# TESI DI DOTTORATO IN CO-TUTELA

**Università degli Studi di Palermo,** Facoltà di Lettere e Filosofia, Dipli, Dottorato in *Letterature moderne e studi filologico-linguistici,* settore scientifico disciplinare L-LIN/04, XXIII ciclo

**Université Stendhal di Grenoble**, UFR des Sciences du Langage, Lidilem, Ecole doctorale *Langues, littératures et sciences humaines*, Spécialité Didactique et Linguistique

# La notion de collocation fondamentale. Etude de corpus en vue d'une exploitation didactique.


Dottoranda
**Veronica Benigno**

Tutor
**Prof. Antonino Velez**

Co-tutor
**Prof. Francis Grossmann**

Coordinatore
**Prof.ssa Laura Auteri**

# REMERCIEMENTS

Cette thèse n'aurait pas pu voir le jour sans la contribution de plusieurs personnes à qui je voudrais exprimer ici toute ma reconnaissance.

Je tiens à remercier Francis Grossmann, pour m'avoir chaleureusement accueillie en France, pour l'ouverture intellectuelle qui a caractérisé nos échanges, pour le temps qu'il m'a consacré et pour avoir dirigé ma thèse de façon attentive et constante.

Mes remerciements vont également à Antonino Velez, pour sa gentillesse, pour son soutien au cours de cette recherche, pour sa confiance et pour m'avoir permis de suivre un parcours singulier.

Je suis reconnaissante envers Olivier Kraif, pour avoir donné forme tangible à mon idée, pour n'avoir pas hésité à offrir son aide pour la partie pratique de ce travail, et pour avoir répondu avec calme et patience à mes nombreuses questions.

Je remercie Vincenzo Lo Cascio, que je considère comme mon point de référence intellectuel, pour m'avoir guidée affectueusement, me laissant pourtant chercher seule quand il le fallait, et pour m'avoir transmis sa passion pour l'étude des collocations. Je le remercie de l'honneur qu'il m'a fait en acceptant d'être le rapporteur de ma thèse.

Merci à Kees Hengeveld pour m'avoir accueillie à Amsterdam au début de ma recherche, et à Piek Vossen pour m'avoir accueillie durant la dernière partie de mon doctorat, me dispensant des conseils précieux.

Mes remerciements vont également à l'équipe du Lidilem de Grenoble, pour les nombreux moments d'échange et de formation.

Je remercie Martin, pour m'avoir rassurée, soutenue, encouragée, et pour avoir passé des nuits blanches sur mes données. De tout l'amour dont je suis capable, je lui suis reconnaissante pour avoir partagé avec moi les moments quotidiens qui ont mené à l'accomplissement de cette thèse.

Mon plus grand remerciement revient à mes parents bien-aimés, sur qui je suis certaine de toujours pouvoir compter : je suis reconnaissante envers vous de m'avoir donné le privilège de réaliser mes études sereinement. A vous je dédie ce travail.

A ma sœur, pour avoir été curieuse et tendrement enthousiaste envers mes succès, vers lesquels elle a toujours su me ramener dans les moments de doute. A elle, à son aimé, et à mon neveu Paolo qui va bientôt changer nos vies, j'adresse toute ma gratitude.

Je remercie mon grand-père Francesco et ma grand-mère Epifania, à qui j'exprime tout mon amour.

Aux correcteurs français, amis ou connaissances, à Luc en particulier, merci pour avoir fait preuve d'altruisme en m'aidant à éliminer les fascinantes erreurs d'une locutrice non native.

A l'équipe du Lidilem de Grenoble, à tous mes amis et collègues, et à tous ceux qui m'ont soutenue, merci de n'avoir jamais douté du succès de mon entreprise.

A Jacqueline, enseignante, qui m'a appris à écrire en français.

« Non ho pensato alla strada, ma al punto di arrivo. Al punto d'arrivo, ripenso alla strada ».

# SOMMAIRE

# LISTE DES TABLEAUX

# LISTE DES FIGURES

# CHAPITRE 1
# Introduction

Ce premier chapitre présente un aperçu de la problématique traitée, et présente le cadre théorique dans lequel elle se situe. Nous synthétiserons ici l'objet et l'objectif de la recherche et expliquerons la méthodologie adoptée et les implications envisagées.

## 1.1 Description de la recherche

Les collocations constituent aujourd'hui une problématique réelle en linguistique et en linguistique appliquée. Elles jouent un rôle essentiel dans le traitement automatique du langage naturel, en particulier en traductologie (dans ce domaine, les champs d'investigation privilégiés sont, par exemple, l'extraction terminologique pour les langues de spécialité et la reconnaissance d'entités nommées dans le domaine de l'extraction d'information). En outre, à cause de leur caractère arbitraire et idiosyncratique, les collocations représentent un obstacle majeur dans l'apprentissage et dans l'enseignement des langues étrangères. Cela explique pourquoi nombre d'études sont publiées régulièrement sur le sujet, et ce par des linguistes aux spécialisations les plus variées. Les collocations ont fait l'objet de nombreuses recherches qui ont tenté de saisir leurs propriétés (parmi des synthèses plus récentes, citons Bartsch, 2004 ; Carter, Schmitt, 2004 ; Cowie, 1998 ; Grossmann, Tutin, 2003 ; Lo Cascio, 1998 ; Williams, 2003), de comprendre la façon dont elles doivent être traitées dans les dictionnaires traditionnels ou électroniques (citons entre autres Benson et al., 1986a ; Fontenelle, 1998 ; Hausmann, 1979 ; Lo Cascio, 2007, 2008 ; Mel'čuk, 1998), de proposer des théories nouvelles qui expliquent le fonctionnement du lexique mental et de l'apprentissage lexical (Lewis, 1993 ; Tomasello, 2003 ; Wray, Perkins, 2000). Cependant, à ce jour, personne n'est parvenu

à établir une définition unique du concept de collocation, la nature arbitraire du phénomène compliquant sans doute la tâche. Quant aux ressources existantes pour l'apprentissage et la didactique de la langue étrangère, bien qu'elles aient déjà permis un important pas en avant dans le traitement des collocations, elles sont encore loin d'être suffisantes.

Le concept de « collocation fondamentale », notamment, a reçu très peu d'attention. L'analyse conduite dans la présente étude élargit la notion de « vocabulaire fondamental »[1] à la dimension syntagmatique et propose une méthode pour identifier et définir ce que nous proposons d'appeler collocation fondamentale. Dans cette étude, nous avons considéré la collocation fondamentale comme l'unité lexico-grammaticale essentielle qui met en contexte de façon plus typique le sens d'un mot. Autrement dit, les collocations fondamentales sont envisagées comme des unités polylexicales significatives (unies par des liens collocationnels[2]) fréquentes (dans l'usage) ou non fréquentes (lorsqu'elles sont pertinentes[3]) qui représentent, pour les locuteurs natifs, les contextes de cooccurrence les plus essentiels et typiques d'un mot pivot donné.

Dans le passé, de nombreuses études[4] ont élaboré des listes de fréquence appelées listes de base, listes réduites, listes simplifiées, élémentaires, etc. dans lesquelles les mots du vocabulaire fondamental sont énumérés et décrits (*Le Français Elémentaire* -1954- est un ouvrage de pionnier pour la langue française). Aujourd'hui, ces études apparaissent incomplètes pour avoir négligé l'analyse des associations typiques et fréquentes qui opèrent en tant qu'unités lexico-grammaticales. Les mots de base sont en effet les plus connus chez les locuteurs natifs car ils leur permettent d'accomplir les actes du quotidien, mais ils n'existent pas en tant qu'unités isolées : bien au contraire, leur sens se réalise en cooccurrence, d'où la nécessité d'étudier les

[1] L'expression « vocabulaire fondamental » est utilisée d'après la dénomination de la part de l'équipe du *Français fondamental* (Gougenheim et al., 1964) du vocabulaire de base de la langue française, et désigne le noyau lexical dont chaque locuteur dispose pour ses actes communicatifs élémentaires et quotidiens.

[2] Les collocations sont extraites en utilisant des mesures associatives telles que l'Information Mutuelle.

[3] Gougenheim et al. (1964) appelaient unités « disponibles » les mots pertinents dans un certain domaine sémantique qui, malgré leur basse fréquence, étaient jugés comme essentiels par les locuteurs natifs.

[4] Pour un état de l'art détaillé sur le sujet, nous renvoyons au chapitre 2 de la présente étude.

associations qu'ils privilégient et les contextes d'occurrences où ils apparaissent le plus fréquemment.

L'étude vise à développer un outil d'extraction des collocations fondamentales, à tracer un profil de la collocation fondamentale à l'aide de critères extralinguistiques et linguistiques, et enfin, à préciser les implications pour la didactique du FLE. Ces trois objectifs trouvent leur origine dans une question unique qui est sous-jacente à l'ensemble de notre recherche : qu'est-ce qui permet d'identifier le caractère fondamental d'une unité polylexicale ?

Notre étude analyse un échantillon d'associations[5] fondamentales produites à partir de dix mots pivots qui sont fondamentaux d'après le *Dictionnaire Fondamental* de Gougenheim (1971), et qui représentent des « événements sociaux » d'après la consultation du *Thésaurus* de Daniel Péchoin (1991) : *colloque, conférence, congrès, conversation, débat, fête, interview, rencontre, réunion, séminaire.* Nous considérons comme étant un « événement social » toute occasion de rencontre avec d'autres personnes ayant lieu dans un processus temporel avec un début et une fin. Les collocations sont repérées dans le corpus *frWaC* (Baroni et al., 2010), un très vaste corpus écrit issu du web. Le corpus est analysé et exploré à l'aide d'un outil d'extraction développé par Olivier Kraif (2011), qui génère en sortie un tableau affichant des mesures statistiques (la fréquence, la dispersion et des mesures associatives telles que l'Information mutuelle) et un concordancier pour l'étude du contexte de cooccurrence des associations.

La méthode adoptée est hybride, étant donné qu'il s'agit d'une analyse qualitative étayée de mesures quantitatives. Nous avons eu recours aux divers outils suivants : un thésaurus, une liste de fréquence, un corpus, un outil d'extraction ainsi qu'un test auprès des locuteurs natifs.

Pour commencer, à partir des mots pivots choisis, et à l'aide de critères statistiques (qui sélectionnent les unités polylexicales significatives fréquentes et non fréquentes), nous avons trié dans le corpus les unités candidates au statut de

[5] Rappelons que, du point de vue statistique, les véritables collocations ne sont pas facilement distinguées des associations libres mais fréquentes. D'ailleurs, cette différence n'est pas toujours utile dans le domaine didactique et lexicographique. Ce point, qui fera l'objet d'une réflexion à l'issue de l'analyse, explique le fait que nous utilisons à la fois le terme « collocation » et les termes « association » ou « unités polylexicale ».

collocations fondamentales ; ensuite, nous avons demandé à quatre-vingt-dix locuteurs natifs du français de sélectionner les associations qui leur apparaissaient essentielles pour la communication, afin d'évaluer la justesse de l'échantillon constitué et de comprendre ce dont dépend l'assignation du caractère fondamental. Nous avons tiré les résultats de la recherche répondant aux questions suivantes :

- Quelle est la corrélation entre fréquence et jugement sur le caractère fondamental des associations tel qu'il est évalué par les locuteurs natifs ?
- De quel facteur autre que la fréquence dépend l'attribution du caractère fondamental de la part des locuteurs natifs ?
- Quelle est l'utilité de la fréquence, des mesures associatives et de la dispersion pour le repérage des collocations fondamentales ?

Enfin, nous avons dérivé quelques implications pour la didactique du FLE et nous avons tiré des considérations finales.

Pour conclure, nous précisons que notre travail veut trouver un équilibre entre deux approches qui caractérisent l'étude des collocations, l'approche phraséologique et l'approche statistique. L'approche phraséologique est généralement marquée par l'utilisation de critères linguistiques (syntaxiques et/ou sémantiques) distinguant les différents types d'unités phraséologiques, tandis que l'approche statistique donne priorité à la fréquence d'occurrence de l'association. Dans notre travail, nous tentons de ne pas faire prévaloir des raisons d'ordre statistique sur des raisons d'ordre linguistique.[6]

## 1.2 Repères théoriques

Dans ce qui suit, nous allons donner un cadre à la problématique de notre recherche. Nous présentons quelques brefs repères théoriques sur le vocabulaire fondamental et sur la collocation, ainsi que la nouvelle vision de l'usage langagier, reconnaissant la langue comme étant en grande partie conventionnelle et composée de « paquets lexicaux » et non de listes de mots isolés.

[6] Cf. chapitre 3 pour un approfondissement sur la différence entre approche statistique et phraséologique.

### *1.2.1 Le vocabulaire fondamental*

Qu'appelle-t-on « vocabulaire fondamental » d'une langue ? Il s'agit du noyau lexical d'une langue, le vocabulaire dont chaque locuteur dispose pour ses actes communicatifs élémentaires et quotidiens. Selon Goddard et Wierzbicka (2007), la notion de vocabulaire fondamental s'explique par l'idée que certains sens sont plus simples que d'autres et aident à expliquer les autres sens plus complexes.

En français, l'ouvrage de référence du vocabulaire fondamental est né d'une initiative de l'U.N.E.S.C.O. visant à diffuser les grandes langues de civilisation. Le Ministère de l'Education Nationale a attribué la réalisation de la liste *Le Français Elémentaire* (1954) à une commission spéciale, dont il a établi le centre d'étude à l'Ecole Normale Supérieure de Saint-Cloud, sous la direction de Gougenheim (Gougenheim et al., 1956), Professeur d'Histoire de la Langue Française à Strasbourg. *Le Français Elémentaire* devint *Dictionnaire Fondamental* en 1958, avec un total d'environ 3 000 mots. Il en résulta par la suite la constitution du *Français Fondamental 1er degré* (1959 ; Gougenheim et al., 1964) et du *Français Fondamental 2e degré* (1959).

Se fondant sur des corpus de langue orale, le *Français Fondamental* représente la liste de fréquence la plus connue et la plus significative, élaborée d'après de grandes enquêtes sur la langue orale (plus récemment, on y a intégré des données concernant la langue écrite). La liste a représenté, et représente encore, une œuvre de référence pour de nombreuses études sur le vocabulaire, notamment : l'*Inventaire thématique et syntagmatique du Français Fondamental* de Galisson (1971), le *LOB* (*Listes orthographiques de base du français* de Catach, publiées en 1984) et bien d'autres travaux. Dans *L'élaboration du français élémentaire* (Gougenheim et al., 1956), les auteurs affirment qu'il s'agit d'une langue qui ne diffère pas du français « normal » (p. 7) et qui ne se rapproche pas du *Basic English* (Ogden, 1930), car les bases pour un parcours d'acquisition du « français complet » y sont jetées. En outre, la liste s'adresse à un public « le plus large possible » et est avant tout un outil pour les professeurs qui enseignent à des adultes ou à des enfants (p. 8). Il est nécessaire ici de signaler le caractère relativement ancien de ces enquêtes et l'urgence d'une mise à jour des données. Les tentatives récentes de réflexion sur le sujet sont

malheureusement peu nombreuses. En outre, les nombreuses applications au niveau de la didactique des langues sont souvent mises de côté.

Un test de corrélation entre le vocabulaire de base et les facteurs qui facilitent l'apprentissage lexical[7] a montré qu'il existe une association entre les deux, étant donné que le vocabulaire de base offre de nombreux arguments pour être facilement appris (Benigno, 2007). Le vocabulaire de base ne doit pas cependant être considéré comme une liste d'items lexicaux, mais il doit être étudié en contexte. Les chercheurs reconnaissent que l'aisance en langue étrangère peut être attribuée principalement à la maîtrise des collocations : celui qui essaie d'apprendre une langue étrangère devrait donc commencer son parcours lexical par le vocabulaire de base et ses collocations les plus utiles afin de réussir dans les échanges quotidiens avec des locuteurs natifs. Carter (1998) affirme que les apprenants qui connaissent les mots de base ont à leur disposition un « survival kit [...] that they could use in any situation »[8]. En résumé, l'apprentissage lexical a son origine dans le vocabulaire fondamental, car il s'agit d'un processus de développement graduel qui va de l'appui sur la pragmatique à l'utilisation des structures syntaxiques. Comme le souligne Carter (1998, p. 171), le vocabulaire de base se compose de « lexical items which are the most central, 'nuclear' or core in the lexicon »[9]. Mais le vocabulaire de base n'est pas une liste de mots isolés : le bagage lexical d'une langue consiste en une variété de formules, *chunks*, phrases de routine, expressions figées, collocations. Le vocabulaire de base peut donc être envisagé comme la structure portante de la compétence lexicale, qui s'élargit de plus en plus dès que les rencontres avec la langue en contexte se font plus nombreuses et dès que la connaissance des associations lexicales d'un mot s'accroît. Les couches successives s'ajouteront par des liaisons privilégiées avec les mots que l'on maîtrise déjà. A l'aide de ces liaisons privilégiées, les mots se voient désambiguïsés.

[7] L'étude cite des facteurs linguistiques (la morphologie simple ou la facilité de prononciation) ainsi que des facteurs extralinguistiques (la fréquence ou le caractère prototypique).

[8] « Un kit de survie [...] qu'ils peuvent utiliser dans n'importe quelle situation » (notre traduction).

[9] « Eléments lexicaux qui sont les plus centraux, 'nucléaires' ou 'de base' dans le vocabulaire » (notre traduction).

L'ouvrage *Le Français Elémentaire* (1954) inclut les mots les plus fréquents, ainsi que des mots moins fréquents (pertinents par rapport à leur spécifique domaine sémantique) qui sont considérés comme essentiels par les locuteurs natifs. Des mots comme *fourchette* et *dent* mériteraient d'être inclus dans une liste fondamentale car, comme l'explique Gougenheim (1964, p. 145), ils sont « […] cependant des mots usuels et utiles ». L'équipe de Gougenheim appelait ces mots, « disponibles ». Ainsi, déjà dans les années 50, cette étude de pionnier révélait que le caractère fondamental n'était pas à attribuer uniquement à la fréquence. Nous reconnaissons le mérite considérable de cette découverte, mais nous voulons pousser plus loin l'analyse : nous nous intéressons à la dimension syntagmatique et nous essayons de comprendre ce dont dépend l'attribution du caractère fondamental, étant donné que la fréquence n'est pas un facteur stable.

