Évaluation de mesures d'association pour les bigrammes et les trigrammes au moyen du test exact de Fisher

Yves Bestgen
CECL, Place du Cardinal Mercier 10, B-1348 Louvain-la-Neuve, Belgique
yves.bestgen@uclouvain.be

RÉSUMÉ

Pour déterminer si certaines mesures d'association lexicale fréquemment employées en TAL attribuent des scores élevés à des n-grammes que le hasard aurait pu produire aussi souvent qu'observé, nous avons utilisé une extension du test exact de Fisher à des séquences de plus de deux mots. Les analyses ont porté sur un corpus de quatre millions de mots d'anglais conversationnel extrait du BNC. Les résultats, basés sur la courbe précision-rappel et sur la précision moyenne, montrent que le LL-simple est extrêmement efficace. IM3 est plus efficace que les autres mesures basées sur les tests d'hypothèse et atteint même un niveau de performance presque égal à LL-simple pour les trigrammes.

ABSTRACT

Using Fisher's Exact Test to Evaluate Association Measures for Bigrams and Trigrams

To determine whether some often-used lexical association measures assign high scores to n-grams that chance could have produced as frequently as observed, we used an extension of Fisher's exact test to sequences longer than two words to analyse a corpus of four million words. The results, based on the precision-recall curve and the average precision, show that simple-ll is extremely effective. MI3 is more efficient than the other hypothesis tests-based measures and even reaches a performance level almost equal to simple-ll for trigrams.

MOTS-CLÉS : Mesures d'association lexicale ; N-grammes de mots ; LL-simple ; IM3.

KEYWORDS: Lexical association measures ; Word n-grams ; Simple-LL ; MI3.

1 Introduction

L'évaluation de l'efficacité des mesures d'association lexicale pour identifier les expressions polylexicales retient l'attention de nombreux chercheurs depuis plus de vingt ans (Church & Hanks, 1990 ; Evert & Krenn, 2005 ; Ramish, 2015). Dans le cas de bigrammes, un grand nombre de mesures ont été comparées et plusieurs synthèses sont disponibles (Evert, 2008 ; Pecina, 2010). Les mesures d'association pour des n-grammes de plus de deux mots ont retenu beaucoup moins l'attention comme l'ont souligné Evert (2008), Gries (2010) et Nerima *et al.* (2010).

Cette situation a probablement conduit certains chercheurs à privilégier la fréquence de cooccurrence brute comme critère de sélection (Wermter & Hahn, 2006 ; Biber, 2009). Dans le cas des bigrammes, cependant, il est bien établi que la fréquence seule est insuffisante pour distinguer les séquences intéressantes de celles que le hasard aurait pu produire en aussi grand nombre qu'observé dans un corpus (Evert, 2008 ; Gries, 2010). Selon Ellis *et al* (2015), la fréquence devrait également être peu

efficace pour identifier des séquences de plus de deux mots. Pourtant, plus une séquence est longue, moins il est probable qu'elle se produise par hasard, même si elle est composée de mots fréquents. La question est donc de savoir quelle longueur doit avoir une séquence pour s'assurer que la fréquence brute ne sélectionne que les séquences qui ne sont probablement pas dues au hasard.

Cette question est en fait bien plus générale. Elle se pose à propos de presque toutes les mesures d'association puisque celles-ci visent principalement à mettre en évidence des séquences dont la force d'association est suffisamment supérieure à ce que le hasard pourrait produire. Si cela est évident pour les mesures d'association basées sur des tests d'hypothèses, tels que *z* ou *LL-simple*, c'est également vrai pour des mesures de taille d'effet, telles que *IM* et ses variantes, car ces mesures comparent également les fréquences observées aux fréquences attendues par le seul effet du hasard, même si elles ne dérivent pas d'un test statistique. La comparaison de la fréquence aux mesures d'association est d'autant plus importante que Biber (2009) a formulé des critiques très fortes contre l'utilisation des mesures d'association pour sélectionner les n-grammes les plus intéressants dans un corpus.

