Collocations et linguistique de corpus : l'intuition des linguistes et les critères quantitatifs convergent-ils ?

Olivier Kraif, Agnès Tutin

Version préliminaire

Collocations et linguistique de corpus : l'intuition des linguistes et les critères quantitatifs convergent-ils ?

Olivier KRAIF, Agnès TUTIN

1. Introduction

Phraséologie et linguistique de corpus sont étroitement liées dans les études linguistiques contemporaines, qui semblent ne pouvoir se passer l'une de l'autre. Les travaux de lexicologie et plus encore de lexicographie sont largement basés sur des extractions partiellement automatisées réalisées à partir de corpus, qui exploitent des méthodes quantitatives. Ces méthodes semblent aller de soi alors qu'il conviendrait de réexaminer selon nous les liens entre ces deux champs d'études, en abordant plus précisément à travers les angles qualitatif et quantitatif une question qui a été peu étudiée :

- Les critères quantitatifs confirment-ils les repérages lexicologiques effectués dans les textes par les experts linguistes ? Autrement dit, les méthodes quantitatives apparaissent-elles toujours adaptées pour extraire à grande échelle des ensembles d'expressions polylexicales ? Par ailleurs, certains repérages peuvent-ils se traduire autrement que par des critères quantitatifs ?

Dans cet article, nous aborderons principalement un type d'expression polylexicale aux contours un peu flous, les collocations, parce qu'elles apparaissent centrales dans la maîtrise de la langue, en particulier dans le cadre de la production en langue étrangère. Ces expressions, qui relèvent de la préfabrication linguistique, ne peuvent pas être considérées comme sémantiquement figées. Il s'agit pour nous d'associations binaires privilégiées (p.e. *grièvement blessé*, *remercier infiniment*, *faire une hypothèse*) dont les éléments entretiennent une relation syntaxique.

Parce qu'elles ne sont pas complètement figées, contrairement à des expressions polylexicales comme *tenir compte*, *bien que*, *autant que faire se peut*, elles posent des problèmes d'identification complexes pour les linguistes. Elles n'en sont pas moins des éléments indispensables dans la maîtrise de la langue, comme en attestent les travaux en acquisition des langues étrangères (par exemple, Nesselhauf 2004 ou Cavalla 2014) ou les nombreuses ressources élaborées récemment, principalement à l'attention des apprenants avancés des langues étrangères (*Oxford collocations Dictionary for Students of English*, *Dictionnaire des combinaisons de mots* de l'éditeur Le Robert ou le *Dictionnaire des cooccurrences* du logiciel Antidote).

Dans cet article, nous faisons l'hypothèse que le phénomène collocationnel échappe à une approche qui serait purement qualitative, au sens usuel où l'on entend les méthodes qualitatives en linguistique. Pour illustrer cette position intermédiaire des collocations dans l'articulation de l'opposition méthodologique entre qualitatif et quantitatif, nous proposons de mettre en œuvre une étude empirique autour d'une tâche de repérage manuel, permettant de confronter le jugement intersubjectif de quelques linguistes experts avec les données quantitatives issues du comptage des fréquences d'occurrence et de cooccurrence.

Dans un premier temps, nous ferons une synthèse de quelques travaux autour des expressions polylexicales en linguistique de corpus et présenterons quelques études qui ont examiné les liens entre mesures quantitatives et expertise linguistique. Dans un second temps, nous présenterons la méthodologie quantitative d'extraction des collocations et les tests effectués par des experts linguistes. Dans un troisième temps, nous comparerons ces résultats et en ferons une analyse.

2. Expressions polylexicales et linguistique de corpus

La phraséologie et la linguistique de corpus entretiennent des rapports anciens et continus, en particulier dans le domaine de la lexicographie. Les premières méthodes exploitant de grandes ressources textuelles pour extraire automatiquement un ensemble d'expressions apparaissent dès les années 70 en Grande Bretagne (travaux de l'OSTI, cf. Krishnamurthy 2005) et en France, avec l'utilisation du corpus Frantext pour l'extraction des « groupes binaires » à l'aide de méthodes statistiques comme la loi de Poisson (Gorcy *et al*. 1970). La linguistique contextualiste anglaise de Sinclair (1991) s'intéresse beaucoup aux phénomènes de cooccurrence lexicale, où le paramètre de la fréquence dans les corpus apparaît déterminant. En Traitement Automatique des Langues, de nombreux travaux ont également exploité les corpus pour extraire les expressions polylexicales, principalement les collocations (Cf. par exemple les travaux de Smadja 1993; Evert 2008 ; Seretan 2011; Ramisch 2014)[1]. Les linguistes s'intéressant aux expressions polylexicales et cherchant à constituer des ressources lexicales,

1. Parmi les nombreux projets de TAL consacrés à ce sujet, nous pouvons citer le projet ANR Parseme fr (http://parsemefr.lif.univ-mrs.fr).

en particulier dans une perspective lexicographique, se sont penchés sur l'évaluation des méthodes quantitatives exploitant des corpus. Ces techniques permettent en effet d'optimiser la constitution de telles ressources et, dans ce cadre, il apparaît évidemment souhaitable de repérer la méthode la plus appropriée pour cette tâche. Pour l'extraction des collocations, ce sont généralement des évaluations qui sont effectuées à partir de listes de collocations repérées à partir d'un schéma syntaxique spécifique (par exemple, la structure Nom-Adjectif comme dans *peur bleue* ou *amour fou*) et de méthodes quantitatives utilisant des mesures d'associations.

Les mesures d'association s'appuient sur le calcul du tableau de contingence, qui résume les occurrences des deux unités lexicales l'une par rapport à l'autre. Ce tableau contient 4 valeurs : f_{12}, le nombre de contextes où les deux unités apparaissent ensemble (fréquence de cooccurrence), f_1 et f_2 les nombres de contextes où apparaissent la première et respectivement la seconde (fréquences d'occurrence), et N le nombre total de contextes de calcul. Par contexte, on peut entendre différents types d'espaces de cooccurrence : fenêtres de largeur fixe, relations de dépendance syntaxiques, phrases, paragraphes, textes, etc. Pour le calcul, différentes mesures ont été proposées dans la littérature : information mutuelle spécifique (Church & Hanks, 1990), rapport de vraisemblance (Dunning, 1993), Chi2, t-score, z-score (Evert, 2007, Pecina et Schlesinger, 2006) ou test de Fischer (Bestgen, 2017). Quels que soient leurs caractéristiques et leurs fondements mathématiques, ces mesures s'appuient sur une comparaison entre la fréquence de cooccurrence observée et la fréquence de cooccurrence attendue (c'est-à-dire, celle que l'on obtiendrait par le seul jeu du hasard) : plus la première est élevée par rapport à la seconde, plus la valeur de la mesure sera haute. Le choix d'une mesure ou d'une autre dépend de nombreux paramètres : taille du corpus, fréquence des associations ciblées, degré de figement de celles-ci, etc. Certaines mesures ont tendance à favoriser les associations rares (comme l'information mutuelle spécifique ou le z-score) et d'autres les associations fréquentes (comme le rapport de vraisemblance) : comme le note Evert (2008), les différentes études effectuées sur le sujet ne permettent pas d'aboutir à une conclusion simple et unilatérale sur la supériorité d'une mesure par rapport aux autres. Pour notre part, nous utiliserons le rapport de vraisemblance, qui représente la meilleure approximation du test exact de Fischer (Bestgen, 2017) et produit de bons résultats pour des collocations assez fréquentes et générales.

