α souhaité par le nombre de tests effectués. Que les tests soient indépendants ou non, cet α' garantit (toute autre chose étant correcte par ailleurs) que la probabilité d'avoir un ou plusieurs tests significatifs par le seul fait du hasard sur l'ensemble des tests effectués est au plus α. Cette correction de Bonferroni est recommandée par Gries (2005) et c'est une des deux réponses de Oakes et Farrow (2007) au problème du trop grand nombre de tests significatifs.

3.2. Discordance entre les unités d'échantillonnage et d'analyse

Il est important de remarquer que la première explication et la correction proposée présupposent que le test produit un taux d'erreur par test correct. En d'autres mots, chaque test effectué doit avoir une probabilité α de rejeter l'hypothèse nulle. Si ce n'est pas le cas, appliquer la correction de Bonferroni ne permettra pas de résoudre le problème. Or, la deuxième raison potentielle pour laquelle on rejette trop souvent l'hypothèse nulle dans ce genre d'analyses met justement en cause ce taux d'erreur par test. Elle porte sur l'unité d'échantillonnage à l'origine de la table de contingence soumise au test. Pour que l'emploi *inférentiel* du test du Khi-2 soit valide, chaque observation reprise dans la table est censée avoir été sélectionnée dans la population correspondante par un processus aléatoire. En d'autres mots, l'unité analysée doit être l'unité d'échantillonnage (Baroni et Evert, 2009; Evert, 2006). Or, ce n'est habituellement pas le cas en linguistique de corpus. L'unité analysée est souvent le mot, parfois la phrase, alors que l'unité d'échantillonnage ayant servi à construire le corpus est le texte (ou un extrait de textes). Ceci ne poserait aucun problème si les textes étaient eux-mêmes composés d'un échantillon aléatoire de mots (ou de phrases), mais ce n'est évidemment pas le cas.

Pourquoi cette discordance entre unité d'échantillonnage et unité d'analyse peut-elle affecter le nombre de tests significatifs? On sait depuis longtemps que la fréquence d'occurrences des mots varie très fortement selon les textes (Church, 2000 ; Lafon, 1980). Il s'ensuit que la présence ou non d'un texte spécifique dans un corpus peut être suffisante pour accroître fortement la fréquence de certains mots et ainsi modifier les mots qui seront considérés comme significativement plus fréquents dans ce corpus selon le test du Khi-2. Ce phénomène est parfaitement illustré dans l'exemple suivant de Oakes et Fallon (2007). Ces auteurs ont observé, dans l'étude décrite ci-dessus, qu'un des mots les plus typiques de l'anglais britannique, selon le test du Khi-2, est *thalidomide*. Ils font toutefois remarquer que chacune des 55 occurrences de ce mot dans le corpus en question se trouve dans un seul et même texte. Contrairement donc à ce que semble indiquer le test du Khi-2, *thalidomide* n'est pas typique de l'anglais britannique, mais seulement d'un texte du corpus britannique. C'est parce que ce texte a été sélectionné en entier pour inclusion dans le corpus que *thalidomide* apparaît comme typique. Si, lors de la constitution de ce corpus, l'unité d'échantillonnage avait coïncidé avec l'unité d'analyse (le mot), *thalidomide* n'aurait eu (pratiquement) aucune chance d'être déclaré typique. Ainsi donc, chaque texte sélectionné risque de provoquer une série de faux positifs.

Si ce problème a été souligné à plusieurs reprises dans la littérature, très peu d'études ont essayé de montrer d'une manière explicite son impact sur le nombre de tests significatifs dû exclusivement à des variations aléatoires. Bien plus, plusieurs des auteurs ayant attiré l'attention sur ce problème suggèrent qu'il est atténué lorsque l'analyse porte sur de très grands corpus (Baroni et Evert, 2009 ; Oakes et Farrow, 2007). On peut toutefois penser que plus un corpus contient de textes et plus le problème se produira puisque chaque texte risque d'apporter son lot de mots typiques dont il accroît indûment la fréquence. Une vérification empirique serait donc bien nécessaire ici. A ma connaissance, seul Kigarriff (1996, 2005) a tenté une telle démonstration en constituant arbitrairement deux pseudocorpus en extrayant deux échantillons aléatoires de textes (et non de mots) d'un seul et même corpus, le *British National Corpus* (BNC). Kilgarriff soutient que dans une telle situation, on doit s'attendre à ce que les écarts entre les fréquences des mots dans les deux pseudocorpus résultent exclusivement de variations aléatoires et donc à ce que le test du Khi-2 ne rejette l'hypothèse nulle d'égalité des fréquences que dans 0.5% des cas si un seuil de probabilité de 0,005 est employé. Kilgarriff observe que «For very many words, including most common words, the null hypothesis is resoundingly defeated» (2005, p. 269). Toutefois, comme l'a souligné Gries (2005), l'expérience de Kilgarriff présente une limitation importante. N'ayant été effectuée qu'une seule fois, il est possible que les deux corpus construits soient relativement différents l'un de l'autre par le seul effet du hasard et donnent une image trop négative de l'efficacité du test.