Pour essayer de répondre à cette question de l'efficacité des mesures d'association pour identifier des séquences de plus de deux mots, cette recherche propose d'employer le test exact de Fisher, le test inférentiel considéré comme le plus adéquat pour analyser la fréquence de bigrammes dans un corpus (Evert, 2008 ; Moore, 2004 ; Pedersen, 1996). Il est important de noter que cette étude vise uniquement à développer un point de vue complémentaire sur l'efficacité de certaines mesures d'association. En aucun cas, l'approche proposée ne pourra remplacer une évaluation fondée sur une liste étalon établie manuellement ou sur les bénéfices que les expressions polylexicales extraites automatiquement peuvent apporter à des applications du TAL.

La section suivante décrit l'approche employée pour étendre le test de Fisher à des n-grammes de plus de deux mots. Nous présentons ensuite une expérience, menée sur un corpus de quatre millions de mots, qui analyse les performances de six mesures d'association pour extraire des séquences de deux à quatre mots, les bigrammes servant de point de repère auquel les autres séquences sont comparées.

2 Étendre le test exact de Fisher à des n-grammes de plus de deux mots

Le test exact de Fisher est considéré comme le test inférentiel le plus adéquat pour calculer la probabilité que le hasard seul a de produire au moins autant d'occurrences d'un bigramme que le nombre effectivement observé dans un corpus (Evert, 2008 ; Jones & Sinclair, 1974 ; Moore, 2004 ; Pedersen *et al.*, 1996). La proposition de Fisher est de considérer l'ensemble des tables de contingence qu'il est possible de construire en respectant les totaux marginaux réellement obtenus et de déterminer la proportion de celles-ci qui donne lieu à un résultat au moins aussi extrême que celui observé, une table plus extrême étant une table dans laquelle la fréquence du bigramme est plus élevée que celle observée. La formule de calcul est basée sur la distribution hypergéométrique (Evert, 2004).

Son utilisation pour analyser des séquences de mots plus longues est toutefois problématique parce que cela implique d'analyser des tables de contingence à trois dimensions ou plus et que les procédures de calculs exacts pour celles-ci (Zelterman *et al.*, 1995) ne sont pas adaptées à l'étude des séquences de mots, en raison de la taille de l'échantillon et du grand nombre de tests à effectuer. Le tableau 1 présente, à gauche, la table de contingence obtenue dans le corpus qui sera analysé ci-après (5 529 378 formes graphiques) pour le trigramme *a_lot_of*. À droite, on y trouve une des très nombreuses tables

Mot_1	Mot_2	Mot_3	
		of	¬of
a	lot	1723	1357
	¬lot	5146	69102
¬a	lot	179	737
	¬lot	35244	5415890

Mot_1	Mot_2	Mot_3	
		of	¬of
a	lot	3996	0
	¬lot	19148	54184
¬a	lot	0	0
	¬lot	19148	5432902

TABLE 1 – Tables de contingence pour le trigramme *a_lot_of*

de contingence les plus extrêmes possible qui respectent l'effectif total et les totaux marginaux univariés (la fréquence totale de chaque mot). Ces tables les plus extrêmes ne sont évidemment qu'une infime fraction de toutes les tables possibles qui respectent les mêmes conditions. À titre d'exemple, Mielke, Berry et Zelterman (1994) ont calculé qu'il y a plus de 3,6 milliards de tables différentes compatibles avec une table de contingence $2\times2\times2$ dont l'effectif total est 1663 et les totaux marginaux 624, 623, et 384, soit des valeurs incomparablement plus petites que celles qui doivent être analysées pour chaque trigramme et quadrigramme du corpus.