Une fois extraites des corpus, les collocations peuvent être évaluées à l'aide de deux méthodes : a) manuellement, à l'aide d'experts ou b) automatiquement par comparaison avec des ressources existantes comme des dictionnaires, quand elles sont disponibles (Seretan 2011). Il ne faut toutefois pas oublier que les ressources lexicales utilisées dans les évaluations automatiques ont généralement été constituées manuellement. Seretan (2011) a effectué des comparaisons assez systématiques en utilisant la méthode des experts. La méthode d'extraction qu'elle a utilisée recourait au rapport de vraisemblance comme mesure d'association, couplée au repérage de relations syntaxiques, en utilisant une méthode de fenêtre de mots (les mots, par exemple les adjectifs, sont situés à n mots du mot pivot, par exemple le nom) ou une analyse syntaxique qui identifie directement le lien entre les mots, par exemple le nom et d'adjectif. L'évaluation, qui portait sur les 500 paires ayant les meilleures mesures sur la partie française du corpus Hansard, a été effectuée par trois linguistes qui ont porté un jugement sur l'intérêt des combinaisons : paire avec erreurs syntaxiques, paire correcte syntaxiquement, paire intéressante. Les combinaisons jugées intéressantes pouvaient être des collocations ou des expressions plus figées. La méthode utilisant la syntaxe obtient des résultats plus satisfaisants pour le repérage des collocations (65,9 % en moyenne) que la méthode basée sur la fenêtre (56,9%). Le recours à la syntaxe dans les méthodes d'extraction apparaît donc apporter des résultats plus proches de ceux des experts, même si la vérification manuelle apparaît indispensable. C'est la méthode que nous emploierons dans notre expérimentation.

D'autres travaux comme ceux de Simpson-Vlach & Ellis (2010) ont une philosophie un peu différente : ils cherchent à proposer une méthode quantitative qui s'approche le plus possible de l'intuition des experts. Pour élaborer leur *Academic Formulas List* (AFL) qui recense des séquences formulaires utiles dans les discours académiques oraux et écrits, ils utilisent des mesures statistiques qui ont été considérées pertinentes (*formula teaching worth*) par les enseignants qui les utilisent. Cela les amène à exploiter des ngrammes[2] associés à des mesures de fréquence et d'une mesure d'association (ici, l'information mutuelle). Ils observent que les expressions sélectionnées par les experts s'approchent des éléments retenus par la mesure d'association, davantage que par la fréquence. Toutefois, ils ne s'intéressent qu'aux ngrammes d'au moins trois mots et excluent donc les collocations binaires. La démarche de ces chercheurs nous paraît tout à fait intéressante, car l'évaluation des expressions est réalisée par des experts pour une tâche précise et dans un genre spécifique, ce qui nous apparaît plus pertinent que des évaluations réalisées sur de simples listes. Dans notre tâche d'évaluation, nous retiendrons cette perspective de tâche spécifique.

Enfin, certains travaux se sont intéressés aux collocations fondamentales. Benigno *et al*. (2016) élargissent aux collocations la notion de « vocabulaire fondamental » (Gougenheim *et al.*, 1964) et nomment « collocations fondamentales » des « unités polylexicales significatives (...) qui représentent pour les locuteurs natifs les contextes les plus essentiels d'un mot donné. »

2. Les ngrammes (ou segments répétés) sont des suites récurrentes de mots identiques.

(Benigno *et al*., 2016, p. 125). Dans l'étude présentée, certaines collocations fréquentes ont été extraites automatiquement à partir d'un vaste corpus tiré du Web (le FrWack, Baroni *et al*., 2010), puis présentées à des locuteurs natifs afin de recueillir leur jugement quant au caractère de collocation de ces expressions (comme *conférence de presse, organiser un colloque, conférence téléphonique,* etc.). L'étude montre une certaine corrélation entre des mesures d'association telles que le rapport de vraisemblance et le jugement des locuteurs : pour une collocation, plus cette mesure est haute, et plus nombreux sont les locuteurs qui l'ont annotée comme étant une collocation fondamentale. Étonnamment, même la fréquence brute semble bien corrélée au jugement des annotateurs : les collocations avec les mots outils ayant été supprimées (ce sont souvent les plus fréquentes, alors qu'elles sont peu significatives), la simple fréquence de cooccurrence devient un bon indicateur de l'association statistique, comme cela avait déjà été observé par Lapata *et al.* (1999) – ceci étant sans doute lié au fait que le corpus est de grande dimension, et que les collocations visées par l'étude sont elles-mêmes des collocations générales plutôt fréquentes.

On constate donc que les chercheurs observent une corrélation entre les jugements effectués par les experts et les extractions effectuées par les méthodes quantitatives. Toutefois, il nous paraît important d'effectuer une évaluation plus naturelle pour les experts linguistes, en leur proposant une tâche précise, liée à une application déterminée, d'une part, et en proposant, d'autre part, un jugement basé sur corpus, l'évaluation de simples listes lexicales hors contexte étant une tâche complexe et selon nous problématique.

3. L'expérimentation

Nous présentons ici la méthode mise en place pour effectuer notre évaluation. L'objet d'étude sélectionné est celui des collocations du lexique scientifique transdisciplinaire. L'évaluation est une procédure d'annotation des collocations, effectuée sur une application en ligne par 9 linguistes experts. Nous avons donc choisi une méthode d'évaluation « naturelle » en demandant aux experts d'observer les phénomènes dans les corpus mêmes.