Pour prendre en compte cette discordance entre les unités d'échantillonnage et d'analyses, Oakes et Farrow (2007) ont proposé d'ajouter un contrôle sur la dispersion des mots déclarés significatifs de façon à éliminer ceux dont la dispersion dans le corpus dont ils sont typiques est insuffisante. Pour ce faire, ils ont divisé chaque corpus en cinq sections et combiné deux indices : le nombre de sections dans lesquelles le mot apparaît et le *D* de Juilland qui est basé sur le coefficient de variation (V) calculé sur les fréquences d'occurrences du mot dans chaque section. Sa formule est

$$[3] \qquad D = 1 - \frac{V}{\sqrt{n-1}}$$

où *n* correspond au nombre de sections délimitées dans le corpus. *D* varie entre 0 et 1 ; plus il est proche de 1 et meilleure est la dispersion du mot. Oakes et Farrow (2007) ont défini empiriquement un seuil pour chacun de ces indices : pour être considéré comme suffisamment dispersé, un mot doit apparaître dans au moins 3 des 5 sections du corpus et le *D* de Juilland doit être supérieur ou égal à 0.30.

3.3. Conclusion

Cette section discute deux raisons pour lesquelles le test du Khi-2, tel qu'il est employé en linguistique de corpus, pourrait produire un nombre beaucoup trop élevé de résultats significatifs. La première, le très grand nombre de tests effectués, est bien établie et un remède connu est préconisé par Oakes et Farrow (2007) : l'emploi de la correction de Bonferroni. La seconde raison, la discordance entre les unités d'échantillonnage et d'analyse, est nettement plus problématique parce qu'elle met en cause l'emploi même du test et que sa gravité est mal connue. De plus, l'efficacité de la solution proposée par Oakes et Farrow, l'emploi d'une mesure de dispersion, n'a pas été évaluée. Leur étude n'apporte pas d'information à ce sujet parce qu'elle porte sur une situation dans laquelle on peut penser que l'hypothèse nulle (absence de différences de fréquence des mots dans les corpus) est fausse. Il est donc normal de rejeter l'hypothèse nulle pour un certain nombre de mots et il est impossible de décider si on la rejette trop souvent ou non.

La recherche présentée ici se propose de répondre empiriquement à ces questions au moyen de simulations basées sur une procédure de rééchantillonnage de type permutation inspirée de l'étude de Kilgarriff (2005). Ces simulations sont basées sur les données employées dans l'étude de Oakes et Farrow (2007) afin de pouvoir évaluer le degré d'efficacité des solutions proposées. Si on peut montrer que leurs solutions sont efficaces, les autres recherches de ce type en linguistique de corpus gagneraient à les employer. Si, par contre, ces solutions sont insuffisantes, les conséquences seraient encore plus problématiques pour ces autres études.