Bestgen (2014) a proposé d'appliquer aux séquences de plus de deux mots la procédure approchée recommandée pour estimer la probabilité du test exact de Fisher lorsque l'utilisation de la formule hypergéométrique n'est pas possible (Agresti, 1992 ; Pedersen, 1996). Cette méthode repose sur une procédure de permutation de type Monte-Carlo qui génère un échantillon aléatoire des tables de contingence possibles, compte tenu des totaux marginaux, et estime la probabilité exacte sur la base de la proportion de ces tables qui produisent une valeur au moins aussi extrême que celle observée. Étant donné qu'un corpus est une longue séquence de formes graphiques, une table de contingence possible est simplement une permutation aléatoire de l'ensemble de ces formes. Toute permutation de ces formes génère donc une table de contingence pour chacun des bigrammes originaux, résolvant le problème du grand nombre de tests à effectuer. Cette procédure d'estimation peut être directement généralisée à des n-grammes plus longs en comptant ceux-ci dans chaque permutation.

Bestgen (2014) a montré que cette procédure permet une estimation presque parfaite de la probabilité du test de Fisher pour tous les bigrammes dans un corpus, avec la limitation que le nombre de permutations effectuées détermine la précision de la probabilité, celle-ci ne pouvant être plus petite que 1 divisé par le nombre de permutations effectuées. La principale faiblesse de la procédure est donc son coût-calcul.

3 Expérience

Cette expérience vise à évaluer, au moyen du test de Fisher, l'efficacité de six mesures d'association qui ont été proposées pour extraire des expressions polylexicales de corpus.

3.1 Matériel et procédure

3.1.1 Corpus

Les analyses ont été menées sur un corpus de 5 529 378 formes graphiques, dont 4 232 259 mots de conversations spontanées extraits du BNC. Ce corpus, similaire à ceux utilisés dans plusieurs études

sur les paquets lexicaux (Biber, 2009), a été choisi, car il est suffisamment grand pour extraire des expressions composées de plusieurs mots, mais pas trop grand de sorte qu'un nombre suffisant de permutations puisse être obtenu en un temps raisonnable.

3.1.2 Procédure

La version du corpus lemmatisée par CLAWS a été utilisée. Toutes les formes graphiques (mots, nombres, symboles et ponctuations) détectées par CLAWS (http ://ucrel.lancs.ac.uk/claws/) ont été considérées comme des éléments à permuter aléatoirement. Vingt millions de permutations ont été effectuées ; la plus petite probabilité qui peut être estimée est donc de 0,00000005. Dans les résultats, seules les chaînes de mots ininterrompues sont prises en compte. Étant donné que les mesures d'association calculées pour des n-grammes très rares ne sont pas fiables (Evert, 2008), nous avons analysé uniquement les séquences qui apparaissent au moins trois fois dans le corpus original.

3.1.3 Mesures d'association évaluées

Pour ces analyses, une série de mesures d'association ont été sélectionnées en fonction de leur popularité dans les études des bigrammes et des séquences plus longues : *IM* (ou information mutuelle ponctuelle) et *t*-score (Church *et al.*, 1991), *z* (Berry-Rogghe, 1973), *LL-simple* (Evert, 2008) et la fréquence brute (*f*). Nous avons également analysé *IM3*, une modification heuristique de *IM*, proposée pour réduire la tendance de celle-ci à attribuer des scores importants à des mots rares qui ne s'observent ensemble que quelquefois (Daille, 1994). Cependant, comme *IM3* viole une convention des mesures d'association en attribuant, dans certains cas, des scores positifs aux n-grammes qui se produisent moins fréquemment que le hasard le prédit (Evert, 2008, p. 1226), nous avons modifié la formule de telle sorte que ces n-grammes reçoivent un score négatif et, par conséquent, apparaissent à la fin de la liste [1]. Un problème similaire se pose pour *LL-simple* ; nous avons utilisé la version signée recommandée par Evert (2008). Les formules de ces indices pour les trigrammes et les quadrigrammes sont identiques à celles utilisées pour les bigrammes, la fréquence attendue étant calculée en utilisant l'extension habituelle de la formule pour les bigrammes (c.-à-d., en multipliant le produit des probabilités marginales de tous les mots de la séquence par la taille du corpus (Ramish, 2015, p. 64)).