3.1 Les collocations du lexique scientifique transdisciplinaire

Dans cette expérimentation, nous nous limitons à un phénomène linguistique restreint, en nous intéressant exclusivement aux collocations lexicales binaires apparaissant de façon transversale dans les écrits scientifiques (cf. Pecman 2004 ; Tutin 2018). Les collocations que nous souhaitons traiter relèvent du lexique scientifique transdisciplinaire, ce lexique transversal utilisé dans le discours scientifique et renvoyant aux méthodes scientifiques, à l'argumentation et à la structuration du discours (Pecman 2007 ; Tutin 2007 ; Hatier 2016 ; Jacques & Tutin 2018). Les collocations sont des constructions binaires, organisées autour d'une relation syntaxique comme les suivantes :

- N – PREP –> N : *hypothèse de travail, formulation d'une hypothèse* ;
- N – MOD –> A : *hypothèse valide, mutation profonde, immense majorité* ;
- V – OBJ –> N : *faire une hypothèse, jouer un rôle* ;

3.2 La procédure d'annotation

L'expérimentation a consisté à confronter un ensemble de linguistes au repérage des collocations en contexte dans un corpus d'extraits de textes scientifiques des sciences humaines. Comme précisé plus haut, l'utilisation de la procédure d'annotation dans les corpus nous paraît préférable à l'évaluation d'une liste d'expressions hors contexte principalement pour deux raisons. Tout d'abord, évaluer des expressions hors contexte est une tâche extrêmement difficile pour des évaluateurs. Le sens d'un mot ou d'une expression se révèle avant tout dans son contexte d'usage. De ce fait, la validité d'une évaluation hors contexte peut être discutée. Deuxièmement, dans l'hypothèse où on appliquerait une méthode d'extraction automatique (ce que nous envisageons dans des travaux ultérieurs), l'évaluation de cette méthode à partir de l'annotation permet de calculer non seulement la précision (la proportion de « bonnes réponses » parmi les réponses sélectionnées) mais aussi la proportion de réponses repérées (le « rappel »), ce que l'évaluation à partir de listes préétablies ne permet pas. Pour cette expérimentation, nous avons sollicité 9 linguistes[3] qui sont des locuteurs natifs ou quasi-natifs du français, connaissant très bien la question de la phraséologie, différenciant les expressions très figées du type *tenir compte* des collocations lexicales du type *hypothèse préalable* et familiarisés avec la notion de « lexique transdisciplinaire ».

3. Les experts sollicités sont tous titulaires d'un doctorat en sciences du langage. Certains de leurs travaux de recherche portent sur la question du lexique et de la phraséologie.

Les évaluateurs ont été amenés à annoter des extraits d'articles scientifiques, dans 5 disciplines des sciences humaines : anthropologie, économie, linguistique, psychologie et sociologie. Chaque texte a été annoté par 4 annotateurs différents. L'article de sociologie, qui contenait essentiellement de la phraséologie disciplinaire, a été retiré de l'évaluation, car il apparaissait au final peu pertinent pour l'évaluation. Le choix de plusieurs disciplines des sciences humaines n'a pas pour but de comparer les disciplines, l'échantillon étant ici trop limité pour cette tâche, mais de proposer une certaine variété lexicale.

Extrait de Corpus	Anthropologie	Economie	Linguistique	Psycho
Taille en nombre de mots	2306	2898	4865[4]	2465

Tableau 1 : Composition du corpus utilisé pour la tâche d'annotation

La consigne proposait aux annotateurs de repérer les collocations correspondant à « des expressions récurrentes comme *faire une hypothèse* ou *effectuer une analyse* qui apparaissent dans tous types de textes scientifiques et qui sont utiles pour rédiger de façon adéquate un écrit scientifique ». Une explication brève rappelait comment distinguer les collocations des expressions complètement figées (du type *tenir compte*). Pour limiter la tâche d'annotation, seules deux structures productives de collocations ont été proposées aux annotateurs : les structures de type V-N (*faire une hypothèse*, *jouer un rôle*) ou de type N-A (*hypothèse valide*, *grand nombre*).

L'annotation a été effectuée en ligne par les annotateurs, qui devaient simplement cliquer sur les composants de la collocation, comme on peut l'observer sur copie d'écran de la figure 1[5]. L'interface permet d'associer deux, trois ou plus unités lexicales afin de les enregistrer comme collocation. Les collocations ainsi enregistrées peuvent éventuellement se chevaucher : par exemple, dans *Ce phénomène joue un rôle crucial*, on peut sélectionner à la fois *jouer un rôle* et *rôle crucial*.

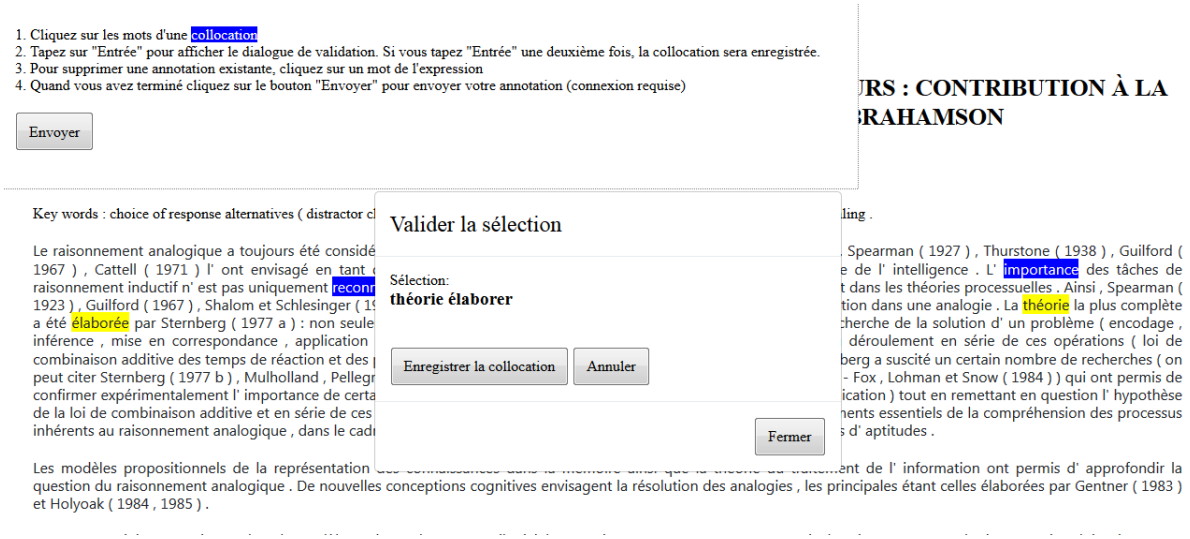

Figure 1 : L'interface d'annotation des collocations

4. Analyse des résultats

4.1 Collocations annotées par les experts

À l'issue de l'annotation, les résultats ont été recueillis et normalisés[6]. Sur l'ensemble des textes, 622 collocations pour 321 types différents ont été retenues par les annotateurs, soit une fréquence moyenne de 1,94 par expression. Parmi les 321 collocations différentes retenues, 130 (soit 40,4%) correspondent au patron N-A (par exemple, *double paradoxe* ou *perspective quantitative*) 191 (soit 59,6%) relèvent de la structure V-N (par exemple, *citer une étude* ou *exclure le cas*). Si l'on observe maintenant l'accord interannotateurs, il n'apparaît pas très élevé (cf. figure 2). Il est difficile de calculer un score