4. Procédure, analyses et résultats

4.1. Reproductibilité des résultats originaux de Oakes et Farrow (2007)

Pour reproduire les analyses de Oakes et Farrow (2007), je suis parti, comme eux, du CD-ROM ICAME (Hofland *et al.*, 1999), qui contient les textes bruts des cinq corpus. Une série de prétraitements ont dû être appliqués comme la segmentation en mots, la suppression des signes de ponctuation et des caractères spéciaux. Ces prétraitements, réalisés au moyen du programme TreeTagger couramment employé en traitement automatique du langage, sont potentiellement différents de ceux effectués par Oakes et Farrow. En conséquence, une première analyse a été effectuée afin de vérifier qu'il est possible de reproduire fidèlement leurs résultats sur les données à ma disposition. Pour ce faire, les mots les plus significatifs obtenus dans la présente étude ont été comparés aux 50 mots les plus significatifs de chaque corpus rapportés dans le tableau 9 de Oakes et Farrow (2007, p. 96). Globalement, la corrélation entre les deux ensembles de valeurs est de 0,998 et l'écart moyen absolu entre celles-ci est inférieur à 4%. Quelques différences ont néanmoins été observées, mais limitées à des mots proches d'une des valeurs seuils de la procédure comme une fréquence attendue minimale de 5 ou un *D* de Juilland d'au moins 0.30.

4.2. Simulations en utilisant le mot comme unité d'échantillonnage

Une première série de simulations vise à confirmer que, lorsque toutes les autres conditions d'application du test du Khi-2 sont remplies, utiliser l'unité d'échantillonnage qui correspond à l'unité d'analyse produit des taux d'erreur de type 1 corrects. Pour ce faire, chacun des mots présents dans les corpus analysés par Oakes et Farrow a été assigné aléatoirement à un des 5 pseudocorpus créés arbitrairement. Ensuite, leur procédure d'analyse a été appliquée[2] et on a déterminé la proportion de tests significatifs pour différents α. On s'attend à ce que cette proportion soit proche de α. Cette simulation a été répétée 1000 fois en faisant varier la valeur initiale (*seed*) du générateur aléatoire.

Lorsque la formule classique du résidu standardisé est employée, le tableau 1 montre qu'on observe un rejet beaucoup trop rare de l'hypothèse nulle. Par contre, la formule corrigée, proposée par García-pérez et Núñez-antón (2003), donne des taux d'erreurs très proches des valeurs attendues, et ce pour tous les α. La troisième colonne du tableau 1 reprend le taux d'erreur de l'ensemble, après donc application de la correction de Bonferroni préconisée par Oakes et Farrow, pour les résidus calculés par la formule corrigée[3]. Les valeurs obtenues sont également proches des valeurs attendues.

	Taux d'erreur par test		Taux d'erreur de l'ensemble
a	Standard	Corrigée	Corrigée
0,05	0,0273	0,0483	0,034
0,025	0,0120	0,0239	0,023
0,01	0,0040	0,0098	0,012
0,005	0,0019	0,0051	0,008
0,001	0,0003	0,0012	0,000

Tableau 1 : Taux moyens d'erreur pour les 1000 simulations basées sur les mots

En conclusion, lorsque l'échantillonnage porte sur l'unité d'analyse, le test du Khi-2, tel qu'adapté par Oakes et Farrow et moyennant l'emploi de la formule corrigée des résidus standardisés, ne produit pas un trop grand nombre de tests significatifs.

Ces résultats sont valables lorsque l'unité d'échantillonnage correspond à l'unité d'analyse, mais ce n'est pas de cette manière que les corpus sont construits. Les simulations suivantes visent à déterminer l'efficacité du test du Khi-2, tel qu'implémenté par Oakes et Farrow, lorsque l'unité d'échantillonnage est le texte.

2 L'analyse de la dispersion n'a pas été effectuée puisque celle-ci est non pertinente lorsque le mot est employé comme unité d'échantillonnage et on a ajouté la correction du résidu standardisé pour la variance non-égale à 1.

3 Etant donné que les taux d'erreur par test sont inadéquats lorsqu'on emploie la formule non-corrigée des résidus standardisés, l'application de la correction de Bonferoni à ceux-ci présente peu d'intérêt.

4.3. Simulations en utilisant le texte comme unité d'échantillonnage

Les analyses présentées dans cette section sont identiques à celles rapportées ci-dessus sauf que l'unité d'échantillonnage est ici le texte et non le mot. Concrètement, chacun des textes analysés par Oakes et Farrow a été assigné aléatoirement à un des cinq pseudocorpus et leur procédure a été appliquée à ceux-ci afin de déterminer la proportion de tests significatifs pour différents α. Cette opération a été reproduite 1000 fois. Dans ces simulations, l'analyse de dispersion a un sens et a donc été appliquée.