4 Analyses et résultats

La principale question de recherche à laquelle cette étude tente de répondre est de déterminer si certaines mesures d'association attribuent des scores élevés à des n-grammes dont la fréquence de cooccurrence est susceptible de résulter du seul effet du hasard. Le seuil de signification a été fixé à 0,05, ce qui implique qu'on accepte le risque qu'une séquence soit déclarée erronément statistiquement significative (donc, trop fréquente) dans 5 % des cas. La procédure séquentielle de Holm (Holm, 1979 ; Howell, 2008, p. 370–371) a été utilisée pour tenir compte du grand nombre de tests effectués qui augmente la probabilité de décider de manière erronée qu'au moins un test est

1. Plusieurs modifications ont été évaluées, mais comme elles ont donné lieu à des performances presque identiques, nous avons employé la formule de IM pour tout n-gramme observé moins souvent que le hasard le prédit.

	Bigramme	Trigramme	Quadrigramme
Nbr. de n-grammes différents	96881	111164	53334
Nbr. de n-grammes significatifs	27063	70942	52825
PM du niveau de base	0.279	0.638	0.990

TABLE 2 – Statistiques pour les trois longueurs de n-gramme

statistiquement significatif. Cette procédure assure un taux d'erreur d'ensemble de 0,05 pour chaque longueur de n-grammes, autorisant des comparaisons entre celles-ci.

Étant donné que la liste des n-grammes pour lesquels le test de Fisher indique qu'ils ne sont pas dus au hasard est considérée comme la référence à atteindre par les autres mesures, les résultats ont été résumés de la manière habituelle (Biemann *et al.*, 2013, p. 40–41) par des courbes précision-rappel (PR) et par la précision moyenne (PM, *average precision*). Comme niveau de base, nous avons utilisé la précision moyenne atteinte par une mesure d'association qui classe les n-grammes au hasard (Bestgen, 2015). Le nombre de n-grammes testés, le nombre de n-grammes significatifs pour le test de Fisher et la PM du niveau de base sont donnés dans le tableau 1. On constate que presque tous les quadrigrammes sont considérés par le test inférentiel comme étant statistiquement significatifs, mais aussi que le niveau de base est beaucoup plus élevé pour les trigrammes que pour les bigrammes.

Ces résultats ont deux implications pour les analyses présentées dans la suite. Tout d'abord, celles-ci n'ont été effectuées que sur les bigrammes et les trigrammes puisque presque tous les quadrigrammes présents dans le corpus au moins trois fois passent ce test inférentiel avec succès. Cette observation répond donc à la première question posée dans cette étude : dès quatre mots, une séquence est suffisamment longue pour garantir que la fréquence brute ne sélectionne (quasiment) que des séquences qui ne sont probablement pas dues au hasard. Toutefois, ce résultat ne démontre pas qu'ordonner les séquences sur la base de la seule fréquence est une procédure efficace pour extraire les meilleures expressions composées de plusieurs mots, les quadrigrammes arrivant en tête de liste n'étant pas nécessairement les meilleurs pour un annotateur humain.

La deuxième conséquence concerne la comparaison des performances des mesures d'association pour les bigrammes et les trigrammes puisqu'une meilleure performance est attendue dans le second cas en raison du niveau plus élevé atteint par le hasard. Pour cette raison, en plus de la PM standard, nous rapportons un nouvel indice appelé *Précision moyenne corrigée pour le hasard* (PMch) obtenu en insérant dans la formule classique du Kappa (Cohen, 1960 ; Powers, 2012) la PM du niveau de base, puisqu'il s'agit du niveau de performance que devrait produire un classement aléatoire des n-grammes. La formule de cet indice est donc : $(PMo - PMnb)/(1 - PMnb)$ dans laquelle PMo représente la PM observée et $PMnb$ celle du niveau de base.