4. L'extrait complet intègre 5634 mots mais nous avons enlevé de notre calcul les nombreux exemples linguistiques qui n'intégraient pas des parties à annoter.
5. L'interface se compose essentiellement d'une feuille de style CSS et de scripts JQuery associés aux fichiers XML-TEI du corpus, chargés d'envoyer les résultats au serveur via des fonctions AJAX.
6. Les déterminants ont été exclus et les structures non conformes (par exemple V Prep N) ont été écartées.

de kappa de Fleiss, comme dans une tâche de catégorisation classique, où les unités sont présélectionnées en amont et où l'annotation ne porte que sur un choix d'étiquettes. Ici, les unités ont été sélectionnées par les annotateurs eux-mêmes dans le texte, le choix d'une unité reposant sur des critères syntaxiques (dépendances entre le nom et le verbe ou l'adjectif), sémantiques (appartenance au lexique transdisciplinaire) et phraséologiques (caractérisation de la collocation comme une unité semi-figée constituée d'une base et d'un collocatif). L'accord entre les annotateurs ne semble pas dépendre fondamentalement de la structure, même si la structure V-N reçoit un accord légèrement meilleur d'au moins 3 annotateurs (31%).

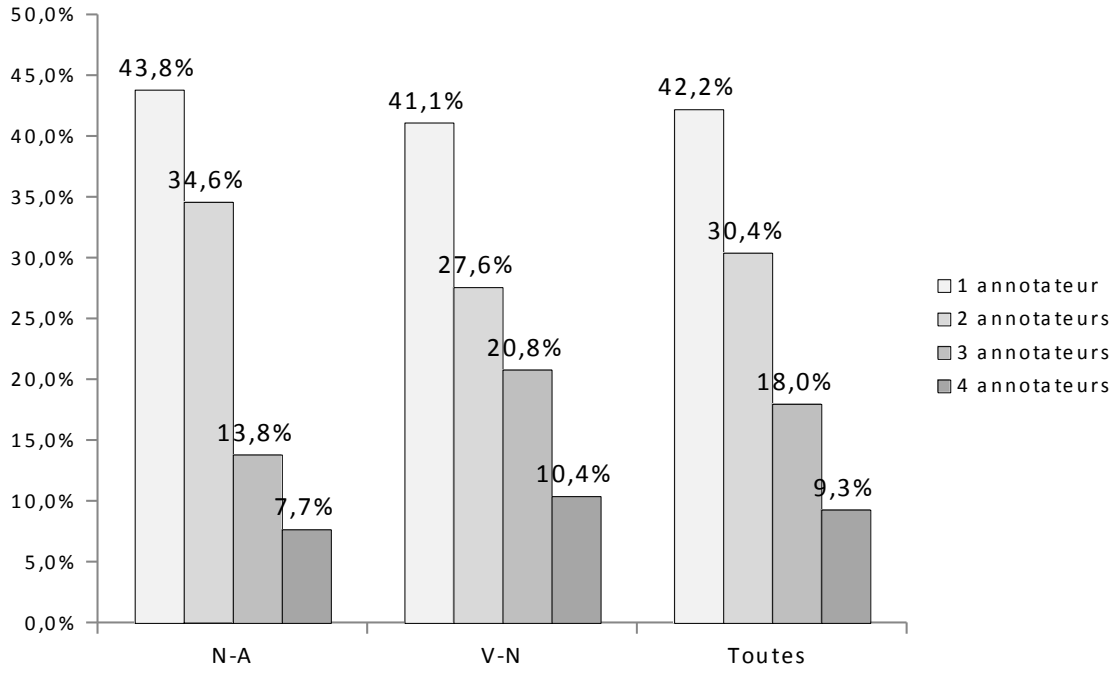

Figure 2 : Collocations et nombre d'annotateurs (en %)

Les sources de variations entre les jugements des annotateurs sont nombreuses :

- degré de figement : certaines combinaisons apparaissaient comme plutôt libres (p.ex. *autre particularité*, *seule variable*, *constater la similarité*) et constituent des cas limites. Ce paramètre semble plus fréquent avec les structures de type N-A.

- degré de généralité sémantique : les collocations très générales, comme *offrir la possibilité*, *ordre alphabétique*, *tenir un discours* ont parfois été annotées. À l'opposé, des outils méthodologiques ancrés dans certaines disciplines comme les tests à *choix multiples* en psychologie, ont parfois été considérés comme transdisciplinaires.

- inattention : certaines collocations ne faisant pas partie du LST ont été annotées quand même, sans doute parce qu'elles satisfaisaient bien les critères syntaxiques et phraséologiques (p.ex. *accident grave*, *interdit réglementaire*).

- omission : au sein d'un texte comptant plusieurs milliers de mots, une collocation peut simplement ne pas être vue. La probabilité d'omission augmente vraisemblablement avec l'éloignement syntagmatique des deux unités.

Le fait d'avoir 42,2% d'unités sélectionnées par un seul annotateur n'est donc pas très surprenant, car parmi les expressions polylexicales, les collocations apparaissent comme des expressions dont le repérage est peu consensuel. Un précédent travail

sur l'annotation des expressions polylexicales à partir d'un repérage semi-automatique va dans le même sens que ces résultats et confirme la complexité de cette tâche (Tutin *et al.* 2015)[7].

Au final, compte tenu du très grand nombre de choix possibles (une sélection aléatoire pouvant donner lieu à des milliers de combinaisons différentes), l'accord n'est pas si mauvais : on constate que 57,8% des collocations ont été annotées par au moins deux annotateurs. Si l'on se penche maintenant sur les collocations qui ont fait l'unanimité (c'est-à-dire, annotées par la totalité des 4 annotateurs), on observe qu'elles présentent un fort degré de « scientificité » et renvoient pour la plupart à des objets scientifiques abstraits (*phénomène, question, cas*, …), des procédures et actions scientifiques (*analyser, étudier, décrire, travail*) ou des propriétés/caractéristiques (*récent, intrinsèque …*). Peu de collocations hormis *attirer l'attention* semblent se situer en dehors de ces sphères scientifiques (cf. Pecman 2004).

Pour les collocations ainsi identifiées, nous avons extrait un certain nombre de statistiques issus du corpus TermITH Transdisciplinaire (Hatier 2016 ; Jacques & Tutin 2018). Ce corpus contient 500 articles (répartis dans 10 disciplines de sciences humaines) qui représentent un total de 4 834 361 mots. Il a été annoté syntaxiquement en dépendances, grâce à l'analyseur XIP (Aït-Mokhtar et al, 2002). À partir de ce corpus, nous avons extrait la fréquence de cooccurrence (nombre de fois que les mots sont liés par une relation de dépendance correspondant à la relation syntaxique visée), la dispersion (nombre de disciplines différentes où la cooccurrence apparaît), ainsi que les deux mesures du t-score et du rapport de vraisemblance[8].