### *1.2.2 Les collocations*

La langue est en grande partie conventionnelle. Ce principe découle d'une conviction désormais partagée par tous les linguistes. L'idée traditionnelle de la langue comme « récipient » ou « conteneur » d'unités lexicales simples (entretenant des relations syntaxiques et dont la somme donne un sens compositionnel) a été désormais remplacée par une nouvelle vision de l'usage langagier, qui reconnait que les phrases que tout locuteur produit se composent en grande partie de morceaux de langue stéréotypés.

La collocation peut être de façon générale définie comme une cooccurrence fréquente de mots dont les composants forment une unité lexicale de niveau supérieur et qui est soumise à des restrictions (de nature sémantique ou syntaxique) dues à l'idiosyncrasie[10] et à l'arbitraire de l'usage linguistique. Comme nous l'avons expliqué auparavant, il existe deux approches principales qui s'intéressent aux collocations : l'approche statistique et l'approche phraséologique. Les défenseurs de la première approche sont Firth (1957), Halliday (1985) et Sinclair (1991) parmi

[10] L'idiosyncrasie est la particularité de l'usage linguistique propre à chaque langue qui se manifeste dans des choix de sélection lexicale arbitraires et souvent non correspondant dans d'autres langues.

beaucoup d'autres. Ils identifient les collocations sur la base de la fréquence textuelle de l'association. D'un autre côté, parmi les représentants les plus significatifs de la seconde approche, nous trouvons Hausmann (1989), Mel'čuk (1998) et Cowie (1998), qui distinguent les collocations des autres types d'associations sur la base de critères syntaxiques et/ou sémantiques. Étant donné que, d'une part, les études purement statistiques sont nombreuses et sont facilement biaisées par le déséquilibre du corpus, et d'autre part, que les études purement phraséologiques adoptent des définitions ou des classements qui risquent d'être peu exhaustifs, toute catégorisation rigide se révèle, à notre avis, peu productive dans une perspective didactique. Reprenons les principales limites concernant les deux approches :

- Dans les approches purement statistiques, on court le risque de ne pas repérer des collocations peu fréquentes mais pertinentes, généralement spécialisées. Ces analyses n'intègrent pas la mesure statistique avec des informations syntaxiques et/ou sémantiques plus détaillées.
- Dans les approches purement phraséologiques, on court le risque contraire d'exclure des associations plutôt libres qui cependant sont très représentatives de l'usage linguistique. Ces analyses se basent sur des critères syntaxiques et/ou sémantiques précis.

Nous renvoyons au chapitre 4 de la présente recherche pour une justification de la définition de travail adoptée dans la présente étude : « la collocation est une séquence polylexicale qui actualise un mot dans une unité sémantico-syntaxique typique, et qui se caractérise par le sémantisme transparent de la base et le sémantisme restreint du collocatif ».

Dans ce chapitre introductif, nous nous limiterons à fournir quelques exemples qui aident à comprendre la notion. En français, on dit *prendre un bain,* tandis qu'en italien on dit *fare un bagno,* ce qui correspondrait par traduction littérale à **faire un bain ;* en français l'appétit est *petit,* en italien l'appétit est *pauvre (scarso appetito)*. Ces exemples mettent en évidence l'importance du choix du mot juste à combiner avec un autre mot : les collocations sont des groupements préférentiels de mots qui « sonnent » de façon naturelle à l'oreille des locuteurs natifs d'une langue, et qui reposent sur les connaissances et les attentes des membres d'une même communauté linguistique. L'opposition de Bally (1951) entre « contrainte de signe »

et « contrainte de contenu » explique la différence entre combinaison libre motivée et combinaison restreinte idiosyncratique, et explique l'importance de l'usage dans la production linguistique : bien que les adverbes *gravement* et *grièvement* soient synonymes, le premier se combine de préférence avec l'adjectif *malade*, et le second se combine uniquement avec l'adjectif *blessé*. Tout ceci est arbitraire : la « contrainte de signe » est de nature lexicale et bloque la possibilité de combiner les mots selon les règles sémantiques communes (« contrainte de contenu »). Sur ce blocage de la libre combinatoire lexicale s'ouvre un grand débat autour de la nature de la collocation, unité phraséologique à statut spécial qui, d'un côté, est soumise à des restrictions syntaxiques et sémantiques qui l'éloignent des combinaisons libres, et de l'autre, n'est pas autant figée que les expressions idiomatiques. Pour compliquer encore la donne, signalons qu'il existe d'autres unités phraséologiques qui se comportent également comme des paquets lexicaux indissociables et qui sont, elles aussi, difficiles à placer tout au long d'un continuum combinatoire imaginaire : les locutions, les mots composés, les formules pragmatiques, les proverbes. Il faut dire qu'à la différence des composants de la collocation, les composants de ces types d'unités phraséologiques ne sont pas autonomes et l'un n'est pas dépendant de l'autre : ils se situent au même niveau et forment un paquet lexical plus figé que celui de la collocation. Le présent travail ne s'intéressera cependant qu'aux collocations.

L'étude des collocations a favorisé le développement de l'idée selon laquelle la signification d'un mot est inférée de son usage, de ses occurrences textuelles réelles. Selon Firth (1957), la signification résulte des relations syntagmatiques qu'un élément entretient avec d'autres éléments d'un même énoncé, d'où sa célèbre phrase « you shall know a word by the company it keeps »[11] (p. 11). De même, Nation (2001) explique que la connaissance d'un mot implique plusieurs éléments, parmi lesquels la connaissance du comportement grammatical et des collocations d'un mot. C'est pourquoi les collocations favorisent les processus de désambiguïsation, étant donné qu'elles représentent les contextes de cooccurrence typiques d'un mot. En résumé, les collocations ont un statut spécial pour la compréhension des mécanismes d'acquisition lexicale.

[11] « On connait un mot par les mots qui lui tiennent compagnie » (notre traduction).

### *1.2.3 L'apprentissage par blocs*

Brent (2009) constate que l'acquisition partielle d'un mot est un cas très commun. Selon l'auteur, l'acquisition lexicale est un procédé de type « meaning-last » (« le sens à la fin ») dont l'acquisition du sens n'est que la dernière étape. Tout locuteur qui utilise pour la première fois un mot, l'utilise dans un contexte spécifique et ne comprend qu'une partie du sens de ce terme. En résumé, un mot ne serait d'abord compris que dans ses contextes lexicaux prévisibles. La connaissance partielle d'un mot serait fonctionnellement suffisante pour les besoins communicatifs essentiels, car nous recourons à un mot très souvent dans les mêmes contextes.

Selon Wray (2002), le nombre restreint de contextes dans lesquels un mot apparaît amène les locuteurs à acquérir les mots en associations typiques et comme des blocs non compositionnels. Ces blocs sont analysés lorsque la compétence des locuteurs s'accroît, et seulement si cela est nécessaire pour la réussite de la communication. Selon l'auteur, ce qui détermine le fait que l'acquisition lexicale s'arrête à une connaissance partielle du sens du mot, ou qu'à l'opposé, elle mène à une pleine compréhension, dépend de la productivité combinatoire, c'est-à-dire « [...] the ease with which words enter into collocational relationships with a wide variety of other words »[12] (p. 132). Plus grande est la productivité collocationnelle du mot, plus grand sera le nombre d'autres mots avec lesquels ce mot se combine. L'acquisition d'une langue est avant tout une question de nécessité pragmatique, et elle est poussée par la volonté d'optimiser l'efficacité de la communication, ce qui explique pourquoi on acquiert le plein sens d'un mot seulement si cela apparaît nécessaire pour l'interaction quotidienne. Si la réussite dans la communication quotidienne est garantie par la connaissance partielle d'un mot, son sens plein ne sera jamais appris.

Pour aller dans le même sens, on peut également citer Stubbs (2002), qui affirme que tout mot apparaît dans des contextes typiques, sélectionne d'autres mots et crée un univers spécifique du discours. Le sens est lié, d'une part, au contexte

[12] « [...] la facilité avec laquelle les mots instaurent un lien collocationnel avec un vaste ensemble d'autres mots » (notre traduction).

social, et d'autre part, au contexte linguistique. Les mots, en tant qu'unités isolées, sont polysémiques ou ambigus, tandis qu'une fois mis en contexte, ils perdent cette ambiguïté.

Le phénomène de la collocation est important d'un point de vue intralinguistique, mais l'est encore plus d'un point de vue interlinguistique, étant donné que des langues différentes ont des structures collocationnelles différentes. La connaissance des collocations est en effet propre aux locuteurs natifs : il n'y a aucune raison qui justifie le fait qu'en italien le *réchauffement* est *global* (*riscaldamento globale*), tandis qu'il est *climatique* en français ; ou encore, qu'en italien, l'*aller* est *seul* (*sola andata*), tandis qu'en français il est *simple*. Les locuteurs, suivant leurs intentions communicatives, combinent les mots selon des restrictions qui dépendent de règles de cohérence sémantico-conceptuelle et morphosyntaxique. Mais la langue est plus préfabriquée qu'on ne l'imagine : la plupart des phrases que nous produisons utilisent des blocs de langue préfabriqués dont les restrictions sont arbitraires, ce qui est le cas des collocations. Les collocations subissent des restrictions non seulement de nature sémantico-conceptuelle et morphosyntaxique, mais aussi des restrictions qui dépendent de l'usage, c'est-à-dire du fait que certains concepts sont préférentiellement exprimés par une association de mots donnée. Ainsi, les collocations seraient envisagées comme des expressions qui correspondent à une manière conventionnelle de dire (Manning & Schütze, 1999).

## 1.3 Méthodologie

Dans ce paragraphe nous allons expliquer la méthodologie adoptée : les étapes suivies pour la constitution, l'évaluation et l'analyse de l'inventaire des collocations fondamentales. La recherche est menée à l'aide d'outils de différents types correspondant aux différentes sections méthodologiques.

Les unités polylexicales ont été repérées dans le corpus *frWaC* (Baroni et al., 2010), un corpus traité issu du Web d'environ un milliard de mots, à partir de dix substantifs pivots[13] représentant la « tête » ou « base » de la collocation.

[13] Les dix substantifs indiquent des « événements sociaux » d'après le *Thésaurus* de Péchoin (1991) et ils sont fondamentaux, d'après la consultation du *Dictionnaire Fondamental* de Gougenheim (1971).

Des mesures associatives telles que l'information mutuelle ont été utilisées pour réduire les fausses cooccurrences dérivant du calcul de fréquence et pour obtenir une analyse plus fine. Le degré de précision de ces mesures dans l'extraction des collocations est étayé par une analyse linguistique manuelle plus approfondie. Le corpus a été interrogé à l'aide des scripts Perl développés par Olivier Kraif (2011). Ces scripts génèrent deux outputs : un tableau affichant les statistiques (fréquence, dispersion et mesures associatives) et une concordance. Nous n'avons retenu que les unités polylexicales les plus fréquentes, d'une part, et des unités polylexicales moins fréquentes mais pertinentes du point de vue de l'acte de communication, d'autre part. Ensuite, un test a été soumis auprès de locuteurs natifs pour évaluer l'échantillon constitué et comprendre quel type d'unités était perçu comme étant fondamental. Le test vise à découvrir la correspondance qui existe entre les unités polylexicales extraites par la procédure automatique et l'intuition native de quatre-vingt-dix locuteurs de nationalité française, suisse ou belge. Nous avons demandé aux locuteurs de répertorier les associations qui leur semblaient les plus familières et les plus essentielles pour la communication (des instructions les aidant à comprendre le concept d'association fondamentale). Cette procédure comparative nous a permis de filtrer l'échantillon extrait automatiquement, et de comprendre ce qui définit le concept de collocation fondamentale, étant donné que la fréquence n'est pas un facteur stable.

Dans ce qui suit nous allons nous attarder sur la valeur de la fréquence, de la dispersion ainsi que des mesures associatives d'une part, et du test soumis auprès des locuteurs natifs, d'autre part.

Le critère principal adopté pour l'élaboration des listes du vocabulaire de base a été la fréquence. La fréquence est un élément objectif, basé sur l'observation empirique de la langue d'usage (Fuster Márquez et B. Pennock Speck, 2008), et qui facilite l'apprentissage lexical (Laufer et Nation, 1995). Toutefois, la fréquence ne suffit pas pour constituer le vocabulaire essentiel d'une langue, et cela pour deux raisons :

- La fréquence n'exclut pas les liens de colligation (pour le mot *rencontre*, *la rencontre* aura une fréquence très élevée bien que l'association avec l'article défini soit peu pertinente) et n'élimine pas les cooccurrences non significatives ou les cas

erronés dérivant de la procédure d'extraction automatique. Nous avons donc utilisé, à côté de la fréquence, la dispersion et les mesures associatives. La dispersion mesure le nombre de textes différents dans lesquels un mot apparaît et suggère d'exclure d'une liste fondamentale les mots qui ne sont fréquents que dans un nombre restreint de textes. Les mesures associatives telles que l'Information mutuelle renseignent sur la force associative entre deux mots.

- Le caractère fondamental ne dépend pas uniquement de la fréquence. Comme nous l'avons expliqué auparavant, le vocabulaire de base d'une langue contient aussi des mots peu fréquents qui sont propres à certains contextes et que les locuteurs natifs considèrent comme utiles dans l'interaction quotidienne. Le critère de la disponibilité a été découvert lors de la constitution des premières listes de fréquence puisque l'on remarquait que certains mots ne ressortaient pas dans les comptages de fréquence malgré leur filiation avec des concepts élémentaires ou des actions quotidiennes. Nous allons appliquer ce raisonnement au concept de collocation fondamentale : à l'aide des mesures associatives, nous repérons aussi des unités polylexicales non fréquentes mais pertinentes. Cependant, les mesures associatives utilisées pour extraire les unités candidates au statut de collocations fondamentales nous informent sur la significativité de l'association et non sur leur utilité communicative, d'où la nécessité d'interroger des locuteurs natifs.

Nous avons donc effectué des enquêtes auprès de locuteurs pour combler les lacunes qui dérivent d'une étude purement statistique et pout jeter une lumière sur ce qui oriente les locuteurs dans leur attribution du caractère fondamental à des unités polylexicales fréquentes ou non. Comme l'affirme Stubbs (2002), la définition de ce qui est « fondamental » se base sur la fréquence, mais aussi sur des critères fonctionnels, par exemple sur qui est plus important au niveau communicatif pour les enfants ou pour les locuteurs non natifs. Etant donné que les collocations fondamentales ne sont pas uniquement des unités fréquentes, nous nous mettons à la recherche d'un autre type de fréquence, celle établie par ses locuteurs.

## 1.4 Apports scientifiques visés

Le but principal visé par la présente recherche est, en premier lieu, la définition d'un concept qui n'a pas encore été analysé par les chercheurs dans les termes que nous proposons. D'autres enjeux innovants sont représentés par les points suivants :

1) l'étude développe un outil d'extraction des collocations fondamentales ;
2) l'étude met en évidence les avantages dérivant de l'exploitation d'un corpus issu du web (surtout pour l'étude de phénomènes non rares) plutôt que ses limites ;
3) la méthodologie pourrait être réutilisée dans d'autres travaux portant sur des langues différentes ;
4) l'étude se pose des questionnements importants sur la didactique des langues étrangères, notamment pour le développement de matériaux didactiques adéquats.

## 1.5 Plan de la thèse

Notre recherche s'organise en huit chapitres. Le présent chapitre décrit la problématique de la recherche, les questions auxquelles nous tenterons de répondre tout au long de l'ouvrage, la méthodologie adoptée et les implications envisagées, ainsi que l'apport que ce travail désire offrir dans la communauté scientifique. Des repères théoriques sont aussi brièvement proposés.

Dans le second chapitre, intitulé « Le vocabulaire fondamental », nous définissons le concept de vocabulaire fondamental et nous présentons les études menées, avec un regard attentif sur la liste publiée dans *Le Français Elémentaire*.

Le troisième chapitre, intitulé « Excursus sur la collocation : théories et approches » propose un état de l'art des études les plus significatives sur les collocations.

Le quatrième chapitre, intitulé « Langage préfabriqué et collocations : une définition de travail » explique les critères adoptés pour définir la collocation fondamentale. Nous décrivons aussi la nouvelle vision de l'usage langagier, basée sur

l'idée que la langue repose sur des séquences préfabriquées, en argumentant ce point de vue.

Le cinquième chapitre, intitulé « Le corpus et l'outil d'extraction », décrit les ressources (ainsi que les démarches suivies) utilisées dans la recherche : le *Thésaurus*, le corpus *frWaC*, l'outil d'extraction développé (notamment les paramètres retenus et les opérations implémentées), et les mesures statistiques.

Le sixième chapitre, intitulé « Méthodologie : la constitution de l'échantillon et le test soumis auprès de locuteurs natifs », explique la méthodologie suivi pour la sélections des unités candidates au statut de collocations fondamentales et le test soumis auprès des locuteurs natifs.

Dans le septième chapitre, intitulé « Evaluation et résultats finaux », nous aborderons en détail les résultats obtenus.

Dans le huitième et dernier chapitre intitulé « Conclusions : contributions de l'étude et implications pour le FLE », nous présenterons les implications applicables à la didactique du FLE et présenterons les conclusions de la recherche.

## 1.6 Conclusions

Dans ce panorama aussi vaste que fascinant offert par l'étude des unités polylexicales, la définition du phénomène de la collocation est encore loin de faire l'unanimité : ses appellations sont nombreuses et l'on se trouve souvent en désaccord au sein de la communauté scientifique. L'une des explications de cette divergence d'opinion est probablement à chercher dans la nature complexe du phénomène ainsi que dans son idiosyncrasie.