Les figures 1 et 2 présentent les courbes PR. Pour chaque mesure, la ligne commence au premier n-gramme qui ne passe pas le test inférentiel. Jusqu'à ce point, la précision est donc de 100 %. Pour faciliter la comparaison des figures, le point de départ de l'axe des *y* a été fixé à une valeur juste inférieure à la PM du niveau de base. Le tableau 2 rapporte les PM et les PM corrigées pour le hasard.

Les courbes PR pour les bigrammes indiquent que *LL-simple* est de loin la meilleure mesure d'association, atteignant un niveau de performance presque parfait. Elle est suivie par *IM3*, *z* et *t* [2]. Comme prévu, la fréquence brute est nettement moins efficace, sa courbe PR montrant qu'elle classe en tête

2. La forme distinctive de la courbe PR pour *t* est due à sa formule qui attribue des scores presque identiques aux n-grammes qui se produisent dans le corpus à la même faible fréquence et qui ont une fréquence attendue proche de 0. Ces n-grammes étant considérés comme trop fréquents par le test inférentiel, ils forment une longue série de scores PR légèrement croissants.

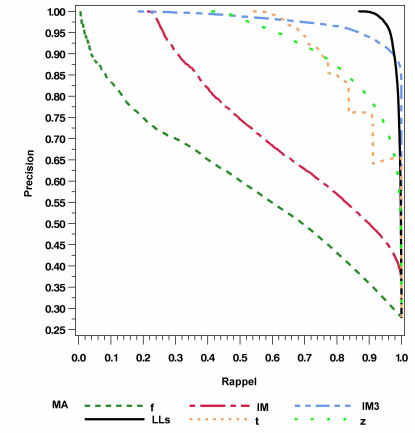

FIGURE 1 – Courbes PR pour les bigrammes

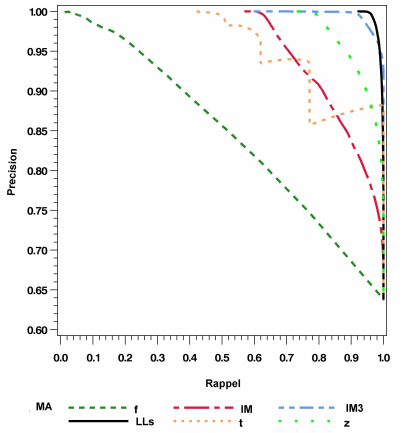

FIGURE 2 – Courbes PR pour les trigrammes

	PM		PMch	
Mesures d'association	Bigramme	Trigramme	Bigramme	Trigramme
f	0.588	0.824	0.429	0.515
IM	0.756	0.957	0.662	0.881
IM3	0.947	0.998	0.926	0.994
z	0.934	0.985	0.909	0.959
t	0.931	0.959	0.905	0.886
LL-simple	0.993	0.998	0.990	0.995

TABLE 3 – PM et PM corrigée pour le hasard (PMch)

de liste plusieurs n-grammes que le hasard pourrait avoir produit aussi fréquemment qu'observé.

Pour les trigrammes, la différence entre *LL-simple* et *IM3* est extrêmement faible, comme le confirme la PM. Ce résultat est inattendu. La performance de *Z* est légèrement moins bonne. *IM* est presque aussi efficace que *t*, un autre résultat inattendu. La fréquence brute est à nouveau la moins efficace.

La comparaison des PM corrigés des bigrammes et des trigrammes montre que *t* ne s'améliore pas du tout alors que *IM*, *IM3* et *z* s'améliorent nettement.

Un objectif de ces analyses est de déterminer si les mesures d'association classent en début de liste des n-grammes que le hasard aurait pu produire aussi souvent qu'observé. Il est donc intéressant d'analyser la position dans le classement du premier n-gramme rejeté par le test inférentiel. Il correspond dans les graphiques au point le plus à gauche des courbes PR. Il est fourni en termes de rappel et doit donc être interprété en fonction du nombre de n-grammes significatifs. Pour *LL-simple*, le premier bigramme non significatif est observé lorsque le rappel atteint 87 %, soit après le 23000e bigramme dans la liste. Pour les trigrammes, c'est au-delà de la 65000e position. Ces valeurs sont clairement meilleures que celles obtenues par les autres mesures d'association, ce qui suggère qu'il pourrait être utile de filtrer les listes de ces autres mesures d'association sur cette base.