4.2 Corrélations avec les propriétés statistiques

Pour examiner la corrélation entre le jugement des annotateurs et ces mesures objectives, une possibilité serait d'examiner s'il existe une corrélation entre ces mesures et le nombre de votes pour une même collocation (de 1 à 4, paramètre quantitatif intersubjectif), comme l'ont fait Benigno *et al*. (2016). Mais comme le montre le tableau 3, ces coefficients sont faibles et ne permettent pas de tirer de conclusion. On pourra juste remarquer une tendance à obtenir un meilleur consensus pour les unités plus dispersées, c'est-à-dire apparaissant dans plusieurs disciplines.

	r
fréquence	0,23
dispersion	0,31
t-score	0,27
rapport de vraisemblance	0,20

Tableau 2 : Corrélation linéaire (coefficient de Pearson) entre les mesures et le nombre d'annotateurs

Sans doute faudrait-il recourir à un plus grand nombre d'annotateurs pour avoir un nombre de votes suffisant : 4 annotations par document ne suffisent pas à obtenir des données statistiquement significatives pour observer des corrélations linéaires fiables. On peut cependant observer certaines tendances, en considérant trois catégories, dont les effectifs sont suffisamment importants pour que le calcul des valeurs moyennes soit significatif :

- collocations identifiées par 1 annotateur : pas de consensus
- collocations identifiées par 2 annotateurs : consensus moyen
- collocations identifiées par 3 ou 4 annotateurs : consensus fort

Pour ces trois catégories, un calcul des moyennes montre une certaine corrélation :

7. Dans cette expérimentation, nous avions observé pour toutes les expressions les accords interannotateurs suivant : 0,683 avec le coefficient kappa de Fleiss pour le texte littéraire, 0,741 pour un rapport scientifique. Le consensus est moins important pour les collocations. Rappelons que le texte est annoté semi-automatiquement, ce qui améliore de toute évidence l'accord obtenu.

8. Le rapport de vraisemblance, en anglais *LogLike* (abréviation de *Log Likelihood Ratio*, parfois noté G2), est utilisé comme test d'indépendance. Cette mesure a été proposée par Dunning (1993) et est d'un usage très répandu pour mesurer le degré d'association entre deux événements. Le *t-score* exprime le nombre d'écarts-types entre la valeur observée et la valeur « attendue » pour une observation distribuée selon une loi normale. Il s'interprète comme le khi2 et le *LogLike*, mais il permet de distinguer les corrélations négatives et les corrélations positives (notons qu'il existe également une version signée du LogLike, dans certaines implémentations).

	Effectif	Fréquence	Dispersion	T-score	rapport de vraisemblance
Pas de consensus 1 annotateur	128	16,08	3,2	2,08	56,76
Consensus moyen 2 annotateurs	87	19,61	3,9	2,61	80,61
Consensus fort 3 ou 4 annotateurs	88	64,23	5,5	4,60	266,97

Tableau 3 : valeur moyenne des statistiques et des mesures par catégorie de consensus

On voit que les unités qui obtiennent le meilleur consensus sont, en moyenne, les plus fréquentes et les plus dispersées dans le corpus TermITH-Trandisciplinaire. Cela se traduit naturellement par une évolution conjointe des deux mesures, t-score et rapport de vraisemblance. Par ailleurs, il est possible d'observer le lien entre les annotations manuelles et les mesures objectives sous la forme de l'évolution du rappel en fonction des seuils appliqués.

Nous proposons de considérer comme annotation de référence l'ensemble des collocations ayant été sélectionnées par au moins deux annotateurs, et réalisant au moins 4 occurrences dans le corpus TermITH. En effet, en deçà de cette fréquence seuil, les mesures statistiques perdent en significativité, et leur usage devient problématique. On obtient un ensemble de référence de 113 collocations différentes (sur les 322 initialement obtenues). Pour cet ensemble de référence, il peut être intéressant d'observer l'évolution du rappel en fonction du seuil qui serait appliqué si on se basait sur une mesure objective pour identifier les collocations (ce qui se fait dans les méthodes automatiques) - le rappel étant la proportion des collocations de références identifiées par application automatique du seuil. Les figures ci-dessous montrent l'évolution du rappel en fonction des seuils de fréquences, de rapport de vraisemblance et de t-score :