Le tableau 2 donne les taux d'erreur par test pour la formule standard et pour la formule corrigée du résidu standardisé, mais seule cette dernière sera commentée puisque García-pérez et Núñez-antón (2003) et les simulations précédentes ont montré que la formule non corrigée ne produisait pas les taux d'erreurs adéquats lorsqu'unités d'échantillonnage et unité d'analyse coincident. Ce tableau est sans appel. Pour tous les α et toutes les conditions, les taux d'erreur par test sont nettement plus élevés qu'ils ne devraient l'être. Pour une probabilité de 0.001, le α choisi par Oakes et Farrow, ce taux est jusqu'à 66 fois trop élevé lorsque le contrôle de la dispersion n'est pas appliqué. Il est encore 47 fois trop élevé lorsque les mots mal dispersés sont considérés comme non significatifs. La prise en compte de la dispersion, telle que proposée par Oakes et Farrow, réduit donc le taux d'erreur, mais elle est nettement insuffisante pour régler le problème du trop grand nombre de tests significatifs.

	Sans dispersion		Avec dispersion	
a	Standard	Corrigée	Standard	Corrigée
0,05	0,1930	0,2384	0,1293	0,1625
0,025	0,1461	0,1841	0,0952	0,1230
0,01	0,0966	0,1356	0,0670	0,0890
0,005	0,0758	0,1051	0,0533	0,0718
0,001	0,0485	0,0658	0,0339	0,0468

Tableau 2 : Taux moyens d'erreurs par test pour les 1000 simulations basées sur les textes

Finalement, il est utile de se demander quel taux d'erreur de l'ensemble est obtenu lorsqu'on applique la procédure de Oakes et Farrow (procédure de Bonferroni et test de dispersion) à des données générées aléatoirement de telle manière que l'hypothèse nulle soit vraie. Dans cette analyse, on a employé le seuil de probabilité fixé par ces auteurs, au moyen de la procédure de Bonferroni, afin de n'avoir un ou plusieurs tests significatifs que dans au plus une des 1000 simulations effectuées, soit une valeur de $1{,}961 \times 10^{-9}$. Les résultats sont tout autre. Toutes les 1000 simulations contiennent au moins un test significatif et même beaucoup plus puisqu'on en observe en moyenne 291 par simulation (min = 234; max = 340). Ceci correspond à une moyenne de 59 tests significatifs par corpus dû au seul hasard. Au bénéfice de Oakes et Farrow, on notera que l'analyse des vrais corpus donne lieu en moyenne à 210 tests significatifs par corpus avec un minimum de 127 pour UK. On peut donc en conclure qu'il y a bien des différences d'emploi du vocabulaire dans ces variétés de l'anglais, mais que la probabilité que certains des mots

qu'ils pointent résultent des seuls effets du hasard, et donc ne reflètent pas une différence réelle, est beaucoup plus élevée que ce qu'ils espéraient.

5. Discussion et conclusion

Les simulations rapportées ci-dessus montrent que le test du Khi-2, tel qu'appliqué en linguistique à la comparaison de corpus, produit des taux d'erreur de type 1 beaucoup trop élevés. La prise en compte d'un seuil minimal de dispersion, remède proposé par Oakes et Farrow (2007), est insuffisante pour résoudre ce problème. Conclure sur un tel constat négatif serait regrettable. Cette section discute trois autres remèdes potentiels.