5 Discussion et conclusion

L'objectif de cette étude est de fournir une perspective différente sur l'efficacité des mesures d'association lexicales pour l'extraction d'expressions polylexicales à partir de corpus. Les analyses mettent en évidence des différences importantes entre des mesures d'association basées sur des tests d'hypothèse. La performance presque parfaite de *LL-simple* est aussi observée pour les trigrammes, étendant à ceux-ci la conclusion d'Evert (2008) selon laquelle *LL-simple* est une bonne approximation du test exact de Fisher pour les bigrammes. *Z* est moins efficace. Le score *t* l'est encore beaucoup moins et ne s'améliore pas lorsqu'on passe des bigrammes aux trigrammes. En ce qui concerne les mesures de la taille de l'effet, *IM3* [3] est nettement plus efficace que *IM* et, pour les trigrammes, atteint une performance équivalente à *LL-simple*. Il mérite certainement une évaluation approfondie.

Le critère utilisé (c.-à-d., une mesure d'association attribue-t-elle des scores élevés à des n-grammes que le hasard aurait pu produire aussi souvent qu'observé) suggère qu'il pourrait être judicieux d'utiliser *LL-simple* pour filtrer les classements générés par d'autres mesures d'association, l'extension du test de Fisher ne pouvant être employée en raison de son coût-calcul. Cependant, avant de recommander une telle procédure, il est nécessaire de s'assurer qu'elle améliore effectivement les performances des mesures d'association dans des applications du TAL et que les séquences retenues ainsi sont considérées comme de véritables expressions par un évaluateur humain.

Une des limitations de cette étude est qu'elle a été menée sur un seul corpus, composé exclusivement de conversations et d'une taille limitée. Si l'analyse d'un corpus plus grand augmente considérablement le temps nécessaire pour obtenir les permutations, il pourrait être intéressant d'analyser des corpus plus petits, aussi semblables que possible à celui utilisé ici, afin d'estimer au moins partiellement l'impact du facteur taille.

Il serait aussi intéressant, lors de la sélection initiale des n-grammes à évaluer, de prendre en compte les catégories morphosyntaxiques des mots qui composent les séquences (Benigno & Kraif, 2016). Il est en effet possible que les performances de la fréquence brute, mais aussi de certaines mesures d'association, s'améliorent nettement si des profils syntaxiques spécifiques, comme *adjectif + préposition + nom*, sont testés.

Enfin, il faut rappeler la principale limitation de l'extension du test de Fisher aux séquences de plus de deux mots : son coût-calcul. La procédure de Monte-Carlo limite fortement la précision des probabilités les plus faibles, produisant donc un grand nombre de séquences obtenant le meilleur score (la plus faible probabilité estimable). Cette extension ne constitue donc pas un indice d'association viable, permettant de classer les séquences de la plus exceptionnelle à la moins exceptionnelle.

Remerciements

Cette recherche a bénéficié du soutien du Fonds de la recherche scientifique (crédit F.R.S.-FNRS J.0025.16) dont l'auteur est chercheur qualifié. Une partie des ressources informatiques utilisées ont été fournies par les installations de calcul intensif de l'Université catholique de Louvain (CISM/UCL) et du Consortium des Équipements de Calcul intensif en Fédération Wallonie Bruxelles (CÉCI) financé par le F.R.S.-FNRS.

3. Des analyses effectuées en utilisant la formule non corrigée ont indiqué que la correction n'améliore les performances que pour les bigrammes.

Références

AGRESTI A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, **7**, 131–153.

BENIGNO V. & KRAIF O. (2016). Core vocabulary and core collocations : combining corpus analysis and native speaker judgement to inform selection of collocations in learner dictionaries. In A. ORLANDI & L. GIACOMINI, Eds., *Defining Collocations For Lexicographic Purposes : From Linguistic Theory To Lexicographic Practice (Vol. 1)*, p. 235–268. Peter Lang.