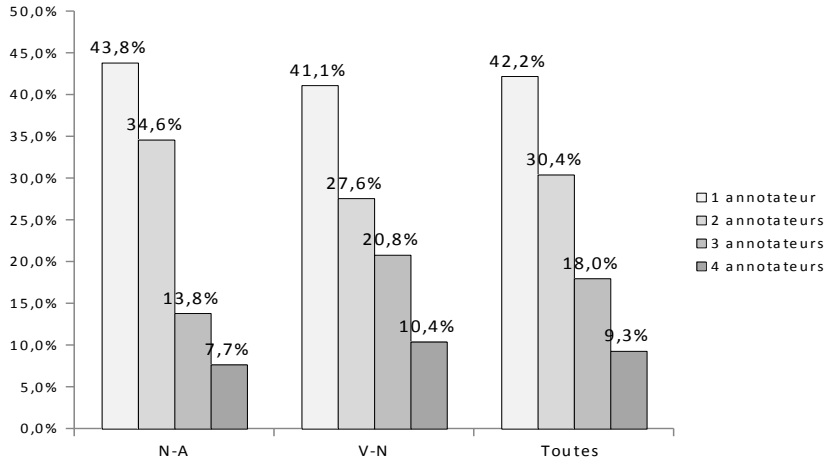

Figure 3: Evolution du rappel en fonction du seuil de fréquence

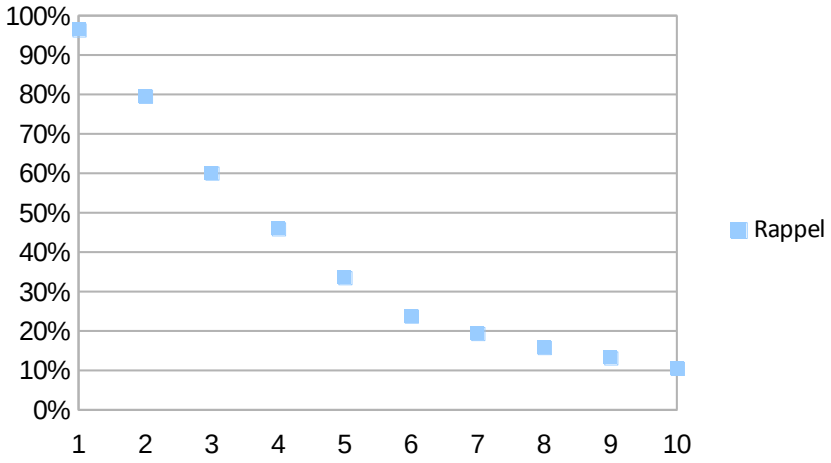

Figure 4 : Evolution du rappel en fonction du seuil de rapport de vraisemblance

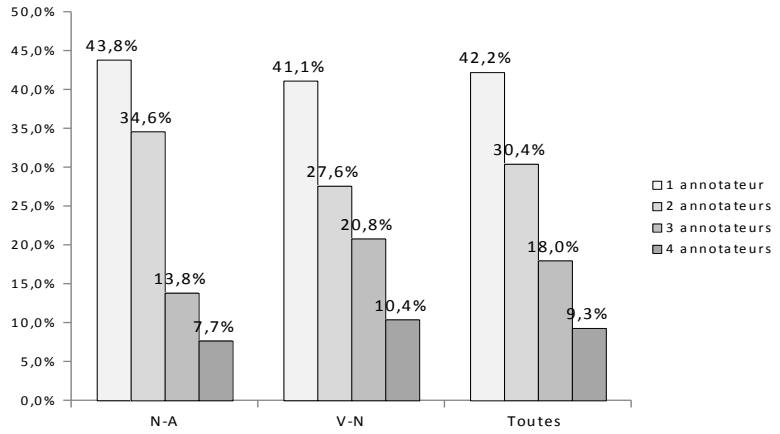

Figure 5 : Evolution du rappel en fonction du seuil de t-score

On voit que ces courbes présentent des profils très différents et donnent des indications intéressantes sur la fixation des seuils quantitatifs pour filtrer des collocations obtenues automatiquement :

- la fréquence absolue (nombre d'occurrences dans le corpus) est un seuil très sensible : entre 4 et 5, on perd d'un coup la moitié des collocations intéressantes. Au-delà de ce palier, la dégradation devient linéaire, avec une perte d'environ 10% par point de fréquence supplémentaire. S'il est intéressant de ne pas considérer les fréquences trop faibles (entre 1 et 3, on considère que les cooccurrences sont peu significatives et comportent beaucoup de bruit), il peut être contreproductif de fixer un seuil de fréquence trop élevé. Notons toutefois que pour des corpus plus vastes, ce palier avec dégradation brutale du rappel, s'il existe, pourrait être repoussé à des valeurs supérieures à 4.

- le t-score présente une dégradation assez constante de 20% par point supplémentaire, la pente s'adoucissant entre 4 et 5. Dans une tâche d'extraction automatique de collocation, Ramisch *et al.* (2010) obtiennent des valeurs bien inférieures au plan du rappel (20,91% pour un seuil de t-score à 1, 6,42% pour un seuil à 5) mais ces chiffres ne sont pas tout à fait comparables avec les nôtres : il s'agit dans leur cas d'une extraction automatique, qui implique qu'une sélection a été faite des candidats collocations, ce qui diminue naturellement le rappel par rapport à une extraction où l'on aurait considéré toutes les paires de mots possibles à l'intérieur des phrases (ce qui correspond théoriquement à notre cas de figure). Par ailleurs, leur extraction est effectuée sur des cooccurrences de surface à partir d'une fenêtre réduite, sans tenir compte des relations syntaxiques, ce qui fait que certaines collocations éloignées dans la phrase sont nécessairement négligées. Quant aux précisions obtenues (le pourcentage de collocations correctes parmi celles identifiées automatiquement), ils obtiennent respectivement des valeurs de 19,96% et 28,93% pour les deux seuils évoqués - nous ne sommes malheureusement pas en mesure de donner une valeur de précision, n'ayant pas automatisé la sélection des collocations en fonction des relations syntaxiques et du t-score.

- le rapport de vraisemblance quant à lui présente un profil différent : on observe un plateau avec des valeurs élevées de rappel (supérieure à 80) jusqu'à 15, suivies d'une dégradation progressive. Cette observation suggère que le filtrage opéré dans cet intervalle cible plus spécifiquement le bruit (les candidats collocations erronés) sans éliminer trop de bonnes collocations. On conseillera donc l'utilisation de cette mesure avec un seuil fixé entre 10 et 15 pour avoir un équilibre optimal entre rappel et précision.

4.3 Discussion

Pour interpréter ces observations sous le prisme de l'opposition qualitatif / quantitatif, il convient au préalable d'éclaircir ces deux notions.

La notion de qualité (par exemple la couleur en tant que qualité sensible) est appréhensible par la langue sous la forme de catégories, établies de manière intersubjective. La langue, en tant qu'elle catégorise le réel, est la principale pourvoyeuse de jugements qualitatifs (« ceci est rouge », « ceci est bleu », ...). Il faut distinguer ce premier niveau linguistique du niveau métalinguistique, qui est propre à l'étude du langage. Les catégories métalinguistiques sont issues de jugements qualitatifs (intersubjectifs) qui portent sur les éléments linguistiques, par exemple : « ces deux énoncés ont le même sens », « cet énoncé n'est pas grammatical », « ces deux unités appartiennent à la même classe morphosyntaxique », « ces deux mots peuvent commuter », « ce mot a un sens plus général ». Les observations qualitatives reposent, en linguistique, sur des exemples (l'*exemplier* est un outil méthodologique qui occupe une place toute particulière dans cette discipline, par rapport à d'autres

sciences humaines) permettant de montrer ou d'illustrer des phénomènes langagiers - et ceci quelle que soit l'origine de ces exemples, qu'ils soient fabriqués ou issus de corpus. La validité de ces démonstrations repose sur l'accord intersubjectif des locuteurs qui interprètent ces exemples, en d'autres termes sur leur compétence linguistique partagée (au sens Chomsky). Les tests linguistiques fonctionnent de la même manière : par les transformations qu'ils opèrent, ils permettent d'identifier des équivalences ou des différences, et de montrer ce qui est acceptable et ce qui ne l'est pas. Ainsi, ces observations sont dites qualitatives parce qu'elles reposent en fait sur des jugements et des catégories métalinguistiques.

Par opposition, on a l'habitude de nommer « quantitative » toute observation reposant sur des données chiffrées : par exemple, la fréquence fondamentale de l'enregistrement d'un phonème, la fréquence d'un mot dans un corpus, ou la longueur d'un énoncé. Mais il s'agit d'une opposition seulement en apparence, car (et c'est le propre des sciences humaines) ces données numériques reposent, dans leur construction, sur des catégorisations qualitatives : si l'on demande à un locuteur de prononcer un certain phonème, il faut au préalable avoir identifié ce phonème en tant que catégorie linguistique. De même, pour compter des mots, il faut s'entendre sur la définition du mot, pour laquelle il faudra convoquer toute une théorie linguistique impliquant des choix intersubjectifs complexes (découpage en morphèmes, association à un lemme, etc.). Enfin, toute tâche d'annotation qui peut donner lieu à des statistiques quantitatives, comme précédemment, repose initialement sur des variables nominales intrinsèquement qualitatives : un mot appartient-il au LST ? telle association correspond-elle à une collocation ? s'agit-il d'une association entre un adjectif et un nom, entre un verbe et un nom en fonction d'objet (ce qui n'est pas évident quand l'adjectif est interprétable comme un participe passé, ou lorsque l'objet devient sujet d'un verbe au passif) ?