La présente analyse se restreint au phénomène spécifique de la collocation fondamentale et vise à la définir à l'aide d'un outil d'extraction qui explore un très vaste corpus écrit issu du Web. Le concept de collocation fondamentale étant primordial dans le domaine didactique pour l'importance de la sélection du syllabus et du choix du lexique à enseigner, le travail dérive en conclusion des implications pour la didactique du FLE (cf. Grossmann et al., 2005 ; et Grossmann, Plane, 2008 pour approfondissement). En effet, dans le passé, la dimension lexicale a été trop souvent ignorée dans la didactique des langues étrangères, attachée à des méthodes

plutôt traditionnelles centrées sur l'enseignement de la grammaire hors contexte. Le vocabulaire fondamental, le noyau lexical que tout locuteur étranger est appelé à apprendre en premier pour accomplir les actes communicatifs de base, a été considéré comme une liste de mots isolés : les études conduites jusqu'à nos jours se sont limitées à en saisir les caractéristiques définitoires et à indiquer les éléments qui facilitent ou entravent son apprentissage. La dimension syntagmatique a été souvent négligée et les outils dont les enseignants ont pu se servir ont ignoré la façon dont l'apprentissage lexical a lieu, voir par la répétition de blocs de langue standardisés et par des associations privilégiées. A l'heure actuelle, l'existence d'un lexique-grammaire sous-jacent la production linguistique est reconnue à l'unanimité et il est évident que l'aisance en L2 peut être attribuée principalement à la maîtrise des collocations, élément clé de l'apprentissage lexical. Cependant, on constate encore une certaine « impréparation » à la mise à jour de la didactique du FLE, étant donné que la plupart des matériaux existants ne traitent pas les collocations de façon adéquate ou ne les traitent presque pas. Une réflexion assez limitée a en particulier concerné ce qu'il faut enseigner. Le lexique est un ensemble très vaste et les mots s'actualisent dans des contextes précis (et même dans des structures sémantico-syntaxiques plus marquées que d'autres). Parmi ses nombreux contextes, il y en a certains qui sont prioritaires. D'après cette considération, nous tenterons de traiter le concept de « collocation fondamentale » en vue d'une exploitation didactique. Il serait très utile pour la didactique du lexique d'avoir à disposition un inventaire des collocations fondamentales couvrant tous les actes communicatifs de base. Cependant, un tel objectif demanderait un énorme travail et ne correspondrait pas aux objectifs de la présente recherche. A travers l'analyse d'un domaine sémantique restreint, notre recherche désire montrer l'utilité du concept de collocation fondamentale en faveur de la didactique du FLE en soulignant l'importance de l'usage et de la fonction sociale de la langue. Les résultats issus de la recherche pourraient être repris dans une étude plus vaste qui vérifierait la possibilité de leur généralisation à tout l'ensemble des collocations fondamentales du français.

# CHAPITRE 2
# Le vocabulaire fondamental

Dans notre travail nous essayons de comprendre ce qu'est la collocation fondamentale, expression qui renvoie au vocabulaire fondamental. Nous commencerons ce chapitre en citant les premières tentatives pour produire des listes de vocabulaires simplifiées ou réduites, en particulier en français, en anglais et en italien, et nous nous intéresserons plus particulièrement au travail le plus connu et le plus significatif qui a marqué les études sur le vocabulaire fondamental en français, *Le Français Elémentaire,* publié pour la première fois en 1954 (paragraphe 2.1). Ce bref retour en arrière nous aidera à mieux comprendre le concept de vocabulaire fondamental, qui sera l'objet du paragraphe 2.2 : nous commencerons par les traits généraux invoqués pour le définir et nous expliquerons l'importance de quelques facteurs cruciaux dans son étude : la fréquence, la dispersion et la disponibilité. Du vocabulaire fondamental est issu le concept de collocation fondamentale qui fait l'objet de notre étude, et que nous présenterons dans le chapitre 4.

## 2.1 Bref excursus sur les listes de fréquence

Ce paragraphe ne se veut pas exhaustif en matière de description des listes de référence existantes pour le vocabulaire fondamental, tout d'abord parce qu'elles sont nombreuses (plus pour l'anglais que pour les autres langues), et ensuite, parce que leur étude ou leur comparaison n'est pas l'objectif de cette thèse. Nous limiterons donc nos références aux travaux les plus significatifs, surtout pour l'anglais, le français et l'italien.

### *2.1.1 Les premières listes de mots pour l'apprentissage des langues étrangères*

De nombreuses recherches sur le lexique ont montré l'importance de disposer d'un vocabulaire fondamental, c'est-à-dire d'un ensemble de mots réputés essentiels pour penser ou exprimer les concepts de base dans une langue. Un esprit fortement pédagogique a accompagné les premières tentatives de définition du vocabulaire fondamental, avec la naissance de listes de vocabulaire simplifiées ou réduites. Ces tentatives provenaient de la volonté d'offrir au locuteur d'une langue étrangère un vocabulaire réduit à l'essentiel qui lui simplifie la tâche d'interaction dans l'autre langue ; en résumé, il s'agissait d'offrir au locuteur non natif un bagage prioritaire pour l'apprentissage. Qu'il s'agisse de listes rédigées sur la base de considérations d'auteurs sur le caractère essentiel du vocabulaire (par exemple le *Basic English* d'Ogden, 1930) ou de listes rédigées à l'aide de mesures statistiques comme les comptages de fréquence (c'est le cas de la liste *French Word Book* de Vander Beke, 1929), toutes ces études sont nées en raison des besoins de l'enseignement de la langue étrangère ; ce n'est que par la suite qu'elles se sont élargies à d'autres domaines, qu'elles ont visé d'autres objectifs, tels que la mesure de la lisibilité des textes ou l'étude du style d'un auteur.

La thèse intitulée *Le vocabulaire fondamental du français. Etude pratique sur l'enseignement des langues vivantes* présentée dans les années 30 du siècle passé à la Faculté des lettres de l'Université de Paris par James Douglas Haygood (1937) vise à « déterminer l'étendue d'un vocabulaire *Français Fondamental* en vue de la lecture, et à choisir les mots qui doivent le constituer » (p. 14). Haygood analyse l'utilité des 2 000 mots les plus fréquents de l'œuvre de Vander Beke (1929) et de 69 mots-outils d'Henmon (1924), deux listes de fréquence très significatives que nous présenterons à la page suivante. Grace à une étude comparative de cinq ouvrages littéraires français du XIXe siècle, il affirme que ces 2 069 mots représentent en moyenne 90% du contenu d'un texte littéraire de difficulté modérée. Ils sont donc « suffisants » pour satisfaire les besoins de communication des étudiants de français langue étrangère (Richards et Savard, 1970, p. 27). Haygood signale, dans ses premières pages, l'actualisation d'un changement de l'enseignement des langues vivantes à la suite d'une réaction, venue d'Allemagne et plus tard des Etats-Unis, aux méthodes

didactiques traditionnelles[14]. Ce grand changement est identifié par l'utilisation de méthodes plus scientifiques et par l'apparition des premières enquêtes statistiques pour établir des listes de mots à enseigner en langue étrangère. Ci-dessous, nous allons présenter quelques ouvrages significatifs, parmi les nombreux cités par Haygood.

- Pour l'allemand, l'œuvre *Häufigkeitswörterbuch der Deutschen Sprache* publiée par Friedrich Wilhelm Kaeding en 1898 (cité par Haygood), d'après la liste de 80 000 mots différents rédigée par un groupe d'étudiants allemands pour l'enseignement de la sténographie, liste elle-même issue d'un corpus gigantesque de près de 11 000 000 de mots. Il semble y avoir un accord général sur l'importance de la liste de Kaeding pour le fait d'avoir établi une vraie et propre technique dans la constitution de listes de fréquence du vocabulaire.

- Pour l'anglais, l'œuvre *The Teacher's Word Book* de l'américain Edward Lee Thorndike (1921) qui consistait en 10 000 mots différents extraits d'un corpus d'à peu près 4 500 000 mots contenus dans des livres de lecture.

- Pour le français, l'œuvre *A french Word Book Based upon a Count of 400,000 Running Words* publiée aux Etats-Unis en 1924 par V. A. C. Henmon, et qui présentait une liste de fréquence d'à peu près 9 000 éléments différents sur la base d'un dépouillement de 400 000 mots tirés de textes littéraires.

Selon Haygood (1937, p. 10), les limites de ces travaux (les plus importantes étant l'absence d'un comptage effectué sur une plus grande échelle et une plus grande variété de textes) seront finalement surmontées grâce à l'adoption de méthodes plus scientifiques adoptées par les comités *American and Canadian Committees on Modern Languages*. Ces derniers publient :

- Le *Graded Spanish Word Book*, réalisé par M. A. Buchanan (1927, cité par Haygood) ;

- L'œuvre *French Word Book*, réalisée par G. E. Vander Beke en 1927 ;

[14] Haygood cite deux travaux qui témoignent de ce changement. En Allemagne, en 1882, William Viëtor publiait une critique des méthodes didactiques allemandes intitulée *Quo Usque Tandem ? Der Sprachunterricht muss umkehren (Il faut que l'enseignement des langues vivantes fasse volte-face*). Aux Etats-Unis, en 1898, le *Report of the Committee of Twelve of the Modern language Association of America*, publié par un comité de douze experts américains, avançait les mêmes critiques pour l'enseignement aux Etats-Unis.

- L'oeuvre *German Frequency Word Book*, réalisée par B. Q. Morgan (1928, cité par Haygood) sur la base de la liste de Kaeding.

### *2.1.2 Quelques travaux de pionniers*

Dans ce qui suit, nous citerons par ordre chronologique quelques travaux de pionniers parmi les plus significatifs pour l'anglais, le français et l'italien, sans aucune prétention d'exhaustivité, puisque les listes produites sont nombreuses.

- Pour le français[15]:
  1) La liste de fréquence *A french Word Book* (1924), publiée aux Etats-Unis par V. A. C. Henmon.
  2) Le travail *French Word Book* de Vander Beke (1929) qui contient une liste de 6 067 mots repérés dans un corpus varié (romans, journaux, textes de science, etc.) de 1 147 748 mots. La liste indique la répartition ainsi que la fréquence. On y trouve des mots à la fois littéraires et archaïques (Richards et Savard, 1970). Cependant, elle représente une œuvre de pionnier qui inspira aussi les auteurs du *Français Fondamental*.
  3) Le travail *Le vocabulaire fondamental du français. Etude pratique sur l'enseignement des langues vivantes* de Haygood de 1937.
  4) *Le Français Elémentaire* (1954), mieux connu sous le nom « Français Fondamental », rédigé par Georges Gougenheim et son équipe. Nous consacrerons à cet ouvrage un paragraphe distinct, puisqu'il s'agit incontestablement d'un grand travail de référence pour le français.
- Pour l'anglais[16] :
  1) La liste *The Teachers' Word Book* de E. L. Thorndike (1921)

[15] Parmi les travaux les plus récents pour le français, citons par exemple les *Listes orthographiques de base* de Nina Catach (1985), la liste de fréquence du *TLF* développée à Nancy (*Dictionnaire des fréquences*, 1971), le *Frequency Dictionary of French Words* de Juilland et al. (1970), l'*Echelle Dubois-Buyse d'orthographe usuelle française* de Mayer, Reichenbach, Ters (1969).

[16] Tous ces auteurs à l'exception de Thorndike (1921), construisirent leur liste sans adopter une authentique méthode scientifique fondée sur des principes fiables, mais en s'appuyant plutôt sur leur expérience en tant qu'enseignants de l'anglais langue étrangère.

2) La liste *Basic English* (Ogden, 1930), où « Basic » indique les initiales de « British », « American », « scientific », « international », « commercial ». Elle est utilisée au début de la seconde guerre mondiale sur tous les continents comme langue simplifiée universelle à base d'anglais (Richards et Savard, 1970, p. 23). Comme l'explique Carter (1998), il s'agit d'une liste de 850 mots qui serviraient à exprimer des idées plus complexes par paraphrase et qui devait amener au développement d'un anglais « général », même s'il n'indique pas comment la liste prétend arriver à ses fins[17]. En effet, il s'agit d'une tentative pour constituer une langue un peu artificielle capable d'exprimer les idées de base en quelques centaines de mots (qui ne sont pas nécessairement les plus fréquents).
3) La *General Service List* de M. P. West (1953) qui avait été publiée pour la première fois en 1936 en compétition avec le *Basic English* d'Ogden (1930). West, qui avait travaillé au Bengale en Inde, destine cette liste à l'enseignement de l'anglais dans les écoles situées dans des pays non anglophones. La liste comprend à peu près 1 490 mots parmi les plus fréquents. Elle a été établie sur des critères tant statistiques que subjectifs (l'utilité des mots, leur universalité, etc.), et elle fournit une distinction des sens, une annotation des fréquences et une annotation des significations que l'apprenant connaît à chaque étape de son apprentissage. Ce travail est issu des rapports *Interim Report*[18] (Faucett et al., 1936) dus aux efforts de Palmer en collaboration avec West, Faucett, un expert qui avait beaucoup travaillé en Chine, et Thorndike (Faucett et al., 1936, cités par Richards et Savard, 1970, p. 24-25). Selon Gougenheim et al. (1964, p. 29), la liste de West, tout en adoptant la même recherche de mots définissants que le *Basic English*, est pourtant supérieure à celle du *Basic English*, car elle réussit à définir tous les

[17] Les règles grammaticales sont réduites à l'essentiel, par exemple seulement le pluriel en « s » est envisagé. Le même principe vaut pour le vocabulaire : la liste comprend seulement 18 verbes, dont par exemple *want* qui est paraphrasé par *have a desire for*. En outre, elle n'inclut pas des termes comme *Goodbye* ou *thank you*. Selon Carter (1998), ce système simplifié a des limites, dont celui d'être peu naturel. Considérons la phrase *I have a desire to go out :* elle est moins naturelle que la phrase *I want to go out*. Toutefois, la valeur de la liste est en général reconnue.

[18] L'*Interim Report on vocabulary Selection for Teaching English as a Foreign language* (Faucett et al., 1936) est connu aussi comme *Carnegie Report* parce que le travail fut financé en partie par la Carnegie Corporation.

mots du vocabulaire anglais. Cependant, « un vocabulaire de `définissants` n'est pas un vocabulaire de base », expliquent les auteurs (p. 30). Comme l'explique Carter (1998), cette liste constitue aujourd'hui la base pour le *Longman Dictionary of Contemporary English -LDOCE-* (1978).

4) La liste *Thousand-Word English* (1937) de H. E. Palmer et A. S. Hornby, élaborée en 1937 au Japon et commissionnée par l'organisme *I.R.E.T.* (« Institute for Research in English Teaching »), qui était censée établir un vocabulaire anglais destiné à des élèves japonais d'onze à seize ans (Richards et Savard, 1970, p. 28).

Pour l'italien[19] :

1) La liste de Thompson *A study in Italian vocabulary frequency*, publiée en 1927 (cité par D'Agostino, 1998) et fruit d'une thèse de doctorat. La liste comprenait environ 500 mots extraits de textes de lecture en italien publiés aux Etats-Unis pour un total de 100 000 mots.
2) La liste *An Italian Word list from literary sources* élaborée de T. C. Knease en 1933 (cité par Russo, 1947). Il s'agit d'une liste de 2 080 mots d'après le dépouillement de textes littéraires pour un total de 400 000 occurrences et qui tient compte de la fréquence et de la répartition.
3) La liste *Der grundlegende Wortschatz des Italienischen. Die 1500 wesentlichsten Worter* de Migliorini (1943), issue elle aussi, d'une thèse de doctorat. Elle inclut 1 500 mots fondamentaux (parmi lesquels aussi des mots grammaticaux) et leur traduction en allemand. Elle a été composée selon l'intuition personnelle de l'auteur et pour des buts fortement didactiques.
4) Le *LIF* (*Lessico di frequenza della lingua italiana contemporanea*), élaboré par Bortolini, Tagliavini e Zampolli (1971). Il est constitué de 5 356 mots différents d'après le dépouillement d'un corpus de 500 000 occurrences (lemmes) très équilibré (cinq typologies textuelles de 100 000 occurrences chacune). Le corpus comptait la fréquence, la dispersion ainsi que l'usage (la fréquence multipliée par la dispersion).

[19] Parmi des ouvrages plus récents, le *Frequency Dictionary of Italian words* de Juilland, Traversa (1973), le *Vocabolario fondamentale della lingua italiana* de Sciarone (1977), le *Livello soglia per l'insegnamento dell'italiano come lingua straniera* de Galli de' Paratesi (1981), etc.

5) Le *Vocabolario di base della lingua italiana* (2003), dont la première édition fut en 1980 et élaboré par une équipe guidée par le linguiste Tullio De Mauro à partir du *LIF* (et qui a connu de nombreuses rééditions). Pour l'italien, c'est la liste de référence du vocabulaire fondamental la plus connue. Elle comprend à peu près 7 050 lemmes divisés en trois « couches » : le « vocabulaire fondamental », le « vocabulaire de haut usage » et le « vocabulaire de haute disponibilité » (des mots qui ne sont pas très fréquents mais qui se révèlent fondamentaux pour les actes communicatifs quotidiens). La liste a été constituée sur la base de trois critères : la fréquence, l'usage, la disponibilité.

### *2.1.3 Le Français Fondamental*

De cet important ouvrage il existe différentes versions, éditions et dénominations, que nous tenterons d'analyser dans ce paragraphe. Il s'agit de la liste de fréquence la plus connue et la plus significative élaborée à partir de grandes enquêtes sur la langue orale, et à laquelle ont été intégrées plus tard, des données concernant la langue écrite. Le parcours qui mène du *Français Elémentaire* (1954) à la dernière édition du *Dictionnaire fondamental* (Gougenheim, 1971) a été assez complexe :

- L'ouvrage communément connu sous le nom de *Français Fondamental* est publié pour la première fois par le Centre National de Documentation Pédagogique du Ministère de l'éducation nationale sous la dénomination de *Français Elémentaire* en 1954. L'ouvrage est né d'une initiative de l'U.N.E.S.C.O. visant à diffuser les grandes langues de civilisation. Le Ministère de l'Education Nationale a attribué sa réalisation à une commission spéciale, dont il a établi le centre d'étude à l'Ecole Normale Supérieure de Saint-Cloud, sous la direction de Gougenheim, Professeur d'Histoire de la Langue Française à Strasbourg.
- En 1956 Gougenheim, Michéa, Rivenc et Sauvageot publient *L'élaboration du français élémentaire. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*
- La liste du *Français Elémentaire* devenait *Dictionnaire fondamental* en 1958 et s'élargissait d'environ 1700 nouveaux mots pour un total d'environ 3 000 mots. Selon les mots de Gougenheim :

« Les mêmes matériaux (enquête statistique sur le français parlé, enquête sur les centres d'intérêt) avaient été utilisés. Cependant, estimant qu'il convenait de tenir compte de la langue écrite, nous avions utilisé aussi le dépouillement de textes écrits (livres, pièces de théâtre, journaux) effectué par George E. Vander Beke » (Gougenheim, 1971, p. 5).