On pourrait, en premier lieu, imaginer de conserver la procédure proposée par Oakes et Farrow, mais en rendant encore plus extrême le seuil de probabilité de sorte d'obtenir un taux d'erreur de l'ensemble conforme à leur attente, soit une chance sur mille d'avoir au moins un test significatif parmi les milliers de tests effectués lorsque toutes les hypothèses nulles considérées sont vraies. Les simulations rapportées dans la section 4.3 permettent de se faire une idée de ce que donnerait cette option. Pour ce faire, le résidu standardisé[4] maximal (en valeur absolue) de chacune des 1000 simulations a été identifié et on a recherché la deuxième plus grande valeur parmi ceux-ci, soit 27,186. Si on fixe la valeur critique du test juste au-dessus de ce nombre, une seule des 1000 simulations effectuées donne lieu à au moins un résultat significatif. Cette valeur est si élevée que les procédures habituelles de calcul de probabilité pour la distribution normale renvoient une probabilité nulle. Quel serait l'impact d'une telle valeur critique sur les résultats réels de Oakes et Farrow? Au lieu des 210 mots significatifs obtenus en moyenne par corpus avec le seuil trop libéral, ils n'en auraient obtenu que 5 par corpus et même aucun dans le corpus britannique. Parmi les termes qui cessent d'être significatifs, on trouve nombre de mots dont on a toutes les raisons de penser qu'ils sont typiques d'une variété d'anglais comme *center*, *behaviour* et *color* pour l'anglais américain, *caste* et *upto* pour l'anglais indien ou encore *Aborigines* pour l'anglais australien. Cette observation confirme simplement un résultat classique : être (beaucoup trop) rigoureux sur le taux d'erreur de type 1 accroît dramatiquement le taux d'erreur de type 2, c'est-à-dire la probabilité d'accepter l'hypothèse nulle alors qu'elle est fausse. Le test perd toute puissance, l'analyse tout intérêt.

Une deuxième solution consisterait à employer un test inférentiel adapté à la véritable unité d'échantillonnage, à savoir le texte. C'est l'approche proposée par Kilgarriff (1996) : utiliser le test non paramétrique de Wilcoxon-Mann-Whitney qui porte, non sur la fréquence totale d'un mot dans chaque corpus, mais sur la fréquence (transformée en rang) du mot dans chaque texte de chaque corpus. Cette proposition n'a pas connu de succès en linguistique de corpus (mais voir Paquot et Bestgen (2009) pour une comparaison avec le test du Khi-2), et a été vivement critiquée par Rayson (Rayson *et al.*, 2004 ; Rayson et Garside, 2000) parce que ce test ne permettrait que l'analyse des mots les plus fréquents et parce qu'il néglige une part importante des informations disponibles en raison de la transformation des fréquences en rangs.

Une troisième solution, plus catégorique, est possible. Elle consiste à considérer l'utilisation du test du Khi-2 pour analyser les différences entre des corpus pour ce à quoi il est réellement employé, un outil exploratoire, relevant de la statistique descriptive, visant à identifier des mots

4 Comme on souhaite comparer la valeur critique aux valeurs réelles obtenues par Oakes et Farrow, on a employé pour cette analyse la formule standard.

qui méritent une analyse approfondie, et non pour un instrument d'analyse confirmatoire, dont l'objectif est d'accepter ou de rejeter des hypothèses spécifiques. Les Khi-2 obtenus (ou les résidus standardisés) sont alors employés comme des indicateurs de l'intérêt potentiel d'une différence entre les corpus. Le recours à l'analyse des correspondances (AC), qui, comme on le sait, est basée sur la distance du Khi-2 (et non le test), trouve dans ce cadre tout son intérêt puisque cette technique a justement pour objectif premier de proposer une représentation graphique des associations sous-jacentes à une table de contingence.

Cette solution n'est en rien neuve. Elle a été proposée et est mise en pratique depuis de très nombreuses années en analyses statistiques des données textuelles. Elle semble, par contre, éprouver beaucoup plus de difficultés pour trouver sa place en linguistique de corpus. L'analyse suivante en souligne, une fois de plus, tout l'intérêt en l'appliquant aux données de Oakes et Farrow (2007). Si on soumet leur table de contingence à une AC, les quatre premiers facteurs obtenus distinguent très nettement les 5 corpus. De plus, les mots, qui contribuent le plus à la détermination de ces dimensions sont ceux qui sont pointés par Oakes et Farrow comme les plus typiques des différents corpus. Ces résultats reproduisent donc leurs observations sans devoir se baser sur des statistiques inférentielles discutables. Plus intéressante encore est l'AC réalisée sur les corpus divisés en cinq sections correspondant à celles employées par Oakes et Farrow pour leur analyse de la dispersion. Les figures 1 et 2 montrent le positionnement de ces sections sur les 4 premiers axes[5]. Sur ceux-ci, les corpus sont identifiés par le code du pays (AU = Australie...) et les sections par le chiffre qui suit. Si les axes 3 et 4 (qui expliquent ensemble 13% de l'inertie) mettent en évidence des oppositions géographiques entre les corpus, les deux premiers axes (qui expliquent ensemble 36% de l'inertie) soulignent les différences entre les sections qui composent chaque corpus, celles-ci correspondant à des genres de textes différents (p.e., les sections 5 de tous les corpus sont composées de textes littéraires). Comme on le voit, l'AC permet la mise en évidence simultanée de différents principes d'organisation des données, un résultat pour le moins difficile à atteindre avec le test du Khi-2 tel qu'il est habituellement employé en linguistique de corpus.