BERRY-ROGGHE G. L. M. (1973). The computation of collocations and their relevance in lexical studies. In A. J. AITKEN, R. W. BAILEY & N. HAMILTON-SMITH, Eds., *The Computer and Literary Studies*. Edinburgh University Press.

BESTGEN Y. (2014). Extraction automatique de collocations : Peut-on étendre le test exact de Fisher à des séquences de plus de 2 mots ? In *Actes de JADT 2014*, p. 79–90.

BESTGEN Y. (2015). Exact expected average precision of the random baseline for system evaluation. *The Prague Bulletin of Mathematical Linguistics*, **103**, 131–138.

BIBER D. (2009). A corpus-driven approach to formulaic language in English : Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, **14**, 275–311.

BIEMANN C., BILDHAUER F., EVERT S., GOLDHAHN D., QUASTHOFF U., SCHÄFER R., SIMON J., SWIEZINSKI L. & ZESCH T. (2013). Scalable construction of high-quality Web corpora. *Journal for Language Technology and Computational Linguistics*, **28**, 23–59.

CHURCH K., GALE W. A., HANKS P. & HINDLE D. (1991). Using statistics in lexical analysis. In U. ZERNIK, Ed., *Lexical Acquisition : Using On-line Resources to Build a Lexicon*, p. 115–164. Lawrence Erlbaum.

CHURCH K. & HANKS P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, p. 22–29.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

DAILLE B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7.

ELLIS N. C., SIMPSON-VLACH R., RÖMER U., O'DONNELL M. & WULFF S. (2015). Learner corpora and formulaic language in SLA. In S. GRANGER, G. GILQUIN & F. MEUNIER, Eds., *Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

EVERT S. (2008). Corpora and collocations. In A. LÜDELING & M. KYTÖ, Eds., *Corpus Linguistics. An International Handbook*, p. 1211–1248. Mouton de Gruyter.

EVERT S. & KRENN B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, **19**, 450–466.

GRIES S. T. (2010). Useful statistics for corpus linguistics. In A. SÁNCHEZ & M. ALMELA, Eds., *A Mosaic of Corpus Linguistics : Selected Approaches*, p. 269–291. Frankfurt au Main, Germany : Peter Lang.

HOLM S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

HOWELL D. (2008). *Méthodes statistiques en sciences humaines*. Bruxelles : De Boeck Université.

JONES S. & SINCLAIR. J. (1974). English lexical collocations. a study in computational linguistics. *Cahiers de Lexicologie*, **24**, 15–61.

MIELKE P. M., BERRY K. J. & ZELTERMAN D. (1994). Fisher's exact test of mutual independence for 2×2×2 cross-classification tables. *Educational and Psychological Measurement*, **54**, 110–114.

MOORE R. C. (2004). On log-likelihood-ratios and the significance of rare events. In *Proceedings of EMNLP 2004*, p. 333–340.

NERIMA L., WEHRLI E. & SERETAN V. (2010). A recursive treatment of collocations. In *Proceedings of LREC 2010*, p. 634–638.

PECINA P. (2010). Lexical association measures and collocation extraction. *Language Resources & Evaluation*, **44**, 137–158.

PEDERSEN T. (1996). Fishing for exactness. In *Proceedings of the South Central SAS Users Group*, p. 188–200.

PEDERSEN T., KAYAALP M. & BRUCE R. (1996). Significant lexical relationships. In *Proceedings of the 13th National Conference on Artificial Intelligence*, p. 455–460.

POWERS D. M. W. (2012). The problem of kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, p. 345–355.

RAMISH C. (2015). *Multiword Expressions Acquisition : A Generic and Open Framework*. Springer.

WERMTER J. & HAHN U. (2006). You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, p. 785–792.

ZELTERMAN D., CHAN I. S. & MIELKE P. W. (1995). Exact tests of significance in higher dimensional tables. *The American Statistician*, **49**, 357–361.