- Un an plus tard, en 1959, cent nouveaux mots étaient ajoutés et les auteurs scindaient l'ouvrage comme suit : *Français Fondamental 1er degré* d'une part, et le *Français Fondamental 2e degré* d'autre part. Chaque degré comptait 1 500 mots. Les 1 500 mots du second degré étaient choisis avec des principes analogues à ceux qui avaient été adoptés pour le choix des 1 700 mots en 1958.
- En 1971 les auteurs publient une nouvelle édition du *Dictionnaire fondamental* pour harmoniser ce dernier et les deux degrés réunis. Les auteurs expliquent que :

« [...] aucun mot de la première édition du *Dictionnaire fondamental* n'a été supprimé dans cette nouvelle édition, même s'il ne figure dans aucun des deux degrés du *Français Fondamental* » (Gougenheim, 1971, p. 5).

Et encore :

« Les mots du *Français Fondamental* qui ne figuraient pas dans notre première édition ont été ajoutés, ce qui nous a amené à rédiger 460 articles entièrement nouveaux. Une centaine d'autres articles ont reçu des additions plus ou moins importantes » (Gougenheim, 1971, p. 5).

Cette dernière édition de 1971 est publiée sous la forme d'un dictionnaire : des définitions sont rédigées à l'aide de « définissants », les différentes acceptions sont distinguées, et les expressions et les proverbes sont ajoutés.

Tentons maintenant de résumer les caractéristiques principales de cet ouvrage de référence pour la langue française :
- « Il n'est pas une langue différente du français 'normal' » (*Le Français Elémentaire*, 1954, p. 7).

- Il est différent du *Basic English*, puisqu'il offre des « moyens d'expressions suffisants » et les bases pour un parcours d'acquisition du « français complet ». Il s'agit d'un point essentiel : les auteurs du *Français Elémentaire* déclarent ici vouloir s'éloigner de la « simplicité » du projet du *Basic English* (*Le Français Elémentaire*, 1954, p. 7).
- *Le Français Elémentaire* s'adresse à un public « le plus large possible ». C'est avant tout un outil pour les maîtres qui enseignent à des adultes ou à des enfants (*Le Français Elémentaire*, 1954, p. 8).
- *Le Français Elémentaire* présente une liste de 1 138 mots (en grande partie des mots lexicaux), rédigée d'après le dépouillement et la transcription de cent soixante-trois conversations recueillies à l'aide de magnétophones et correspondant à un total de 312 000 occurrences (et d'à peu près 8 000 mots différents). Nous précisons que dans l'opuscule du *Français Elémentaire*, la liste de fréquence dressée n'était pas montrée pour ne pas « influencer les maîtres » (p. 11). Dans l'édition de 1956, une liste de mots classés par fréquence décroissante et une liste de mots classés par ordre alphabétique sont présentées. Les 1 138 mots du *Français Elémentaire* atteignent, dans la dernière mise à jour de la liste (le *Dictionnaire fondamental* de 1971), le nombre de 3 500.
- La liste comprend des mots fréquents ainsi que des mots disponibles, c'est-à-dire des mots peu fréquents mais essentiels pour la communication : pour repérer ces mots l'équipe de Gougenheim effectue une enquête dans des écoles et demande aux élèves quels mots (appartenant à seize centres d'intérêt principaux et à d'autres secondaires, comme par exemple « le mobilier ») sont considérés comme étant les plus utiles.
- L'œuvre présente une grammaire « conçue uniquement comme un ensemble de prescriptions délimitant ce qui devra être enseigné et ce qui pourra être laissé de côté ou remis à plus tard » (*Le Français Elémentaire*, 1954, p. 11).

Dans la présentation de l'opuscule *Le Français Elémentaire*, le Ministre de l'Education Nationale André Marie, s'exprime de la façon suivante :

« Voici donc terminée l'œuvre d'élaboration du Français élémentaire. Si cette entreprise, recommandée au Gouvernement par la Commission de la République Française pour

l'Education, la Science et la Culture, et dont je me félicite d'avoir, grâce au soutien des assemblées parlementaires, assuré l'accomplissement, a suscité au départ certaines inquiétudes, et s'il n'est pas possible d'être assuré que sa réalisation recueille tous les suffrages, je ne doute point que cette œuvre de bonne foi, respectueuse du génie de notre langue et au service de son rayonnement, ne recueille l'hommage unanime des esprits non prévenus » (1954, p. 3).

On remarque que l'importance attribuée à l'entreprise du *Français Fondamental* est accompagnée par la crainte que des esprits « prévenus » puissent s'acharner contre un travail « de bonne foi ». Il est vrai que, comme tous les travaux de pionnier, l'ouvrage a été très critiqué : pensons au petit volume de Marcel Cohen de 1955 intitulé *Français Elémentaire ? Non.* L'auteur, entouré d'autres illustres linguistes de l'époque, y exprime un refus catégorique de cette entreprise et de sa signification politique et sociale : le *Français Elémentaire* était considéré comme un travail imparfait qui avait abouti à une langue imprécise et réduite, ainsi comme une tentative d'imposer la langue française sur les langues locales. Malgré cette critique très sévère, le caractère novateur et les mérites de cet ouvrage sont indiscutables, et personne ne pourra nier la valeur de la méthode adoptée et l'importance de l'effort mené pour le recueil d'interviews orales, surtout avec les moyens dont on disposait à l'époque.

### *2.2 Définition du concept*

Qu'est-ce que le vocabulaire fondamental d'une langue ? C'est le noyau lexical d'une langue, le vocabulaire dont chaque locuteur dispose pour ses actes communicatifs élémentaires et quotidiens. L'exigence de retrouver les mots les plus centraux de la langue dérive de la conviction qu'il existe un noyau lexical essentiel dont on peut se servir pour comprendre et communiquer des idées de base dans la vie quotidienne : ceci explique les nombreuses tentatives pour rédiger des listes simplifiées, réduites, élémentaires, etc. (voir paragraphe 2.1).

Mais est-il réellement possible de trouver un noyau central et stable valable dans toutes les situations communicatives ? La réponse est non, et ce pour un motif

très simple : l'apprentissage du vocabulaire est un parcours progressif et soumis à des variations individuelles qui dépendent des expériences subjectives de chaque individu. Par exemple, le bagage lexical d'un astronome inclut des termes de la langue spécialisée qu'un plombier ne connait pas (mais ce dernier en connait d'autres que l'astronome ignore, et dont il se sert dans son métier, tandis que ses expériences de vie personnelles lui en font utiliser encore d'autres). L'apprentissage du lexique d'une langue est un procès articulé, tant pour les locuteurs natifs que pour les locuteurs étrangers (qui ont souvent le désavantage d'avoir des « rencontres » limitées avec l'usage de la langue cible). Un inventaire idéal et stable de mots fondamentaux est donc impossible à rédiger. Le vocabulaire fondamental devrait être envisagé comme un ensemble de mots non exhaustif et qui ne peut en aucune manière prétendre couvrir tous les contextes d'expériences et tous les besoins de l'expression langagière.

Cependant, on peut se rapprocher beaucoup de cet idéal. On peut rédiger une liste de mots qui se retrouvent dans tout type de contexte (les mots très fréquents) ou une liste de mots essentiels (qui inclut les mots très fréquents mais aussi les mots peu fréquents mais importants pour la conceptualisation de base, appelés « disponibles » par Gougenheim) ; ou encore, on peut tenter de rédiger une liste qui soit idéale pour le but qu'on poursuit (lexicographique, pédagogique, orthographique, etc.).

Bien que le vocabulaire de base ne soit pas un ensemble cohérent, des traits généraux peuvent en être extraits (voir paragraphe suivant). Par exemple, pour l'italien, Thorton, Iacobini et Burani (1997) présentent des données quantitatives et aussi qualitatives : la distribution dans le vocabulaire de base des catégories grammaticales, des classes de flexion, du genre, des affixes ; ou encore les structures accentuelles prédominantes, la longueur moyenne en syllabes, etc.

Dans le prochain paragraphe, nous indiquerons des traits qui peuvent aider à trouver les mots qui appartiennent au vocabulaire de base d'une langue, tandis que dans les paragraphes successifs nous discuterons brièvement l'importance de trois autres facteurs qui ont été beaucoup utilisés pour la constitution des listes de fréquence du vocabulaire de base : fréquence, dispersion et disponibilité. Dans notre étude, ces trois facteurs contribueront à la définition de collocation fondamentale.

*2.2.1 Traits généraux*

Carter (1998) explique qu'il est possible de soumettre un mot à des tests pour vérifier si celui-ci appartient au vocabulaire fondamental. Ces tests, qui utilisent des informateurs natifs, essaient de relativiser la subjectivité de l'intuition personnelle et aident à trouver les traits généraux (formels ou non) qui caractérisent le vocabulaire de base. L'auteur envisage la possibilité que les tests se réduisent en nombre (à cause de leur recoupement partiel), qu'ils soient améliorés, hiérarchisés ou intégrés par d'autres tests. Ci-dessous nous en proposons une brève synthèse :

1) Des tests portant sur les relations syntaxiques et sémantiques, qui permettent de savoir si un mot est plus intégré que d'autres dans le système linguistique :
   a) Test de « substitution syntaxique » (« syntactic substitution »)

      Un mot fondamental ne peut pas être remplacé par d'autres mais il peut se substituer à d'autres mots. Le mot *eat,* par exemple, qui exprime un trait sémantique de base de *dine* et *devour,* peut se substituer à eux, mais il ne se laisse pas définir par ces mots. Un test a montré que quatre-vingts pour cent d'informateurs à qui on demandait de définir *guffaw, chuckel, giggle, laugh, jeer, snigger* utilisaient le mot de base *laugh* pour expliquer ces termes. Ce sont évidemment les mots génériques ou superordonnés qui passent le test de la substitution syntaxique.
   b) Test d' « antonymie » (« antonymy »)

      Plus un mot est fondamental, plus il est probable qu'il ait un antonyme : considérons par exemple l*augh/cry* et *fat/thin.* Par contre, il est plus difficile de trouver l'antonyme de mots non fondamentaux tels que *skinny*.
   c) Test de « collocabilité » (« collocability »)

      Plus un mot est fondamental, plus il s'associe avec d'autres mots : *bright* peut être associé à *sun*, *light*, *sky*, *idea*, *future*, *colours* et d'autres mots encore ; tandis que *gaudy* ne trouve qu'un seul collocatif indiscutable, *colours*. Ces tests doivent toujours tenir compte des effets stylistiques et devraient toujours être conduits sur la base de mesures de fréquence.

d) Test d' « extension » (« extension »)

Les mots « nucléaires » (terme utilisé par Stubbs, 1986, cité par Carter) ont une propriété d'extension, c'est-à-dire qu'ils ont un degré d'association très élevé. Ce test est similaire au précédent mais inclut tout type de relation syntagmatique : collocations, noms composés, phrases idiomatiques, etc. Stubbs remarque que dans le *Collins English Dictionary* (CED) on trouve *well* en 150 associations différentes.

e) Test de l'« hyperonymie » (« superordinateness »)

Un hyperonyme est très souvent un mot fondamental par rapport à ses hyponymes : *flower*, par exemple, est un mot ressenti comme générique et non marqué par rapport à *tulip* et *rose*. Toutefois, l'auteur fait remarquer que la fonction connotative de la langue peut parfois renverser cette relation et produire un hyperonyme plus marqué que ses hyponymes. Ce test est donc à perfectionner.

2) Des tests de neutralité, qui montrent si un mot est plus neutre ou moins expressif que d'autres dans un contexte pragmatique et discursif :

a) Test de « neutralité culturelle » (« culture-free »)

Plus un mot est fondamental, plus il est culturellement neutre. Cela expliquerait pourquoi l'anglais, par exemple, emprunte à d'autres langues des mots comme *pouf* ou *chaise-longue*, mais non des mots essentiels comme les parties du corps.[20]

b) Test du « résumé (« summary »)

Les mots fondamentaux sont les plus utilisés dans les résumés. Carter cite des enquêtes conduites par Stubbs (1982) sur les mots utilisés par un groupe d'informateurs dans leurs résumés du récit *Cat in the rain* d'Hemingway : le mot *cat* était à l'unanimité préféré à *kitty* et *feline*. Cela est aussi un indicateur de la neutralité stylistique, rhétorique ou évaluative du résumé en tant que genre textuel.

[20] Nous suggérons d'être prudents lorsque l'on définit le vocabulaire de base comme culturellement neutre. En effet, les mots s'actualisent par des liaisons et des associations privilégiées et les sens qu'ils expriment en contexte jouent le rôle d'indicateurs des préférences d'usage spécifiques d'une langue par rapport à une autre : en italien, par exemple, les associations autour du mot *pasta* sont nombreuses et fréquentes, ce qui n'est pas le cas dans une autre langue où cet aliment n'est pas fondamental dans les habitudes alimentaires de ses locuteurs.

c) Test d'« association (« associationism »)

Ce test dérive de l'échelle de différenciation sémantique d'Osgood et al. (1957, cité par Carter) et demande aux informateurs de placer les mots tout au long d'un continuum sémantique à échelle (qui va par exemple de l'informel au formel), et d'y attribuer une valeur dans une échelle établie. Il en résulte que les mots de base se situent au milieu des échelles évaluatives, c'est-à-dire qu'ils sont les plus neutres. Un test mené par Carter montre que le mot *thin* se classe au centre de l'échelle, tandis que les mots *emaciated, skinny, lean, slender, slim, weedy* ont une force associative plus forte et se placent très différemment selon l'informateur.

d) Test de « neutralité de discours » (« neutral field of discourse »)

Les mots fondamentaux ne concernent pas un sujet spécifique, par exemple la médecine. Comme Carter l'explique, en anglais *galley* relève du domaine nautique tandis que *kitchen* n'appartient pas à un domaine spécifique, même si les deux sont des cuisines, excepté que l'une est une cuisine de bord et l'autre une cuisine « standard ».

e) Test de « neutralité de registre » (« neutral tenor of discourse »)

Les mots fondamentaux sont neutres en ce qui concerne l'opposition informel-formel. Par exemple *fat* est jugé par les informateurs comme plus neutre que *podgy* et *corpulent*.

Pour conclure, le vocabulaire fondamental d'une langue apparaît comme générique, neutre, non marqué. Nous précisons que selon Carter, aucun test ne peut seul suffire en tant que mesure statistique du caractère fondamental d'un mot, et le vocabulaire de base ayant des contours flous et ambigus, il vaudrait mieux discuter en termes d'échelle. A notre avis, les facteurs servant à définir le vocabulaire fondamental sont de différentes natures et il faudrait leur attribuer une valeur différente selon l'objectif de la sélection du vocabulaire (lexicographique, pédagogique, etc.). Quoi qu'il en soit, certains traits reçoivent une attention constante de la part des rédacteurs de listes de vocabulaire fondamental : ce sont la fréquence, la dispersion et la disponibilité. Ces traits seront pris en compte dans la présente recherche pour définir ce qu'est une collocation fondamentale.

### *2.2.2 La fréquence, moteur de la mémorisation des items lexicaux*

La fréquence est un élément basé sur l'observation empirique de la langue d'usage qui sert à identifier des éléments de type prototypique, général, non marqué. Selon Nattinger et DeCarrico (1992), la fréquence serait un facteur primaire dans l'apprentissage car la « ritualisation » joue un rôle fondamental dans l'acquisition du langage comme dans tous les autres types de comportement humain. La fréquence facilite l'apprentissage lexical dans la langue maternelle comme dans une langue étrangère puisqu'elle favorise les procès de mémorisation, comme cela a été souligné par de nombreuses recherches en neurolinguistiques. Selon l'hypothèse du niveau seuil de Paradis (2004), l'activation d'un item dans le lexique mental dépend de la quantité d'impulsion neurale reçue, qui à son tour dépend de la fréquence d'occurrence de l'item. Lorsqu'un item lexical est activé, son niveau seuil s'abaisse et cela facilite sa réactivation. D'après Tomasello (2003), la fréquence est un facteur crucial dans l'apprentissage lexical. Elle explique pourquoi des constructions irrégulières, et donc difficiles à apprendre de la part d'un locuteur étranger, sont malgré tout mémorisées : cela dépendrait du fait qu'elles sont souvent utilisées et donc apprises comme des blocs.

Tous les rédacteurs de listes simplifiées ou réduites se sont servis de la fréquence. De même, la fréquence est la première caractéristique invoquée dans toute définition du vocabulaire fondamental. Une étude très significative sur les indices d'utilité du vocabulaire fondamental français a été effectuée au Québec par Jack Richards et Jean-Guy Savard en 1970. Ouvertement animés par la volonté d'établir l'utilité des mots fondamentaux de la langue française par l'étude de la fréquence, de la répartition, de la valence[21] et de la disponibilité, les auteurs ont voulu offrir un guide aux enseignants du français langue étrangère. Leur but était celui d'« aider ces enseignants à mettre de côté leurs hésitations et leurs incertitudes » tout en soulignant qu'« un vocabulaire restreint n'est qu'un outil qui vaudra ce que vaudront les ouvriers qui consentiront à l'utiliser » (p. 3). Selon

[21] « Elle permet de mesurer l'utilité des autres éléments du vocabulaire pour des fins précises : quelques mots sont utiles parce qu'ils peuvent remplacer d'autres mots » (Richards, Savard, 1970, p. 10).