Si l'on revient aux observations précédentes, on peut constater :

- qu'il est possible de trouver un certain consensus intersubjectif concernant des propriétés complexes telles que l'appartenance au LST, l'aspect préfabriqué, le semi-figement, - et ceci tout en laissant la tâche d'annotation relativement vague et ouverte ;

- que ce consensus pourrait peut-être être amélioré, en encadrant plus finement la tâche des annotateurs, grâce à des exemples plus nombreux et à des tests permettant de départager certaines ambiguïtés, afin de guider leur intuition. Ceci dit, la mise en œuvre de consignes d'annotation plus complexes peut soulever de nouveaux problèmes, comme perdre les annotateurs par la manipulation de règles trop complexes, ou encore orienter leurs choix dans une direction déterminée à l'avance, ce qui pourrait introduire des biais - d'où notre choix de limiter le guide d'annotation à une page.

- que les choix d'annotation peuvent être corrélés à des mesures « objectives » telles que la fréquence, la dispersion ou le rapport de vraisemblance.

Jugement intersubjectif et observations quantitatives se rejoignent, et pourtant dans le cas des collocations, il n'est pas toujours possible d'asseoir le jugement intersubjectif sur des catégories métalinguistiques qualitatives, telles que les différentes dimensions de la préfabrication (limitation du paradigme combinatoire, …), mais on peut émettre l'hypothèse que l'usage et l'exposition à ces expressions, dans des genres spécifiques, apparaît déterminant. En d'autres termes, ce que le jugement humain identifie est peut-être aussi, d'une certaine manière, un jugement d'ordre quantitatif, une forme d' « enracinement » bien connu dans le cadre de la linguistique cognitive (voir par exemple, Croft & Cruse 2004). Par exemple, la collocation *principale caractéristique* identifiée par 4 annotateurs, a été jugée « récurrente » et « utile », comme formulée dans le guide d'annotation, alors qu'elle n'apparaît pas du tout contrainte.

5. Conclusion

Cette modeste étude empirique s'intéresse à un phénomène linguistique qu'il est difficile de caractériser qualitativement : le phénomène collocationnel. Comme le note Evert (2008), il s'agit d'une notion « controversée », qui fait l'objet de deux approches différentes :

- une approche que l'on pourrait qualifier de quantitative, selon l'école firthienne ou néo-firthienne, qui interprête la collocation « comme des observations empiriques relatives à la prédictibilité des combinaisons lexicales : elles évaluent quantitativement "l'attente mutuelle" [*mutual expectation*] (Firth 1957, 181) entre les mots, et l'influence statistique qu'un mot exerce sur son voisinage »[9] (Evert 2008 : 2113)

9. "Collocations in this Firthian sense can also be interpreted as empirical statements about the predictability of word combinations: they quantify the "mutual expectancy" (Firth 1957, 181) between words and the statistical influence a word exerts on its neighbourhood."

- d'autre part une approche plus qualitative où « le terme "collocations" a été utilisé dans le domaine de la phraséologie pour désigner des combinaisons de mots semi-compositionnelles et lexicalement déterminées, telles que *stiff drink* [boisson rafraîchissante] (avec un sens particulier de *stiff* limité à un ensemble particulier de noms), *heavy smoker* [gros fumeur] (où *heavy* est le seul intensificateur acceptable pour *fumeur*), *give a talk* [donner une conférence] (plutôt que *make* ou *hold*) et *a school of fish* [un banc de poissons] (plutôt que *group*, *swarm* ou *flock*). Ce point de vue a été défendu avec force par Hausmann (1989) et a été de plus en plus largement accepté ces dernières années (par exemple Grossmann et Tutin 2003). » [10] (Evert, 2008:2114)

Cette dualité dans les approches du phénomène collocationnel est selon nous révélatrice du fait qu'il se trouve dans une zone intermédiaire entre deux types de phénomènes linguistiques :

- des phénomènes susceptibles d'être identifiés par des catégories méta-linguistiques claires : par l'application de tests simples, on peut par exemple identifier que **une peur très bleue*, **un troupeau de poisson*, ?*un fumeur lourd* ou ?*créer une hypothèse* s'écartent de l'usage normal de langue.

- des phénomènes qui n'ont pas d'autres caractéristiques que d'être fréquents et surreprésentés dans certains corpus : c'est le cas de collocations plus libres, ou des routines sémantico-rhétoriques, comme *il est frappant de constater que...*. Aucun test ne permet vraiment de rattacher ces expressions aux catégories du figement syntaxique ou de l'opacité sémantique. Simplement, elles sont préfabriquées dans certains discours. Elles sont maîtrisées par les locuteurs experts de ces discours, d'une manière inconsciente, le plus souvent. Elles assument sans doute une fonction intégrative au sein d'une communauté : la maîtrise de ces éléments discursifs est perçue (là encore de manière plus ou moins consciente) par le récepteur comme une marque de l'appartenance à une certaine communauté et de la maîtrise de son sociolecte.

Pour ces phénomènes, en définitive, il ne reste que deux modes d'appréhension, qui échappent aux catégories métalinguistiques classiques : l'observation quantitative de la forte récurrence de ces phénomènes, et à l'opposé, la reconnaissance intuitive des locuteurs « experts » qui utilisent et manipulent ces expressions, sans pour autant pouvoir constater une combinatoire « anormale » au plan lexicogrammatical. C'est ce qu'indiquent, selon nous, les résultats de notre étude empirique : d'une part les annotateurs sont capables de reconnaître, avec un certain consensus, des collocations difficilement caractérisables par des tests linguistiques et des catégories métalinguistiques traditionnelles (limitation combinatoire, etc.), d'autre part, ces collocations se trouvent corrélées à des phénomènes statistiques purement quantitatifs.

Pour compléter cette étude, dans de futures recherches, nous envisageons de confronter les annotations manuelles avec un système d'extraction automatique se basant à la fois sur la syntaxe, pour reconnaître les relations de dépendances entre les unités, et sur des critères purement statistiques, tels que dispersion et rapport de vraisemblance. Nous pensons pouvoir montrer ainsi de manière plus complète l'adéquation entre le jugement des locuteurs et les phénomènes de récurrence purement quantitatifs.

Remerciements

Nous remercions chaleureusement les annotatrices et annotateurs qui ont participé à cette étude : Magdalena Augustyn, Cristelle Cavalla, Francis Grossmann, Évelyne Jacquey, Sylvain Hatier, Mojca Pecman et Hoai Tran.