5 Ces graphiques ont été produits par le logiciel académique DTM-Vic (http://www.dtmvic.com/).

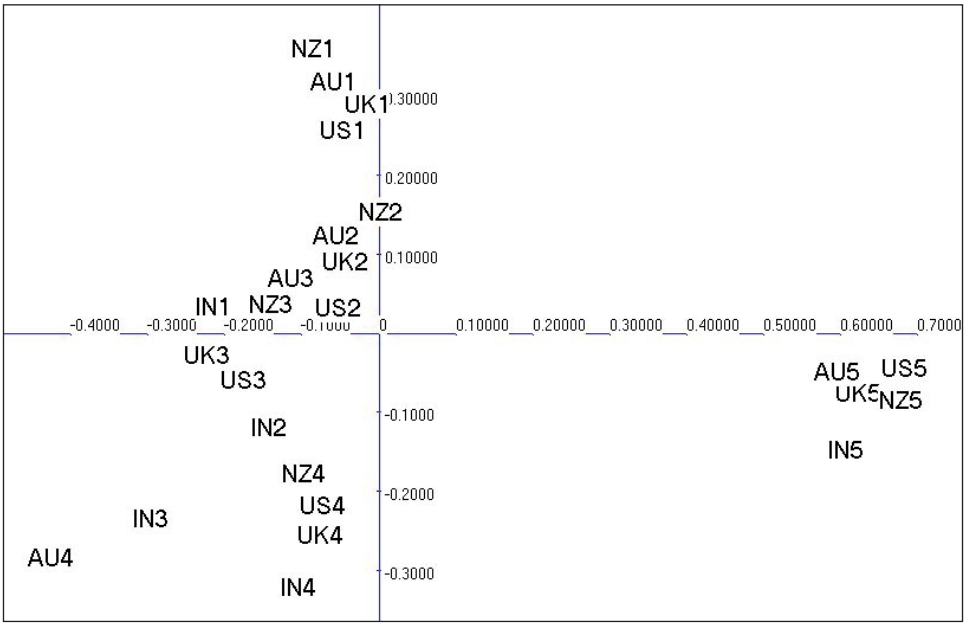

Figure 1 : Positionnement des sections des corpus sur les axes 1 et 2 (AC)

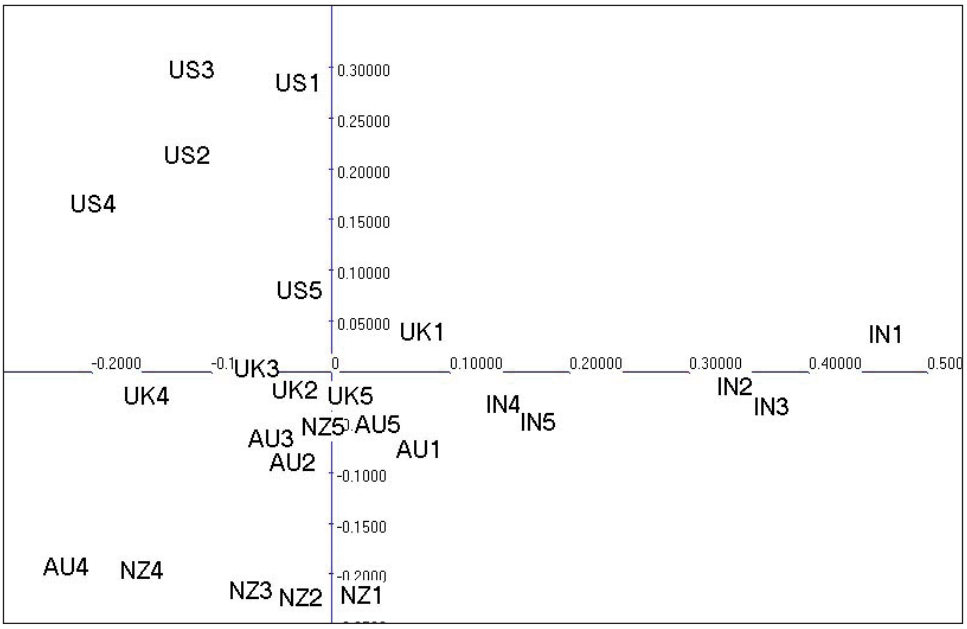

Figure 2 : Positionnement des sections des corpus sur les axes 3 et 4 (AC)