Richards et Savard, « sur le plan linguistique les éléments choisis pour l'apprentissage d'une langue seconde doivent correspondre aux éléments de la langue orale et écrite qui du point de vue statistique sont les plus fréquents » (p. 9). Cependant, comme les auteurs le font remarquer, la fréquence apparaît inadéquate pour mesurer le vocabulaire de situations spécifiques : rappelons que certains mots de basse fréquence sont en réalité fondamentaux pour l'interaction quotidienne bien qu'ils n'apparaissent que dans un ou deux contextes particuliers, et que pour cette raison, ils ne sont jamais pris en compte dans les comptages de fréquence. Ce sont les « mots de disponibilité », des mots forts importants comme *fourchette* et *dents,* dont on ne se sert que rarement en production : la fourchette est un instrument qu'on utilise tous les jours, mais qu'on nomme bien moins souvent qu'on n'en fait usage. Le même vaut pour *dents*, dont on n'en parle que si on en souffre. La fréquence n'est pas un facteur stable, sauf dans les mots les plus fréquents et les plus génériques qui ne sont pas influencés par un contexte de situation particulier. Tous les autres mots ont une fréquence assez variable. Il est donc nécessaire, dans la constitution d'un vocabulaire fondamental, de faire usage d'un certain empirisme pour insérer les mots peu fréquents mais nécessaires à une conceptualisation de base de la langue. Comme l'expliquent Gougenheim et son équipe (1964 p. 154), tenir compte uniquement de la fréquence dans la constitution d'un vocabulaire de base signifierait « se condamner à faire une œuvre fragile et subjective » et à ignorer les mots concrets qui, à cause de leur faible usage linguistique, sont exclus des listes de fréquence. En résumé, la valeur de la fréquence comme élément pour la sélection du vocabulaire fondamental n'est pas remise en question, mais la fréquence en elle-même n'est pas suffisante pour sélectionner le vocabulaire essentiel d'une langue.

### *2.2.3 La disponibilité, mesure de l'utilité communicative*

Pour établir des listes fondamentales, Gougenheim et son équipe (Gougenheim et al., 1956) ont voulu intégrer à la fréquence, la disponibilité. Comme nous l'avons déjà expliqué, les mots disponibles sont des mots concrets plutôt rares, mais liés à des circonstances ou à des thèmes de conversation spécifiques de la vie quotidienne. Ce

ne sont pas des mots techniques, car les mots *dents* et *fourchette* (pour reprendre les exemples déjà utilisés) sont connus même par un enfant de trois ans. Ce qui est plus important est qu'ils sont issus exclusivement du dépouillement de corpus oraux et surtout pas écrits. Les auteurs du *Français Elémentaire* repérèrent donc les mots disponibles auprès des locuteurs, comme cela a également été fait plus tard par l'équipe du *Vocabolario di base* (De Mauro, 1980). L'équipe du *Vocabolario di base* a repéré les mots de « haute disponibilité » de façon un peu « brute » : après une sélection manuelle dans le dictionnaire *Zingarelli* des mots considérés comme utiles, ils ont cherché à identifier quels mots étaient pensés le plus souvent. L'équipe du *Français Fondamental* a soumis des interviews à des d'écoliers, des ouvriers menuisiers et au public de la revue *Vie et Langage*, en obtenant une liste de mots concernant seize « centres d'intérêt » (domaines sémantiques). Gougenheim et al. (1964) constatent que « tandis que dans la liste générale des fréquences les verbes sont le plus souvent stables et les noms concrets instables, les enquêtes sur les associations d'idées autour d'un centre d'intérêt font ressortir la stabilité des noms concrets et l'instabilité des verbes » (p. 152) ; et que par conséquent, seule une combinaison de vocabulaire fréquent et disponible produit le vocabulaire fondamental. On peut continuer la citation :

> « On constate que des mots très utiles ne sont fréquents ni dans la langue écrite ni dans la langue parlée et que, de plus, cette faible fréquence est instable. Ce sont en général des mots concrets. Les mots concrets, en effet, ne sont prononcés ou écrits que dans des circonstances particulières, tandis que les mots de caractère général, notamment les mots grammaticaux et les verbes, ont une fréquence beaucoup plus forte et beaucoup plus stable. Contrairement à ce que l'on croit à première vue, on ne peut obtenir par la fréquence des mots tels que *veston*, *autobus*, *timbre*, *épicier* » (*Le Français Elémentaire*, 1954, p. 9).

Richards et Savard (1970, p. 37) affirment que la disponibilité d'un mot mesure le vocabulaire d'un certain domaine. Les auteurs expliquent que l'apprentissage du vocabulaire est lié aux domaines sociaux et culturels que le locuteur rencontre. Ces domaines sont institutionnalisés et se caractérisent par des comportements linguistiques spécifiques. Ainsi, on peut mesurer la familiarité d'un mot par

l'importance que lui accorde le locuteur. L'évaluation subjective ne correspondra pas au degré d'importance qui résulte d'une liste de fréquence.

### *2.2.4 La dispersion ou homogénéité de distribution*

Les mots qui peuvent être adoptés dans des contextes divers (c'est-à-dire qui sont distribués de façon homogène dans tout le corpus de référence plutôt que dans quelques sections du corpus) sont sans doute plus utiles que ceux qui ont un usage restreint à des contextes particuliers. Considérons le verbe *faire* : il se retrouve pratiquement dans n'importe quel type de texte, tandis que le verbe *grimper* se trouve dans un nombre restreint de textes. La dispersion est un critère statistique très important qui sert à relativiser la fréquence et à exclure de l'inclusion dans le vocabulaire fondamental les mots qui sont très fréquents mais qui n'apparaissent que dans des contextes limités. Une des façons les plus simples pour mesurer la dispersion est notamment la répartition, c'est-à-dire le calcul du nombre de textes différents dans lesquels le mot apparaît. D'après Richards et Savard (1970) :

> « [...] un mot qui ne figure que dans un petit nombre de textes, mais qui s'y trouve maintes fois répété, est évidemment lié au contenu de ces textes. Sa fréquence résulte de circonstances particulières. Il ne devrait pas trouver place dans une bonne liste de fréquence, dont le rôle est de mettre en évidence l'effet des causes générales » (p. 9-10).

## 2.3 Conclusions : des mots isolés aux mots en contexte

Dans notre étude nous nous intéressons aux collocations fondamentales d'après l'observation de l'importance du vocabulaire fondamental dans l'apprentissage linguistique. Etant donné que les mots ne sont pas des unités isolées, nous étudions la dimension syntagmatique, qui révèle la présence d'associations privilégiées entre les mots. L'étude des mots en contexte constitue la base de toute approche de l'étude du vocabulaire, qu'elle soit simplement descriptive, pédagogique ou lexicographique. Ici, nous « oublions » les longues listes de mots du passé et nous nous lançons dans le domaine fascinant des unités phraséologiques. Dans le chapitre

suivant, nous allons présenter un état de l'art des études sur la phraséologie et en particulier sur les collocations.

# CHAPITRE 3
# Excursus sur la collocation : théories et approches

Dans ce chapitre, nous nous proposons de situer notre approche dans le cadre plus général des études phraséologiques, aujourd'hui en plein essor, et nous nous centrons plus particulièrement sur le phénomène de la collocation.

Dans la première section, nous commencerons par expliquer la notion de figement (paragraphe 3.1.1) ; ensuite, nous citerons les deux approches principales qui caractérisent l'étude de la phraséologie (paragraphe 3.1.2) ; enfin, nous rappellerons les caractéristiques des deux approches principales (3.1.3).

Dans la seconde section, nous nous centrerons uniquement sur la collocation : en premier lieu, nous ferons un excursus historique des études les plus significatives sur les collocations et nous présenterons les définitions du concept qui ont été proposées (3.2.1) ; en second lieu, nous réfléchirons sur la validité des critères traditionnellement caractérisant les deux approches principales (3.2.2) ; enfin, nous proposerons une solution hybride conciliant ces deux approches (3.2.3).

## 3.1 Le vaste domaine de la phraséologie

La phraséologie reste un domaine de recherche fascinant et répandu en linguistique théorique et appliquée, puisque la connaissance des formules conventionnelles est indispensable pour que l'expressivité et l'aisance de la production se manifeste en langue première comme en langue étrangère.

Le terme « phraséologie » réfère aux séquences polylexicales qui se caractérisent par un certain degré de figement syntaxique ou sémantique et qui sont mémorisées comme des unités à part entière. Par figement, on entend ce qui n'est pas compositionnel (dérivable de la somme des sens des composants d'une

association de mots) et qui manque souvent d'autonomie syntaxique et sémantique, à la différence des syntagmes libres.

Gross (1996) affirme que le figement est une propriété des langues naturelles qui a été traitée dans toutes les grammaires mais dont l'importance a été longtemps ignorée. Les nombreuses études sur la phraséologie ont mis en évidence deux aspects principaux de la langue : en premier lieu, le caractère omniprésent des blocs lexicaux dans la langue en tant qu'unités lexico-syntaxiques (Sinclair, 1991) ; en second lieu, la maîtrise de la phraséologie comme clé pour aboutir à une compétence native en langue étrangère (Nattinger & DeCarrico, 1992).

### *3.1.1 La notion de figement*

Charles Bally (1951), le premier à faire de la phraséologie une véritable discipline au sein de la lexicologie, inclut sous le terme de phraséologie toute association de mots se caractérisant par une autonomie sémantique en tant qu'unité, et dont les composants perdent leur signification partiellement ou totalement :

« Si dans un groupe de mots, chaque unité graphique perd une partie de sa signification individuelle ou n'en conserve aucune, si l'association de ces éléments se présente seule avec un sens bien net, on peut dire qu'il s'agit d'une locution composée. [...] c'est l'ensemble de ces faits que nous comprenons sous le terme général de phraséologie » (p. 65-66).

Presque un siècle plus tard, Gaston Gross (1996) énumère les critères qui caractérisent l'expression figée, ainsi décrite :

« Une séquence est figée du point de vue syntaxique quand elle refuse toutes les possibilités combinatoires ou transformationnelles qui caractérisent habituellement une suite de ce type. Elle est figée sémantiquement quand le sens est opaque ou non compositionnel, c'est-à-dire quand il ne peut pas être déduit du sens des éléments composants. Le figement peut être partiel si la contrainte qui pèse sur une séquence donnée n'est pas absolue, s'il existe des degrés de liberté » (p. 154).

Ci-dessous, nous présentons les conditions nécessaires pour définir la notion de figement selon l'auteur.

1) La polylexicalité

Cette propriété est attribuée aux séquences composées de plusieurs mots qui ont par ailleurs leur propre autonomie. La polylexicalité n'inclut pas les cas plus généraux de contrainte syntaxique qui existent dans toutes les langues, par exemple le cas des arguments qui appartiennent à leurs propres prédicats.

2) L'opacité sémantique

Cette propriété est attribuée aux séquences sémantiquement figées ou contraintes lexicalement, c'est-à-dire les séquences dont le sens n'est pas transparent car il n'est pas dérivable de la somme des sens de ses composants. La phrase *Les carottes sont cuites*, par exemple, quand elle acquiert un sens imagé signifiant qu'une situation est désespérée, n'est pas interprétable à partir de la somme des sens de ses composants. L'unité *clé anglaise*, elle aussi, n'a pas de sens compositionnel et indique un type de clé.

3) Le blocage des propriétés transformationnelles

Cette propriété est attribuée aux séquences syntaxiquement figées, c'est-à-dire aux séquences qui ont une ou plusieurs restrictions syntaxiques : par exemple, pour le verbe, une restriction sur la passivation ou sur la pronominalisation ; pour le substantif, une restriction sur l'adjonction des adverbes intensifs. Considérons le verbe *regarder* : il ne peut pas être utilisé au passif, comme dans l'exemple **Nous sommes tous regardés par cette affaire*. L'auteur souligne la corrélation entre opacité sémantique et blocage syntaxique, et la nécessité d'une analyse combinée des deux aspects.

4) La non-actualisation des éléments

Cette propriété définit le concept de « locution », une séquence polylexicale dont les composants ne peuvent pas être actualisés, comme dans la suite figée idiomatique *prendre une veste,* qui signifie « perdre aux élections », où le nom *veste* ne se réfère pas à un vêtement et où *prendre* ne garde pas non plus son sens habituel. Cette suite figée est soumise à une contrainte sur le déterminant *une,* car le nom *veste* ne peut pas être actualisé par n'importe quel déterminant mais exclusivement par *une* dans ce sens particulier. Par contre, dans la séquence non figée *Paul a pris une*

*veste*, le nom *veste* peut être actualisé par *sa, cette, ta,* etc. sans que le sens de la séquence ne change.

5) La portée du figement

Il s'agit d'une propriété des suites polylexicales qui concerne l'extension du figement sur un ou plusieurs de ses composants. Dans des séquences comme les proverbes, tous les composants de la suite sont figés. Dans d'autres séquences comme la phrase *Vous lui avez tiré les vers du nez* (1996, p. 15), « une partie seulement de l'ensemble peut faire l'objet d'un figement » (1996, p. 16) car le sujet *vous* et le complément *lui* ne sont pas figés et peuvent être substitués.

6) Le degré de figement

Il s'agit d'une propriété des suites polylexicales qui concerne l'ampleur du figement. En effet, il est plutôt rare qu'une suite lexicale figure comme une entrée lexicale indépendante et qu'elle soit complètement bloquée sur le plan sémantique et syntaxique. Une suite qui est enregistrée comme une entrée à part dans les dictionnaires est, par exemple, *fait divers*, qui n'est pas transparente ni prédicative, et qui n'admet pas de nominalisation. La suite *fait historique*, par contre, n'est pas totalement figée. Les suites du type *fait historique* sont beaucoup plus fréquentes que les suites du type *fait divers,* car la présence de paradigmes est très commune dans la langue. Par exemple, dans la suite opaque *rater le coche* (qui signifie « laisser passer une occasion »), *rater* peut être remplacé par *louper* ou par *manquer*. Gross affirme à ce propos que « les suites totalement figées sont très minoritaires par rapport à celles qui ont des restrictions partielles » (1996, p. 22).

7) Le blocage des paradigmes synonymiques

Cette propriété est attribuée aux séquences dont un des composants ne peut pas être substitué par un synonyme, comme dans *faux sens,* où *faux* ne peut pas être remplacé par *mauvais* et ainsi produire **mauvais sens*. Le blocage du paradigme synonymique est un mécanisme lexical qui concerne toute catégorie grammaticale, et consiste dans le fait que la substitution d'un des composants de la suite figée dénature le sens de l'expression.

8) La non-insertion

Cette propriété est attribuée aux séquences dont la portée du figement ne peut pas être modifiée par l'insertion d'éléments nouveaux, ce qui s'applique surtout aux

nominales contraintes : il est correct de dire *un pré très vert,* mais nous ne pouvons pas dire **un col très blanc*, où *col blanc* signifie « bureaucrate ».

9) Le défigement

Le « défigement » marque la déstructuration de suites figées et permet la création de paradigmes « insolites » pour créer un effet de surprise. Il s'agit d'un mécanisme très utilisé dans la presse, car il opère sur la connaissance des paquets linguistiques partagée par une même communauté linguistique. Il peut concerner des proverbes (*aide-toi*), des chansons (*la ville en rose*), des formules[22] (*interruption volontaire de la carrière*), etc. Il est évident que l'effet du défigement prouve l'existence du figement.

10) L'étymologie

Il s'agit de la possibilité que la suite figée soit motivée par son origine historique. Par exemple, *pomme de discorde*, fait référence à un événement de la mythologie grecque. Il peut aussi arriver que certaines suites se bloquent dans un état de langue antérieur à cause de mécanismes historiques internes : par exemple, l'expression *chercher noise à quelqu'un* ou *chercher des noises*[23], (qui signifie « chercher la querelle délibérément avec quelqu'un ») a gardé la même forme qu'elle avait dans le français ancien, quand l'article n'était pas utilisé.

Bolly (2005) recense les principales études menées autour du phénomène phraséologique et liste les quatre critères attribués traditionnellement aux séquences figées.

1) La polylexicalité

Il s'agit d'un critère orthographique. Les séquences figées sont composées de lexèmes séparés par un blanc (González Rey, 2002, p. 53).

2) La fixité

Il s'agit d'un critère syntaxique. Les séquences figées sont soumises à des contraintes sur l'ordre de leurs composants et sur leur possibilité transformationnelle.

[22] Ici, le terme « formule » n'est pas utilisé en sens technique.

[23] Selon le *Dictionnaire historique de la langue française Le Robert* (2000), le terme *noise* viendrait du latin *nausea* (« nausée »). Le mot, qui signifie « querelle », a disparu du français courant.

3) L'opacité

Il s'agit d'un critère sémantique. Les séquences figées ont un sens non compositionnel.

4) La restriction paradigmatique ou lexicale

Il s'agit d'un critère lexical. Les séquences figées sont soumises à des contraintes sur leur possibilité de substitution synonymique.

Les trois derniers critères permettent de placer les unités phraséologiques tout au long d'un continuum qui va des séquences libres aux séquences plus figées.

### *3.1.2 Les deux grands courants de recherche sur la phraséologie : approche statistique et approche phraséologique*

Avant que le débat autour du concept de collocation ne commence, la recherche linguistique essayait de classifier et de définir le concept plus général d'unité phraséologique. Comme l'explique Cowie (1998), de nombreuses théories ont abordé la question de la classification des unités phraséologiques, assez problématique à cause de leur hétérogénéité. L'intérêt pour ce phénomène s'est considérablement accru grâce au développement des théories sur l'acquisition lexicale en langue maternelle, seconde ou étrangère. Le débat qui s'ensuivit a porté sur les limites de la vision atomique de la théorie générativiste, qui considérait le langage comme un système composé d'unités minimales et géré par des règles générales d'interprétation sémantique.

Cowie cite trois théories principales dans le domaine des études sur la phraséologie : la tradition classique russe, la tradition anthropologique et la tradition anglo-saxonne. La tradition anthropologique (qui souligne l'importance de l'élément culturel dans l'unité phraséologique, et dont Veronika Teliya est une des représentantes majeures) étant un peu plus à l'écart, les deux grands courants qui occupent la scène de la recherche sur les unités phraséologiques restent la tradition anglo-saxonne et la tradition classique russe. La tradition classique russe est le fondement de l'« approche phraséologique » ou « approche continentale » (selon la terminologie de Williams, 2003) ; la tradition anglo-saxonne a donné naissance à l'« approche statistique » ou « approche contextualiste », dont le travail de pionnier

est celui de Firth. Dans notre recherche, nous nommerons les deux approches « approche phraséologique » et « approche statistique ».