Bibliographie

AÏT-MOKHTAR, Salah, CHANOD, Jean-Pierre, ROUX, Claude (2002), Robustness beyond shallowness : incremental deep parsing, *Natural Language Engineering*, 8(2-3), 121–144.

BARONI Marco, BERNARDINI, Silvia, FERRARESI, Adriano, PICCI, Giovanni (2010), Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation, in Xiao R. (Ed.), *Using Corpora in Contrastive and Translation Studies*, Newcastle, Cambridge Scholars Publishing, 259–274.

BENIGNO, Veronica, GROSSMANN, Francis, KRAIF, Olivier, VELEZ, Antonino (2016), La notion de collocation fondamentale: une étude de corpus. *Cahiers de lexicologie, 1*(108), 125-146.

10. " the term "collocations" came to be used in the field of phraseology for semi-compositional and lexically determined word combinations such as stiff drink (with a special meaning of stiff restricted to a particular set of nouns), heavy smoker (where heavy is the only acceptable intensifier for smoker), give a talk (rather than make or hold) and a school of fish (rather than group, swarm or flock). This view has been advanced forcefully by Hausmann (1989) and has found increasingly widespread acceptance in recent years (e.g. Grossmann and Tutin 2003). "

BESTGEN, Yves (2017), Évaluation de mesures d'association pour les bigrammes et les trigrammes au moyen du test exact de Fisher, *Actes de TALN 2017*, 10-18.
CAVALLA, Cavalla (2014). Collocations transdisciplinaires : réflexion pour l'enseignement, in Gonzalez-Rey (ed.), *Outils et m éthode d'apprentissage en phraséodidactique*, EME, 151-169.
CHURCH, Kenneth Ward, HANKS, Patrick (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.
CROFT, William, CRUSE, D. Allan (2004) , *Cognitive linguistics*. Cambridge, Cambridge University Press.
DROUIN, Patrick (2007), Identification automatique du lexique scientifique transdisciplinaire, *Revue française de linguistique appliquée*, *12*(2), 45-64.
DUNNING, Ted (1993), Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.Evert, Stefan (2008) Corpora and collocations, in A. Lüdeling and M. Kytö, *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin, 1212–1248.
FIRTH, John Rupert (1957),A synopsis of linguistic theory 1930–55, In *Studies in linguistic analysis*, The Philological Society, Oxford, 1–32.
GORCY, GERARD, MARTIN, Robert, MAUCOURT, J., VIENNEY, R. (1970), Le traitement des groupes binaires. *Cahiers de Lexicologie, n° 2*, 15-46.
JACQUES, Marie-Paule, TUTIN, Agnès (2018), *Lexique transversal et formules discursives des sciences humaines*, Londres, ISTE Editions.
KRAIF, Olivier (2016), Le lexicoscope: un outil d'extraction des séquences phraséologiques basé sur des corpus arborés, *Cahiers de lexicologie*, (108), 91-106.
KRISHNAMURTHY, Ramesh (ed) (2004), *The OSTI Report*, London/New-York, Continuum.
LAPATA, Maria, MCDONALD, Scott, KELLER, Franck (1999), Determinants of adjective-noun plausibility, *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999)*, 30–36.
NESSELHAUF, Nadja (2005), *Collocations in a learner corpus*, Amsterdam, John Benjamins Publishing.
PECINA, Pavel, SCHLESINGER, Pavel (2006), Combining association measures for collocation extraction, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Poster Sessions, 651–658, Sydney, Australia. ACL.
PECMAN, Mojca (2004), *Phraséologie contrastive anglais-français: analyse et traitement en vue de l'aide à la rédaction scientifique*, Thèse de doctorat, Université de Nice.
RAMISCH, Carlos (2014), *Multiword expressions acquisition: A generic and open framework*, Dordrecht, Springer.
RAMISCH, Carlos, Villavicencio, Aline, Boitet, Christian (2010), Mwetoolkit: a Framework for Multiword Expression Identification. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 2010, Valetta, Malta. European Language Resources Association, 662-669.
SERETAN, Violeta (2011), *Syntax-based collocation extraction*, Dordrecht, Springer.
SIMPSON-VLACH, Rita, ELLIS, Nick C. (2010), An academic formulas list: New methods in phraseology research, *Applied linguistics*, *31*(4), 487-512.
SINCLAIR, John (1991),*Corpus, concordance, collocation*, Oxford, Oxford University Press.
SMADJA, Franck (1993), Retrieving collocations from text: Xtract. *Computational linguistics*, *19*(1), 143-177.
TUTIN, Agnès (2007), Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, *12*(2), 5-14.
TUTIN, Agnès, ESPERANÇA-RODIER, Emmanuel, IBORRA, Manuel, REVERDY, Justine (2015), Annotation of multiword expressions in French, *European Society of Phraseology Conference (EUROPHRAS 2015)*, 60-67.
TUTIN, Agnès (2018) « Les expressions polylexicales transdisciplinaires dans les articles de recherche en sciences humaines : retour d'expérience», *in* Jacques, M. P. ; Tutin, A. (dir.), *Lexique transversal et formules discursives des sciences humaines*, ISTE Editions, Londres, pp. 73-89.

LIDILEM – Linguistique et Didactique des Langues Étrangères et Maternelles, Grenoble
Olivier.kraif@univ-grenoble-alpes.fr; agnes.tutin@univ-grenoble-alpes.fr

Résumé
Dans cet article, nous faisons l'hypothèse que le phénomène collocationnel échappe à une approche qui serait purement qualitative, au sens usuel où l'on entend les méthodes qualitatives en linguistique. Pour illustrer cette position intermédiaire des collocations dans l'articulation de l'opposition méthodologique entre qualitatif et quantitatif, nous mettons en œuvre une étude empirique autour d'une tâche de repérage manuel des collocations dans un corpus, permettant de confronter le jugement intersubjectif de quelques linguistes experts avec les données quantitatives issues du comptage des fréquences d'occurrence et de cooccurrence dans un corpus de référence. Les résultats de l'étude confirment en grande partie notre hypothèse.

Mots-clefs :
Collocations – méthodes quantitatives – annotation de corpus

Abstract
In this article, we assume that collocations cannot be purely analysed with the help of a qualitative analysis, in the usual sense of qualitative methods in linguistics. In order to illustrate this intermediate position of collocations half-way between qualitative and quatitative criteria, we implement an empirical study with a task of manual identification of collocations in a corpus. This allows to compare the intersubjective judgment of some experts with the quantitative data resulting from

quantitative data (frequencies, asscociation measure, dispersion) in a reference corpus. The results of the study largely confirm our hypothesis.

Keywords :
Collocations – quantitative method – corpus annotation