References

Agresti A. (1990). *Categorical data analysis.* New York: John Wiley.

Baker P. (2004). Querying keywords: questions of difference, frequency and sense in keyword analysis. *Journal of English Linguistics* 32: 346–359.

Baroni M. and Evert S. (2009). Statistical methods for corpus exploitation. In A. Lüdeling and M. Kytö (eds), *Corpus Linguistics. An International Handbook* (Article 36, pp. 777-803). Mouton de Gruyter..

Bibby J. (1978). Reviewed Work(s): The Analysis of Contingency Tables by B.S. Everitt. *The Mathematical Gazette*, 62: 138-139.

Bondi M. and Scott M. (2010). *Keyness in texts*. John Benjamins.

Church K. (2000). Empirical Estimates of Adaptation: The Chance of Two Noriegas is Closer to p/2 than p2. In *Proceedings of the 17th Conference on Computational Linguistics*, pp. 180-186.

Evert S. (2006). How Random is a Corpus? The Library Metaphor. *Zeitschrift für Anglistik und Amerikanistik*, 54: 177-190.

García-pérez, M.A and Núñez-antón, V. (2003). Cellwise Residual Analysis in Two-Way Contingency Tables, *Educational and Psychological Measurement*, 63: 825-839.

Gries S. (2005). Discussion Note: Null hypothesis significance testing of word frequencies: a follow-up on Kilgarriff, *Corpus Linguistics and Linguistic Theory*, 1: 277-294.

Groom N. (2010). Closed-class keywords and corpus-driven discourse analysis. In M. Bondi and M. Scot (Eds.), *Keyness in texts* (pp. 59-78). John Benjamins.

Haberman S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, *29*, 205-220.

Hofland K. and Johansson, S. (1982). Word frequencies in British and American English. The Norwegian Computing Centre for the Humanities, Bergen, Norway

Hofland K., Lindebjerg A. and Thunestvedt, J. (1999). ICAME Collection of English Language Corpora. The HIT Centre, University of Bergen, Norway.

Howell D. (2008). *Méthodes statistiques en sciences humaines*. De Boeck Université.

Kilgarriff A. (1996). Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved LOB-Brown comparison. In *Proc. ALLC-ACH Conference*, pp. 169-172.

Kilgarriff A. (2005). Language is never, ever, ever random, *Corpus Linguistics and Linguistic Theory*, 1: 263-275.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1: 127-165.

Lebart L., Piron M. et Steiner J.-F. (2003). *La Sémiométrie*. Dunod.

Leech G.. and Fallon R. (1992). Computer corpora: what do they tell us about culture? *ICAME Journa*l, 16: 1-22.

Oakes M. and Farrow M. (2007). Use of the Chi-Squared Test to examine vocabulary differences in English language corpora representing seven different countries. *Literary and Linguistic Computing*, 22: 85-99.

Paquot M., & Bestgen Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In A.H. Jucker, D. Schreier and M. Hundt (Eds), *Corpora: Pragmatics and Discourse* (pp. 247-269). Rodopi.

Rayson P., Berridge D. and Francis B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. *Proceedings of the 7th International Conference on Statistical analysis of textual data*, pp. 926-936.

Rayson P. and Garside R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora*, pp. 1-6.

Rayson P., Leech G. and Hodges M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus, *International Journal of Corpus Linguistics*, 2: 133-152.

Scott M. (1999). WordSmith Tools Help Manual. Version 3.0. Mike Scott and Oxford University Press.

Tribble C. (2000). Genres, keywords, teaching: towards a pedagogic account of the language of project proposals. In L. Burnard and T. McEnery (eds), *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora* (pp. 75-90). Peter Lang.