D'un côté, la tradition classique russe, qui se développe entre les années 40 et 60, et construit l'idée du continuum phraséologique qui va des unités plus figées aux unités plus transparentes. Cette tradition, inaugurée par des linguistes comme Klappenbach, Weinreich, Arnold et Lipka (respectivement : 1968, 1969, 1973, 1974 cités par Cowie, 1998, p. 4), s'est concentrée sur la distinction entre les unités sémantiques de type « mot » et les unités pragmatiques de type « phrase ». Les premières subissent des modifications syntaxiques, par exemple *in the nick of time* ou *break one's journey* ; tandis que les secondes sont représentées par les proverbes, les formules, etc., par exemple *There is no fool like an old fool* ou *You don't say !* (p. 4). Le premier linguiste à opérer ce type de distinction fut Chernuisheva (1964, cité par Cowie, 1998, p. 4), qui appelait les unités pragmatiques « expressions phraséologiques ». Fortement influencé par la tradition classique russe, Cowie fait la distinction entre les unités sémantiques qu'il appelle « composites » et les unités pragmatiques qu'il appelle « functional expressions », c'est à dire une distinction entre expressions plus ou moins figées et vraies formules. Cette perspective est suivie par d'autres auteurs comme Howarth (1996), qui distingue « composite units » et « fonctional expressions », Gläser (1988), qui différencie les « nominations » et les « propositions », Mel'čuk (1998), qui oppose les « phrasèmes sémantiques » aux « pragmatèmes ». Le fait de considérer les unités pragmatiques (proverbes, citations, clichés, etc.) comme des unités au niveau de la phrase permet d'opérer une distinction entre les unités sémantiques polylexicales à valeur syntaxique et les unités à but pragmatique. De la tradition russe se développera l'approche syntaxique qui caractérise l'étude des collocations, que nous allons analyser plus en détail dans le paragraphe 3.2.1.

D'un autre côté, la tradition anglo-saxonne trouve son origine dans le contextualisme anglais, illustré notamment par le linguiste J. R. Firth (1957), dont le travail fut continué par son disciple M. Halliday (1985) et par le père de la linguistique de corpus J. Sinclair (1991), fondateur du projet *COBUILD* (1987). Cette approche est bien plus large que l'autre et inclut, dans le vaste groupe des unités phraséologiques, toute association récurrente de mots ayant une fréquence

d'occurrence plus élevée que le simple hasard. Selon Jones et Sinclair (1974), la collocation est significative si elle est une « [...] regular collocation between items, such that they occur more often than their respective frequencies and the length of the text in which they occur would predict » (p. 19)[24]. Cela témoigne de l'importance fondamentale de l'usage des blocs lexicaux figés dans la langue, toujours à disposition du locuteur. De la tradition contextualiste se développera l'approche statistique qui caractérise l'étude des collocations (voir paragraphe 3.2.1).

### *3.1.3 Collocations, associations libres et expressions figées : quelques classements*

Nous conclurons cette première section en présentant quelques exemples de classification des différents types d'unités phraséologiques. Les unités regroupées sous l'étiquette générale de phraséologie sont multiples et complexes, et leur étude fait aujourd'hui l'objet de nombreuses recherches qui ont abouti à des modèles de classification. Cependant, il reste assez problématique de trouver des critères homogènes qui permettent de distinguer les collocations des associations libres et des expressions plus figées.

Charles Bally (1951), disciple de Saussure à qui on attribue le concept de collocation, dans son *Traité de stylistique française* de 1909 différait entre trois types de séquences phraséologiques :

1) « groupements libres » ou « groupements passagers »
2) « groupements usuels » ou « séries phraséologiques »
3) « unités phraséologiques ».

Les groupements libres sont issus d'un libre choix grammatical et sémantique. Par contre, les groupements usuels et les unités phraséologiques sont des unités lexicales figées, les premières étant compositionnelles (l'unité phraséologique correspondant à la somme des sens de ses composants), les secondes étant non compositionnelles (Gonzalez-Rey, 2002).

[24] « […] collocation régulière entre des items lexicaux qui se présentent plus fréquemment que leurs fréquences respectives et que la longueur du texte dans lequel ils apparaissent le prédirait ».

Sur les groupements usuels, Bally remarquait :

« Il y a série ou groupement usuel lorsque les éléments du groupe conservent leur autonomie, tout en laissant voir une affinité évidente qui les rapproche, de sorte que l'ensemble présente des contours arrêtés et donne l'impression du "déjà vu"» (1951, p. 70).

Selon Grossmann et Tutin (2002), la formalisation nécessaire au traitement automatique et lexicographique des collocations impose une nette distinction entre les collocations et les notions voisines, notamment les expressions libres et les expressions figées. Si au niveau syntaxique, il est possible de distinguer les collocations des notions voisines sur la base de leur degré de figement, au niveau sémantique, il est par contre erroné d'envisager la différence en terme d'échelle. Sur le plan syntaxique, les auteurs affirment qui il n'y a pas d'opposition binaire entre expressions libres et expressions figées. En effet, des critères syntaxiques comme la non-insertion de modifieurs ou la contigüité des éléments ne peuvent être appliqués qu'à un sous-ensemble des expressions figées, et ne sont pas des critères définitoires (comme pour les collocations). Sur le plan sémantique, les auteurs envisagent la classification melčukienne comme une solution meilleure que le continuum de figement, car elle permet de distinguer des cas de figure plus fins dans le cadre d'une typologie.

En effet, la classification melčukienne permet de distinguer les « phrasèmes complets » (non compositionnels, par exemple les phrases idiomatiques), des « semi-phrasèmes » (compositionnels, notamment les collocations, où un des éléments conserve son sens habituel), des « quasi-phrasèmes » (où les composants perdent leur sens autonome en faveur du sens de l'unité, par exemple *donner le sein* qui veut dire « nourrir de son lait »). Comme l'expliquent les auteurs :

« [...] s'il existe indéniablement un continuum dans le degré de figement sémantique des collocations, il apparaît néanmoins souhaitable de proposer une typologie plus fine à l'intérieur de cette classe permettant des traitements linguistiques adaptés à chaque type » (Grossmann, Tutin, 2002, p. 8).

Grossmann et Tutin (2003) présentent les unités phraséologiques de la langue sous deux catégories principales :

1) « Séquences lexicales phrastiques »

Elles comprennent des unités complètement figées qui fonctionnent comme phrases indépendantes, des unités textuelles : les proverbes ou parémies, les lieux communs, les maximes, les slogans et enfin les formules usuelles liées à des contextes spécifiques (les « pragmatèmes » mel'čukiens).

2) « Séquences lexicales syntagmatiques »

Elles se présentent sous la forme de syntagmes plus ou moins figés (nominal, verbal, adjectival ou adverbial). Par rapport au degré de figement, Grossmann et Tutin reprennent la terminologie de Bally et font la différence entre « unités phraséologiques » et « groupements usuels » :

a) « Unités phraséologiques »

Elles sont des associations figées et non-compositionnelles dont le sens est complètement opaque (par exemple *cordon bleu,* qui veut dire « bon cuisinier »), même si parfois déductible de la métaphore ou de la métonymie qui est à la base de l'association (par exemple, *manger les pissenlits par la racine*, qui veut dire « être mort et enterré »). Cette sous-catégorie est très riche et représente l'extrême du plus figé au long du continuum sémantique généralement proposé par beaucoup de modèles descriptifs des collocations.

b) « Groupements usuels »

Il s'agit des collocations (voir paragraphe 3.2.1), expressions semi-figées binaires dont le collocatif est sélectionné pour exprimer un sens donné en association avec sa base, qui elle possède un sens autonome. Des exemples de collocation sont *chaleur suffocante* et *prendre une décision* (construction à verbe support).

En conclusion, les auteurs expliquent que les traits sémantiques caractéristiques des expressions figées, qui les distinguent nettement des collocations, sont deux : la non-compositionalité des expressions figées et le fait de renvoyer à un référent unique. Du premier trait dérive le statut d'unité linguistique qui est accordé aux expressions figées par la lexicologie explicative et combinatoire de Melčuk (1998). Toutefois, ils précisent que parfois des procès métaphoriques ou métonymiques rendent l'expression interprétable. Le second trait concerne en particulier les

expressions figées de type nominal. Il s'agit d'un critère un peu problématique lorsqu'on a affaire à des collocatifs qualifiants, comme dans la collocation *célibataire endurci* qui n'indique pas un référent unique, mais une qualité de *célibataire*. Par contre, les collocatifs typants ne sont pas problématiques, par exemple *café noir* indique une sorte de café, c'est un hyponyme de café.

Nattinger et DeCarrico (1992) s'occupent également de la classification des unités phraséologiques, parmi lesquelles ils soulignent l'importance des unités phraséologiques pragmatiques, qu'ils nomment sous l'étiquette de « phrases lexicales » (en anglais « lexical phrases »). Comme ils l'expliquent, il s'agit d'unités phraséologiques souvent ignorées dans la didactique des langues, du type *how do you do ?, a ____ ago* and *the ___er the ___er* en anglais (utilisées pour dire bonjour, exprimer des relations temporelles, faire des comparaisons respectivement). Notamment, elles sont définies comme :

« [...] lexico-grammatical units that occupy a position somewhere between the traditional poles of lexicon and syntax : they are similar to lexicon in being treated as units, yet most of them consist of more than one word, and many of them can, at the same time, be derived from the regular rules of syntax, just like other sentences. Their use is governed by pragmatic competence, which also selects and assigns particular functions to lexical phrase units »[25] (Nattinger, DeCarrico, 1992, p. 36).

Comme les auteurs l'expliquent, les « phrases lexicales » diffèrent des phrases qui n'ont aucune fonction pragmatique : ces dernières sont des unités très figées comme les phrases idiomatiques ou les clichés, par exemple en anglais *it's raining cats and dogs* (qui veut dire « il pleut fort ») and *kick the bucket* (qui veut dire « mourir »).

Les auteurs font la distinction entre « phrases lexicales », « collocations » et « séquences syntaxiques » (en anglais « syntactic strings »):

[25] « [...] unités lexico-grammaticales qui occupent une position à mi-chemin des poles traditionnels du lexique et de la syntaxe : elles sont similaires au lexique car elles sont traitées comme des unités, pourtant la plupart d'entre elles se composent de plus d'un mot, et en meme temps, un grand nombre d'entre elles peuvent être dérivées des règles habituelles de la syntaxe, comme les autres phrases. Leurs usage est régi par la compétence pragmatique qui, en outre, sélectionne et assigne des fonctions partuclières aux unités lexicales phrasales » (notre traduction).

1) « Phrases lexicales » (canoniques ou non canoniques)

Les phrases lexicales peuvent être de deux types :

a) Elles se composent d'éléments lexicaux spécifiques non productifs et non substituables au niveau syntagmatique et pragmatique. Par exemple, *at any rate* (canonique) et *by and large* (non canonique, étant donné qu'une préposition se combine avec un adjectif) en anglais (p. 36).

b) Elles sont des structures productives généralisables, qui se composent de constructions syntaxiques ou sémantiques spécifiques « remplies » par des éléments lexicaux. Par exemple, en anglais, la structure « a + nom de temps + ago » dans *a month ago, a while ago, a year ago* (canonique) pour indiquer le temps ; la structure « adverbe indiquant la direction + with + syntagme nominal » dans *down with the king* ou *away with all bureaucrats* (non canonique), pour exprimer la désapprobation (p. 36).

2) « Collocations »

Il s'agit de séquences d'éléments lexicaux qui ne remplissent pas de fonction pragmatique. Par exemple, en anglais, *rancid butter,* où la base et le collocatif s'associent plus souvent que le simple hasard (ce qu'on peut mesurer à l'aide de mesures associatives comme l'Information mutuelle (Nattinger, De Carrico, 1992, p. 21-22).

3) « Séquences syntaxiques »

Il s'agit d'associations libres, symboles catégoriels (catégories grammaticales) générés par la compétence syntaxique du locuteur. Elles sont canoniques, c'est-à-dire elles correspondent à des suites syntaxiques standard, par exemple la structure « sujet + auxiliaire + participe passé ».

Selon la théorie des profils textuels de V. Lo Cascio (2000), chaque locuteur, pour exprimer un message, utilise les nombreux profils textuels mis à disposition par sa langue spécifique. Tout locuteur y a accès d'une façon quasi-automatique parce que ces profils sont enregistrés comme des unités dans le lexique mental. Le locuteur idéal devrait connaître tous les profils possibles, mais le locuteur réel n'en connait que certains, les profils qu'il utilise de préférence. La difficulté dans le choix du profil adéquat est plus grande lorsqu'on utilise une langue étrangère. En effet, le choix combinatoire est dicté par une dimension socioculturelle arbitraire qui ne

permet pas de prédire l'association exacte dans la langue cible. Donnant crédit à l'idée du continuum combinatoire, Lo Cascio distingue entre les types suivants d'unités lexicales : polyrhématiques, proverbes, structures idiomatiques, associations libres et collocations. Parmi ces unités lexicales, les structures idiomatiques, les collocations et les associations libres sont les plus difficiles à définir[26] :

1) « Structures idiomatiques »

Les structures idiomatiques sont des blocs lexicaux non compositionnels qui résultent d'un procès de translation ou métaphorisation partiellement transparent ou totalement obscur. Généralement, elles subissent des restrictions syntaxiques : par exemple, *poser un lapin* en français (qui veut dire « ne pas aller à un rendez-vous sans prévenir la personne qui nous attend ») et *pull someone's leg* en anglais (qui veut dire « se moquer de quelqu'un) n'admettent pas de complément objet au pluriel.

2) « Associations libres »

Les associations libres sont des associations de mots qui se basent sur des règles de cohérence encyclopédique, grammaticale, syntaxique et sémantique. En dehors de ces règles, il n'existe pas de restriction particulière, ce qui donne la possibilité au locuteur d'associer les mots selon ses exigences de créativité.

3) « Collocations »

Les collocations, à la différence des associations libres, subissent des restrictions qui ne sont pas seulement de nature encyclopédique ou spécifiquement linguistique, mais aussi de nature sociale et conventionnelle. En effet, elles réfléchissent la façon dont chaque langue exprime, de manière arbitraire, un concept particulier. Nous renvoyons à la section 3.2 pour une description plus détaillée des collocations selon Lo Cascio.

[26] Les proverbes peuvent être assimilés aux maximes ; les polyrhématiques, très souvent confondues avec les noms composés, sont des blocs lexicaux considérés comme une seule unité linguistique et traités dans les dictionnaires comme des entrées lexicales à part: par exemple *feu rouge* ou *coup de foudre* en français *et light bulb* en anglais. Elles subissent des restrictions syntaxiques étant donné que, par exemple, il est impossible de produire **coup fort de foudre* et **light little bulb*. Elles subissent aussi des restrictions sémantiques car elles résultent d'un procès de lexicalisation et ne sont pas compositionnelles (bien qu'elles soient plus ou moins transparentes selon le procès qui a amené les composants à s'unir).

## 3.2 La collocation : une unité phraséologique à statut spécial

Dans cette section, nous centrerons notre attention sur le phénomène de la collocation. En premier lieu, nous présenterons un historique des théories les plus influentes s'inscrivant dans l'approche statistique ou phraséologique (3.1.1) ; ensuite, nous soulignerons les différences principales entre les deux approches dans le traitement des collocations (3.1.2) ; enfin, nous réfléchirons sur la validité des critères adoptés par les deux approches et nous proposerons une solution d'équilibre entre statistique et description formelle (3.1.3).

Le fondateur du débat autour des collocations fut J. Firth (1957), mais comme nous l'avons expliqué dans la section précédente, le concept de « collocation » est déjà apparu en 1909 avec Bally (1951). Le travail de Firth, qui s'inscrit dans le courant contextualiste, fut continué par ses successeurs, en premier lieu par Halliday (1985) et Sinclair (1991). Hausmann (1989) avancera une critique au modèle contextualiste et proposera une formalisation des collocations basée sur une classe restreinte de structures syntaxiques. Le débat autour de ce phénomène linguistique trouvera des applications lexicographiques grâce aux auteurs du *BBI Dictionary of English* (Benson et al., 1986a) qui essaient, comme Hausmann, d'intégrer les collocations dans les dictionnaires. La théorie des fonctions lexicales de Mel'čuk (Mel'čuk et al., 1984) présentera, pour la première fois, un modèle structuré pour la description sémantique des collocations. Ce modèle sera suivi par d'autres typologies basées sur une description sémantique, par exemple celle de Cowie (1998). Enfin, les études modernes visent en particulier à la distinction entre collocations, associations figées et associations libres, et s'intéressent aux applications en TAL. Notre excursus historique se terminera avec la description de deux contributions à l'étude des collocations très significatives : l'approche plutôt phraséologique de Grossmann et Tutin (2002) et l'approche plutôt statistique de Lo Cascio (1998). Les théories que nous allons décrire ne sont que les plus significatives dans un panorama de recherche qui est aujourd'hui devenu très vaste.

*3.2.1 Historique des théories principales sur les collocations*

Si on fait généralement remonter l'introduction formelle du terme « collocation » dans la discipline de la linguistique au linguiste John Rupert Firth, et à Charles Bally la naissance du concept de « collocation » au début du XXème siècle, c'est à H. E. Palmer (1933, cité dans Cowie, 1999) qu'on doit la première utilisation du terme dans les années 30.

Harold Edward Palmer fut l'auteur du dictionnaire combinatoire *A grammar of English Words* (1938, cité dans Cowie, 1999) et professeur d'anglais au Japon. Il menait des recherches sur l'enseignement de l'anglais langue étrangère et s'exprimait comme suit :

« […] it is not so much the words of English nor the grammar of English that makes English difficult, but that vague and undefined obstacle to progress in the learning of English consists for the most part in the existence of so many odd comings-together-of-words » (Palmer, 1933, p. 13, cité dans Cowie, 1999, p. 53).[27]

Fondateur de la lexicographie de l'anglais langue étrangère ensemble à A. S. Hornby, Palmer appelait les collocations « comings-together-of-words » (dont la traduction littérale serait « mots qui vont ensemble »). Son intérêt pour l'étude des collocations naît d'un projet de recherche (auquel Hornby participa plus tard) lancé dans les années 20, qui visait à recueillir et classifier les expressions à mots multiples. Le rapport final du projet prit le nom de *Second Interim Report on English Collocations* (Palmer, 1933). Bien que Palmer utilise le terme collocation pour une grande diversité d'unités phraséologiques et qu'il ne prenne pas en compte les différents degrés de figement des unités phraséologiques, son travail peut être considéré comme innovateur pour de nombreuses raisons. En premier lieu, il fait la différence entre collocations et associations libres, en affirmant que ces dernières dérivent de la libre combinatoire lexicale, c'est-à-dire qu'elles résultent de l'application des règles de grammaire les plus communes (Palmer, 1933, cité dans

[27] « [...] ce ne sont pas tant les mots ou la grammaire qui rendent l'anglais difficile. L'obstacle indéfini et vague au progrès de l'apprentissage de l'anglais provient en grande partie de l'existence de nombreux et étranges groupements de mots » (notre traduction).

Cowie, 1999, p. 211). Cowie souligne en outre l'importance de l'*Interim Report* que nous venons de citer, où une classification détaillée des catégories syntaxiques des associations étudiées est présentée. Il s'agit surtout de structures phrasales ou prédicatives du type « préposition - adjectif possessif - nom », dont ils fournissent les exemples suivants : *at one's ease, at one's feet, at one's leisure,* etc. Enfin, Palmer eut aussi le mérite de souligner l'importance des collocations pour l'apprentissage de la langue.

La théorie fonctionnaliste, se développe auprès de l'Université de Londres grâce aux travaux de J. R. Firth (1957), professeur de linguistique. Les bases méthodologiques de l'analyse statistique et de l'étude de corpus se développent au sein du contextualisme anglais, qui attribue un rôle fondamental au contexte de situation dans l'interprétation du langage. A partir de sa célèbre phrase « You shall know a word by the company it keeps », Firth a le mérite de s'être occupé non seulement des relations lexicales paradigmatiques, mais aussi des relations syntagmatiques. Il insiste sur l'importance d'évaluer le sens en contexte, d'où son « meaning by collocation » : Selon l'auteur (Firth, 1957, p. 196), la collocation est une abstraction au niveau syntagmatique et ne concerne pas directement l'approche conceptuelle au sens. Il explique que le sens de *night* est dans son association avec *dark* et vice-versa.

A partir des années 60, le disciple de Firth, M. A. K. Halliday (1985), et le père de la linguistique de corpus, John Sinclair (1991), font faire un bond en avant à l'analyse des collocations. Halliday accorde, comme Firth, une grande importance au contexte. Selon lui, la collocation peut être purement sémantique et dériver de relations telles que la synonymie, l'hypomimie, etc. ou lexicale et dériver des liens collocationnels. Etant donné que la collocation joue un rôle de cohésion textuelle, elle correspond aussi à des associations lexicales paradigmatiques du type « médecin… hôpital ». Il développe la « Grammaire Systémique Fonctionnelle », un modèle de corrélation entre syntaxe et sémantique/pragmatique (tandis que la grammaire générative de Chomsky et le structuralisme se concentraient exclusivement sur la syntaxe) et considère la langue comme une action sociale, d'où son affirmation :

« Language is a part of the social system » (Halliday, 1978, p. 39)[28].

Si le sens dépend des associations sur l'axe syntagmatique, alors la sémantique est liée à la syntaxe et vice-versa. Le fonctionnalisme et la linguistique cognitive visent à la description de la structure de la langue mais sans la séparer de la sémantique et représentent donc une réaction aux deux grandes théories structuraliste et générative qui ne s'intéressent pas à la question du sens. La « Linguistique Systémique Fonctionnelle » de Halliday (1996) élargit l'analyse de Firth à tous les niveaux linguistiques de l'énonciation et ne prévoit pas un lexique indépendant du système grammatical : dans son article *Lexis as a linguistic level*, l'auteur établit une interdépendance entre grammaire et lexique, voit la langue comme un moyen pour véhiculer un sens social, et souligne la fonction communicative du langage, à la différence des approches structuralistes qui négligent cet aspect.

Dans le même courant s'inscrit John Sinclair (1991), qui a été Professeur d'anglais à l'Université de Birmingham et responsable du grand projet lexicographique *COBUILD* (1987), acronyme de *Collins Birmingham University International Language Database*. Né de la collaboration entre le groupe *English Language Research* de l'Université de Birmingham et l'éditeur *Collins Publisher*, le *COBUILD* représente le premier dictionnaire basé sur une base de données construite sur des exemples annotés. Avec le *COBUILD*, l'étude des collocations devient plus précise et un rôle très important est accordé à la fréquence. Comme l'explique Sinclair (1991, p. 5-6), le langage doit se fonder sur des éléments de preuve apportés par les textes. L'analyse des corpus permet de mettre de côté les faits linguistiques mineurs et peu fréquents, et de retenir à la surface les usages très communs de la langue. En effet, ces usages sont parfois peu analysés, tant ils sont routiniers.

Ce qui rend un texte « naturel » réside à la fois dans l'existence d'une bonne structure textuelle et dans son authenticité. C'est pour cette raison que Sinclair est réticent devant l'utilisation d'exemples inventés dans les ouvrages lexicographiques pour prouver l'existence d'un fait linguistique. Les exemples inventés se réfèrent à un contexte qui n'existe pas dans l'usage langagier, d'où sa phrase « [...] one does not

[28] « Le langage est une partie du système social » (notre traduction).

study all of botany by making artificial flowers »[29] (p. 6). Sinclair ne voit aucune différence entre grammaire et lexique : à un niveau très abstrait on choisit un sens, et ce sens se réalise dans une structure lexico-grammaticale qui dépend de l'usage. La vision selon laquelle un même sens est représenté de façon identique par toutes ses formes a été contestée par Sinclair, Jones et Daley (2004), selon qui l'étude de corpus montre qu'une ou quelques formes grammaticales sont toujours prédominantes par rapport aux autres. Les auteurs expliquent que tous les mots se trouvent dans une structure collocationnelle, même les mots grammaticaux. Cependant, les associations qui incluent des mots grammaticaux (en général des mots très fréquents) subissent seulement des restrictions positionnelles, et sont soumises à des blocages grammaticaux (cf. *in the house* et **house the in*) ; tandis que les collocations qui n'incluent pas de mots grammaticaux subissent un blocage d'ordre lexical (cf. *hard work* vs **work hard*). Les mots grammaticaux ne produisent pas de collocations significatives, et sont générés par des mécanismes d'ordre exclusivement grammatical. Les vraies collocations sont, en conclusion, les collocations lexicales :

« Collocation in the purest sense … recognises only the lexical cooccurrence of words »[30] (Sinclair, 1991, p. 170)

Pour les collocations grammaticales, l'auteur suggère de parler de « colligation ». Pour expliquer l'existence dans la langue de ces deux mécanismes de production linguistique, Sinclair formule deux principes :

1) L' « open choice principle »

C'est le « principe du choix ouvert », basé sur l'association de mots d'une façon additionnelle, selon des règles syntaxiques et sémantiques compositionnelles.

2) L' « idiom principle »

C'est le « principe idiomatique », basé sur le repérage direct du lexique mental d'expressions toutes faites, de blocs linguistiques, d'unités textuelles. Dans ce second

[29] « [...] on n'étudie pas l'ensemble de la botanique en fabriquant des fleurs artificielles » (notre traduction).

[30] « La collocation au sens le plus pur du terme... ne reconnait que les cooccurrences lexicales » (notre traduction).

cas, l'arbitraire de certaines phrases préconstruites fait que le locuteur sélectionne l'unité textuelle même si les règles de cohérence syntaxique et sémantique ne sont pas respectées. En résumé, le locuteur suit le seul principe de l'usage : par exemple, en anglais, une éclipse est *total* ou *full*, mais non **absolute, *complete, *entire, *whole* (Sinclair et al., 2004, p. 29) : ces adjectifs, qui seraient sémantiquement employables, ne le sont pas sur le plan collocationnel. L'auteur explique que le second principe est plus répandu qu'on ne peut l'imaginer et que 80 % du discours écrit et oral est produit selon le « principe idiomatique » et non selon des règles d'ordre syntagmatique et grammatical (Sinclair, 2000, p. 197). En conclusion, chaque locuteur se sert régulièrement de phrases préconstruites qui constituent des « singles choices » et qui ne peuvent pas être segmentées. Ce mécanisme de fonctionnement linguistique suit le même principe des affaires, le principe de l'économicité, ou dépend des exigences réelles de chaque événement communicatif. Sinclair (1991, p. 110) explique que chaque locuteur a à sa disposition un nombre considérable de phrases semi-figées qui représentent des choix uniques, bien qu'elles apparaissent comme analysables en segments plus petits.

Une des réactions les plus significatives au modèle contextualiste vient de Franz Joseph Hausmann (1984), un des représentants les plus importants de l'approche phraséologique. Hausmann fait la différence entre « syntagmes figés », c'est-à-dire les locutions, non compositionnelles ; et « syntagmes non figés **»,** c'est-à-dire les syntagmes compositionnels. Ces derniers peuvent être de trois types (p. 398) :

1) « Ko-Kreation »

Ce sont les « co-créations », les associations libres.

2) « Konter-Kreationen »

Ce sont les « contre-créations », des associations libres marquées par le choix stylistique du locuteur, peu usuelles et typiques du style littéraire.

3) « Kollokation »

Ce sont les « collocations », unités préfabriquées et fréquentes dans l'usage.

Plus tard, l'auteur ajoutera une autre distinction, celle entre éléments lexicaux « autosémantiques », éléments autonomes au niveau sémantique car leur sens est indépendant de leur contexte syntagmatique ; et élément lexicaux « synsémantiques », éléments dépendants sémantiquement de leur contexte

syntagmatique, c'est-à-dire de leur base. Dans les associations libres, les deux constituants sont autosémantiques. Dans les locutions, les deux composants sont synsémantiques parce qu'ils se sélectionnent réciproquement. Dans les collocations, seulement le collocatif est synsémantique : en association, la base garde son sens, tandis que le collocatif se définit par rapport à la base.

« La base n'a pas besoin du collocatif pour être clairement définie. Il en va tout autrement pour le collocatif qui ne réalise pleinement son signifié qu'en association avec une base. La base complète la définition du collocatif, alors que le collocatif se contente d'ajouter une qualité à une base en elle-même suffisamment définie » (Hausmann, 1979, p. 191-192).

Selon Hausmann, l'analyse statistique des collocations faite par les fonctionnalistes regroupe trop d'associations et manque d'une caractérisation syntaxique ou sémantique qui puisse retenir les vraies collocations. A ce propos, Sinclair et al. (2004) avaient déjà reconnu les limites dérivant d'une analyse purement statistique :

« Statistics, however, only tells us that the cooccurrence of two (or more) items is probably not accidental, that there should be a reason for it » (p. xxii).[31]

En réaction au modèle statistique, Hausmann (1989, p. 1010) présente une liste des structures syntaxiques possibles dans lesquelles les collocations apparaissent et il affirme :

« On appellera collocation l'association caractéristique de deux mots dans une des structures suivantes :
a) substantif + adjectif (épithète) ;
b) substantif + verbe ;
c) verbe + substantif (objet) ;
d) verbe + adverbe ;
e) adjectif + adverbe ;

[31] « Cependant les statistiques nous disent seulement que la cooccurrence de deux (ou plusieurs) items n'est probablement pas accidentelle, et qu'il doit y avoir une raison à son existence » (notre traduction).

f) substantif + (prép.) + substantif »

Plus tard l'auteur ajoutera à cette liste une autre structure, « verbe + (prép.) + nom ».

Grossmann et Tutin (2003, p. 7) citent la classification esquissée par Hausmann et la considèrent comme incomplète en se référant au *Dictionnaire Explicatif et Combinatoire* (Mel'čuk et al., 1984), où d'autres types syntaxiques apparaissent. Par exemple, les bases nominales se combinent non seulement avec des épithètes adjectivales, mais aussi avec des syntagmes tels que « prép + substantif » dans *de joie*, « adj. + prép. + substantif » dans *ivre de colère*, « V + nom » dans *avoir quelqu'un en visite*.

En outre, dans son dictionnaire des collocations, Hausmann intègre l'analyse syntaxique des collocations par un système sémantique simple qui explique les différents sens de chaque base. Orliac (2004) cite l'exemple de *doute* :

1. [N+v] **NAITRE**, **EXISTER** : naître, surgir, m'envahit, plane, subsiste, persiste ; **DISPARAITRE** : s'évanouir, s'envoler. 2. [v+N] **AVOIR** : avoir, concevoir, éprouver, il me vient des doutes ; **FAIRE NAITRE** : inspirer ; **EXPRIMER** : émettre, formuler ; **FAIRE DISPARAITRE** : lever, écarter, éclaircir, dissiper, balayer. 3. [v+prép+N] (être) assailli de doutes, rongé, tourmenté par le doute ; être, laisser dans le doute ; mettre, révoquer en doute. 4. [(a)+N+(a)] : légers -, - affreux, subits, persistants, bien fondés. 5. [N+prép+N] : le supplice du doute.

Pour conclure, selon Hausmann, la collocation est une association contrainte et arbitraire orientée, ayant un collocatif dépendant d'une base. Selon Nesselhauf (2005), Hausmann a le mérite d'avoir été le premier à reconnaitre que les composants de la collocation ne possèdent pas tous le même statut, ce qui avait était négligé par les statisticiens :

« En effet, dans la collocation *célibataire endurci*, le signifié de la base (*célibataire*) est autonome. La base n'a pas besoin du collocatif (*endurci*) pour être clairement définie. Il en va tout autrement pour le collocatif qui ne réalise pleinement son signifié qu'en association

avec une base (*célibataire*, *pécheur*, *âme,* etc.) » (Hausmann 1979, p. 191, cité par Nesselhauf).

En vérité, comme l'explique Dubreuil (2008), une définition statistique de la collocation n'exclut pas la possibilité d'une classification plus formelle, comme par exemple la typologie sémantique du lexicographe A. P. Cowie (1981). L'auteur, qui avant tout sépare unités pragmatiques et non pragmatiques, fait la distinction parmi les unités non pragmatiques, entre :

1) « Open collocations » (« collocations ouvertes »)
2) « Restricted collocations » (« collocations restreintes »)
3) « Figurative idioms » (« expressions idiomatiques imagées »)
4) « Pure idioms » (« expressions idiomatiques pures »).

Les « open collocations » sont des associations dont les deux composants gardent leur sens littéral : par exemple, l'association du verbe *to run* avec *machine*, ou *business* ou *horse* ou *program* ; et l'association *broken window*, dont les composants ont leur propre sens autonome.

Les « restricted collocations », ressemblent à celles ouvertes, d'une part ; mais elles sont très similaires aux expressions idiomatiques, d'autre part. Considérons, par exemple, le sens spécifique de *jog* dans l'association *jog one's/sb's memory* : il ne se présente pas dans d'autres contextes. C'est pourquoi seules les collocations ouvertes ne sont pas inclues dans son *Oxford Dictionary of English Idioms* – ODEI - (Cowie, Mackin, 1983) :

«Typically also, in open collocations, each element is used in a common literal sense. [...] We have been careful to exclude them from the Dictionary » (ODEI, 1983, p. xiii).[32]

Les « expressions idiomatiques imagées », par exemple *catch fire* (qui veut dire « prendre feu »), ne sont pas complètement figées et leur sens peut être interprété.

Les « expressions idiomatiques pures », par exemple *blow the gaff,* n'admettent aucune modification syntaxique et interprétation sémantique.

[32] « En outre, d'habitude, dans les collocations ouvertes, chaque élément est utilisé dans un sens littéral habituel » (notre traduction).

En ce qui concerne la différence entre collocations et expressions idiomatiques, Nerima et al. (2006, p. 97) sont de la même opinion. Selon ces auteurs, ce qui distingue les expressions idiomatiques des collocations est la plus grande transparence et la plus grande flexibilité syntaxique des secondes, comme le montre l'expression *décerner un prix*, qui admet les transformations suivantes :

- modification adjectivale : *décerner un important prix*
- passivation : *le prix Nobel de la Paix 2005 a été décerné hier*
- relativisation : *le prix qui lui a été décerné l'année passée*
- clivage : *c'est le prix le plus important qui sera décerné demain soir*
- interrogation : *quels prix ont été décernés lors de ce festival ?*

Dans ce panorama, il ne faut pas oublier la théorie Sens-texte de I. Mel'čuk, qui publie le *Dictionnaire Explicatif et Combinatoire du Français Contemporain –DEC-* (1984), où la base X d'une collocation s'associe à son collocatif Y sur la base d'une fonction lexicale, une relation sémantique abstraite qui s'exprime par la formule f(X)=Y. La fonction lexicale « Magn », par exemple, exprime l'intensité. Appliquée à la base *peur,* cette fonction produit des collocatifs comme *bleue*, *intense*, *grande*, *panique* (Grossmann, Tutin, 2003) ; appliquée à la base *to condemn*, cette fonction produit un collocatif comme *strongly*, appliquée à la base *thin,* cette fonction produit comme collocatif le syntagme *as a rake*, qui traduit en français donne *maigre comme un clou* (Mel'čuk, 1998, p. 32).

Selon Grossmann et Tutin (2003, p. 7-8), les modèles syntaxiques de dépendance qui relient des mots comme la théorie « Sens-Texte » mel'čukienne sont préférables aux modèles qui se basent sur les constituants. En effet, les collocations peuvent relier des constituants contigus formant un syntagme, mais cela n'est pas toujours le cas. Par exemple, l'association « nom + épithète » *steak bleu* peut apparaître séparée dans la phrase *Mon steak est trop bleu*. En outre, un collocatif et sa base, même si contigus, peuvent ne pas former un constituant au sens traditionnel : dans l'association du type « substantif (sujet) + verbe » *Le problème réside…*, les modèles à base de constituants ne prévoient pas que le sujet et le verbe forment une unité. Enfin, la base de la collocation n'est pas systématiquement la tête du syntagme, comme dans le syntagme *de dépit,* où la base est *dépit*, mais la tête du syntagme est *de*.

Toujours dans la tradition lexicographique, une grande contribution à l'étude des collocations a été ajoutée par Benson M., Benson E. et Ilson, qui publient le dictionnaire combinatoire *BBI Dictionary of English* (1986a). Ces auteurs s'inscrivent dans l'approche statistique car ils considèrent les collocations comme « recurrent phrases » (« phrases fréquentes »), mais ils tiennent aussi compte de critères formels. En effet, ils distinguent les collocations lexicales et les collocations grammaticales d'un côté, et de l'autre les collocations, les associations libres et les phrases idiomatiques. Les collocations sont des « loosely fixed combinations » (« associations semi-figées) comme *commit murder* qui, à la différence des associations libres, sont fréquentes et ne peuvent pas s'associer avec de nombreux verbes synonymiques. En outre, à la différence des phrases idiomatiques, elles sont transparentes et leur sens dérive de la somme des sens de leurs composants (Benson et al., 1986b, p. 252-253). Comme expliqué par Orliac (2004, p. 28-36), le BBI se présente comme un dictionnaire d'apprentissage destiné aux apprenants de l'anglais. Deux importants ouvrages lexicographiques destinés aux apprenants de l'anglais existaient avant la publication du BBI : le *Oxford Advanced Learner's Dictionary of Current English* – *OALDCE* – édité par Hornby (1974), publié pour la première fois en 1948 et dont la dernière réédition fut en 2000 ; et le *Longman Dictionary of Contemporary English* - *LDOCE*- (1978). Par rapport à ces deux ouvrages, les auteurs du BBI veulent intégrer le traitement des collocations au traitement de l'information grammaticale (nombre et type de compléments régis), et faciliter leur repérage. En résumé, le BBI ajoute ainsi la combinatoire lexicale à la combinatoire grammaticale des entrées. Selon Orliac (2004), « on peut seulement regretter que le dictionnaire n'ait pas inclus une caractérisation des propriétés sémantiques des collocations, caractérisation jugée également nécessaire par Hausmann » (p. 36).

L'excursus historique que nous venons de présenter se termine avec une description de deux contributions à l'étude des collocations que nous considérons comme étant particulièrement significatives : l'approche plutôt phraséologique de Grossmann et Tutin (2002), qui propose des critères sémantiques permettant de décrire les différents types d'unités phraséologiques et pouvant servir aux applications en linguistique et linguistique appliquée ; et l'approche plutôt statistique de Lo Cascio (1998), qui interprète la collocation de façon plus large, et s'oriente

davantage vers des applications lexicographiques et didactiques. Le fait que les deux approches s'éloignent significativement l'une de l'autre est cependant justifié par un objet de recherche différent.

F. Grossmann et A. Tutin (2003) s'inscrivent plutôt dans le courant phraséologique que dans le courant statistique. Ils adoptent une conception étroite du terme « collocation », qui se base sur les études de Cruse (1986), Mel'čuk (1998) et Hausmann (1989). Leur conception s'éloigne de celle des contextualistes, qui considèrent la collocation comme un phénomène statistique, et ne font pas la distinction entre les associations fréquentes et typiques et celles purement grammaticales, ni entre les associations lexicales paradigmatiques et celles syntagmatiques. Selon les auteurs, les critères qui définissent de manière plus précise ce qu'est une collocation sont les suivants :

1) L'arbitraire

C'est la non prédictibilité de l'association lexicale au niveau sémantique (*pluie torrentielle* mais non * *précipitations torrentielles*) et/ou syntaxique :

« Néanmoins, si l´arbitraire caractérise souvent les réalisations lexicales, on relève que les collocations suivent des patrons syntaxiques précis. Par exemple les collocations construites autour d´un nom incluront des adjectifs ou des verbes, mais probablement pas des adverbes ou des conjonctions » (Grossmann, Tutin, 2002, p. 9).

2) La transparence et le non-figement sémantique

C'est la possibilité de comprendre le sens de l'association à partir du sens de ses composants. Les auteurs citent les exemples de Hausmann *célibataire endurci* et *feuilleter un livre*, tout à fait compréhensibles pour un locuteur non natif du français. Mais le fait que le sens soit beaucoup moins transparent dans certains cas de figure comme *peur bleue* ou *colère noire* ne permettrait pas de généraliser ce critère à l´ensemble de la classe des collocations.

3) Le caractère binaire de la collocation

C'est un trait déjà reconnu par Hausmann (1989) et Mel'čuk (1998) qui indique que l'association se compose de deux éléments, aujourd'hui généralement appelés « base » et « collocatif ». Les auteurs proposent de parler non de constituants mais

de lexies, de façon à intégrer des syntagmes comme *fort comme un turc,* où *comme un turc* est l´élément qualifiant qui s'associe à *turc*.

4) La dissymétrie des composants de la collocation

Ce trait indique que le statut des composants de la collocation est inégal : la base a un sens autonome, tandis que le collocatif perd son sens habituel et acquiert un nouveau sens en association avec sa base.

5) La sélection lexicale ou cooccurrence restreinte

Ce trait indique que la base impose la sélection de son collocatif pour exprimer un sens donné. Par exemple, dans l'expression *peur bleue*, la base s'associe avec le mot *bleue* pour lexicaliser l'intensité. La cooccurrence des deux mots est donc restreinte.

Comme les auteurs l'expliquent, l'utilisation de ces paramètres ne va pas sans poser de problèmes :

« Les cinq paramètres mis en évidence permettent de caractériser des collocations prototypiques comme *célibataire endurci* ou *pluie torrentielle*, mais n'engloberont pas des expressions comme *peur bleue, il pleut à verse, l'âne brait* … parce que le critère d'imprédictibilité pourra paraître contestable avec *l'âne brait*, alors qu'inversement la transparence ne caractérise pas clairement *peur bleue* ou *il pleut à verse* » (Grossmann, Tutin, 2003, p. 4).

C'est pour cette raison que les auteurs utilisent les cinq critères dans un but différent.

D'un côté, ils se servent des trois derniers critères (caractère binaire, dissymétrie et sélection lexicale) pour définir le concept de collocation. Ils reformulent ainsi la définition proposée par Mel'čuk et affirment que :

« Une **collocation** est l'association d'une **lexie (mot simple ou phrasème[33])** **L** et d'un **constituant C** (généralement une lexie, mais parfois un syntagme par exemple *à couper au couteau* dans *un brouillard à couper au couteau*) entretenant une relation syntaxique telle que :

[33] C'est-à-dire une expression complètement figée dans la terminologie mel'čukienne.

- C (le collocatif) est sélectionné en production pour exprimer un sens donné en cooccurrence avec L (la base)
- Le sens de L est habituel » (Grossmann, Tutin, 2003, p. 5).

De l'autre côté, les auteurs utilisent les critères de la transparence et de l'arbitraire pour établir le degré de figement de la collocation, et ils font la distinction entre :

1) Collocation « opaque »

Il s'agit d'une association arbitraire et non transparente. Le collocatif a un sens différent quand il est en association avec sa base, comme par exemple dans *peur bleue*. Ce type de collocation est proche des expressions figées, mais la différence est que dans les expressions figées la base conserve toujours son sens. Comme l'expliquent les auteurs, le collocatif est souvent unique : « par exemple, *bleu* dans le même sens d''intense' n'apparaît qu'en cooccurrence avec *trouille* ou *frousse*, mais de façon moins productive »).

2) Collocation « transparente »

Il s'agit d'une association arbitraire mais transparente. Le sens du collocatif peut être interprété, comme dans *faim de loup* ou *grièvement malade*. C'est la collocation prototypique, qui peut être devinée par le locuteur non natif au niveau de la réception, tandis qu'il y a moins de chances pour qu'il la produise naturellement.

3) Collocation « régulière »

Il s'agit d'une association motivée et transparente et elle peut être de deux types :

a) Le sens du collocatif inclut celui de la base, comme par exemple le mot *bissextile,* qui inclut le sens d'*année* ; le verbe *braire*, qui est un des traits sémantiques de l'âne ; le mot *aquilin*, qui ne s'associe qu'avec *nez* et *profil*, bien qu'*aquilin* signifie originairement « de l'aigle ».
b) Le sens du collocatif est très générique, et l'association produite est très proche d'une expression libre. Considérons, par exemple, le collocatif *grand*, qui porte sur beaucoup de noms et semble être l'intensif plus commun pour les émotions, par exemple dans *grande tristesse*.

Cette distinction faite, réfléchissons sur quelques points critiques de cette définition. De nombreux auteurs intègrent les associations du type *nez aquilin*, *l'âne brait* et *année bissextile* parmi les expressions libres, puisqu'il s'agit d'expressions compositionnelles et motivées, marquées par une solidarité lexicale. Les mots *braire*, *aquilin* et *bissextile* ne prennent pas leur sens en association avec leurs bases, mais ont bien un sens autonome. En outre, le fait qu'ils ne se présentent pas en association avec d'autres bases n'indique pas nécessairement l'existence de la collocation parce que le collocatif n'a pas d'existence autonome. Cette interprétation est, à notre avis, tout autant valide.

On pourrait en outre se demander en quoi la collocation « opaque » diffère des phrases idiomatiques, étant donné que *peur bleue* ressemble à une expression figée, car le sens de *bleue* est difficile à comprendre de la part d'un locuteur étranger (sauf si dans sa langue il existe une collocation similaire). En réalité, cette critique n'est pas recevable, car ce qui différencie la collocation de la phrase idiomatique est son figement moindre. En effet, dans la collocation la base garde son sens, ce qui n'arrive pas pour les phrases idiomatiques, où les deux éléments ont le même statut et l'association n'est pas orientée de la base vers le collocatif. Considérons, par exemple, la phrase idiomatique *croiser les doigts,* où aucun des deux composants ne garde leur sens, et où l'expression en tant qu'unité signifie « espérer que quelque chose se passe bien ».[34]

Les auteurs expliquent leur typologie en concluant :

« La typologie esquissée sera sûrement mise à mal par quelques cas-limites mais cette catégorisation nous paraît toutefois préférable à la solution du continuum qui permet difficilement d'adopter des solutions concrètes pour un traitement formel » (Grossmann, Tutin, 2003, p. 6).

[34] En outre, il ne faudrait pas oublier qu'une unité phraséologique peut posséder une double valeur, selon que la base et le collocatif gardent ou pas leur sens originaire en association. En italien, par exemple, *vedere le stelle (*littéralement *voir les étoiles)* peut signifier « observer les étoiles » (expression libre) ou « ressentir une forte douleur » (phrase idiomatique). En français, *avoir le bras long* a un sens littéral dans la phrase *Il avait le bras assez long pour attraper le chapeau sur l'arbre*, mais acquiert un sens imagé et signifie « avoir de l'influence » dans la phrase *Il a le bras assez long pour faire sauter la manifestation*.

V. Lo Cascio (1997), s'inscrit plutôt dans le courant statistique. L'auteur offre une autre contribution très significative à l'étude des collocations, en particulier dans les domaines de la lexicographie et de la didactique des langues étrangères. Il suggère que (surtout dans une perspective interlinguistique) soit explicité l'univers conceptuel de chaque base, c'est à dire le mini-système incluant ses cooccurents les plus typiques. Ce mini-système inclut les associations basées sur une solidarité sémantique et sur une combinatoire libre, ainsi que les collocations idiosyncratiques communément appelées « collocations », de même que des expressions plus idiomatiques et figées. En effet, l'auteur considère que se limiter aux expressions figées est réducteur, tandis que présenter seulement les associations libres n'est pas fonctionnel.

Les langues codent différemment la diversité propre à la réalité encyclopédique, ce qui ne permet pas de prédire quelle association est difficile à apprendre pour un locuteur étranger. Pour cette raison, le traitement lexicographique devrait présenter toutes les associations « par défaut » (standard) qui construisent l'univers du discours d'un mot : par exemple, pour le mot *pain*, les collocatifs *manger*, *rompre*, *enfourner*, *intégral*, *complet,* etc. D'ailleurs, cette approche ne veut pas nier l'existence et la validité des critères linguistiques qui permettent de différencier les nombreux types d'unités lexicales qui habitent le complexe univers phraséologique. Lo Cascio inclut, dans l'entrée « idéale » d'un dictionnaire, les informations suivantes :

1) Le modèle syntaxique : distribution-attribution catégorielle, attribution des rôles catégoriels et thématiques, et des cas syntaxiques.
2) Les associations préférentielles sémantiques et catégorielles.
3) L'encyclopédie.
4) La fréquence.
5) Le comportement pragmatique et stylistique.
6) Les préférences de forme syntaxique au niveau de la phrase (passive, impersonnelle, etc.).

Par exemple, pour le mot pivot *pain*, il faudrait reconstruire son mini-système en indiquant les « propriétés » suivantes représentées par ses collocatifs :

- opérations sur > verbes - *pain* objet
par exemple : *faire le pain*
- opération de > verbes - *pain* sujet
par exemple : *le pain s'émiette*
- caractéristiques > adjectifs
par exemple : *pain noir*
- spécification :
par exemple : *pain intégral*
- quantification :
par exemple : *morceau de pain*

Ce système permet de construire des tables comparatives pour une langue ou pour plusieurs langues en comparaison. Il s'agit d'une base pour développer des instruments lexicographiques ou de traduction adéquats.

De la même façon, en didactique de langue étrangère, l'enseignant devrait reconstruire les associations les plus fréquentes et les plus typiques contenues dans l'univers du discours autour d'un mot donné, afin d'aider l'apprenant à réussir dans sa communication de base.

En conclusion, ce que cette approche met en évidence est que la tendance d'un mot à choisir un autre mot (avec une fréquence plus significative que le simple hasard) devrait être une priorité de la perspective de la didactique et de l'approche lexicographique.

### *3.2.2 Validité des critères linguistiques*

Dans ce paragraphe, nous allons considérer la validité de six critères linguistiques souvent utilisés pour définir les collocations au sein de l´approche syntaxique. Ces critères sont l'asymétrie, l'arbitraire, le caractère binaire, l'absence ou les restrictions de commutabilité sémantique, la non transparence du collocatif, l'absence ou les restrictions de commutabilité syntaxique :

1) L'asymétrie.[35]

1) L'arbitraire.[36]

2) Le caractère binaire.[37]

Ces trois premiers traits sont assez stables et ne sont pas généralement mis en discussion.[38]

3) L'absence ou les restrictions de commutabilité sémantique.

Il s'agit d'un critère souvent mal interprété, mais lui aussi assez stable. Ce critère nous aide à différencier les collocations et les associations libres. La commutabilité sémantique est la possibilité de substituer un élément de la séquence polylexicale à un autre. Dans les expressions libres, la commutabilité est toujours possible si les règles communes de cohérence syntaxique et sémantique sont respectées. Par exemple, en anglais il est possible de dire *to buy a book* ou *to purchase a book*, mais personne ne dira *to drink a book*, car le fait d'être bu n'est pas un trait sémantique de *livre*. Dans les collocations, qu'elles soient imagées ou non, deux cas sont possibles :

a) La commutabilité sémantique est absente.

Par exemple, en anglais *to take a shower* n'a pas d'expression synonymique ; en italien, dans la collocation *caldo infernale,* le collocatif *infernale* ne peut pas être remplacé par son synonyme *diabolico.*

[35] La présence d'une base autonome et d'un collocatif sémantiquement dépendant.

[36] Le fait que l'association n'est pas motivée, comme dans *heavy rain*, où des synonymes de *heavy* tels que *big et strong* ne sont pas possibles.

[37] Le fait que la collocation se compose d'un pivot et d'un collocatif qui s'y associe typiquement.

[38] A vrai dire, il existe un débat encore ouvert sur la nature des composants de la collocation. Comme le font remarquer Grossmann et Tutin (2002), les composants de la collocation devraient être considérés comme des lexies et non comme des constituants. En effet, les modèles à base de constituants n'acceptent pas une structure du type « substantif en fonction de sujet + verbe » comme *Le problème réside…*. En outre, les auteurs expliquent que la notion de « partie du discours » est inappropriée, car par exemple les collocatifs qui portent sur les adjectifs ne sont pas seulement des adverbes au sens strict, mais aussi des constituants ayant un fonctionnement adverbial. Considérons, par exemple, *ivre mort*, *saoul comme une barrique* et *bête à manger du foin* : un adjectif s'associe avec un autre adjectif, un syntagme prépositionnel et une infinitive respectivement. Cependant, on pourrait discuter de la présence d'une collocation pour les deux derniers exemples, s'agissant-il plus probablement de phrases idiomatiques.

b) la commutabilité sémantique est restreinte.

Par exemple, en italien, *accesso d'ira* et *attacco d'ira* (traduit en français par *accès de colère*) sont possibles, mais *assalto d'ira* ne l'est pas ; en français, *essuyer un échec* et *subir un échec* font partie de l'usage, mais *recevoir un échec* non. Comme nous le remarquons, la substitution synonymique est possible, bien qu'elle soit limitée. A remarquer que, généralement, un des collocatifs est moins fréquent.

Cowie (1998) limite, avec raison, la commutabilité aux synonymes et aux quasi-synonymes, à condition qu'ils maintiennent le sens de la collocation : par exemple, la séquence polylexicale *auburn hair* est restreinte au niveau de la commutabilité car il n'y a pas d'autre expression possible ayant le même sens. D'ailleurs, le fait que la base *hair* s'associe avec de nombreux autres mots (par exemple *dry*, *comb,* etc.) nous montre simplement la largeur du contexte d´usage de la base. En conclusion, quand on considère le critère de commutabilité (à la façon de Cowie), il faut vérifier que le sens de l'expression soit gardé dans le test de commutabilité du collocatif, et non de la base.

4) La non transparence du collocatif

Il s'agit d'un critère assez problématique. La *non transparence* du collocatif est un critère distinctif des collocations par rapport aux expressions figées. En effet, selon des nombreux auteurs (voir par exemple Grossmann et Tutin, 2002), l'opacité sémantique concerne seulement le collocatif dans les collocations, et la base et le collocatif dans l'expression figée. Des associations non transparentes comme *peur bleue,* par conséquent, peuvent être considérées comme des collocations, car *peur* garde son sens littéral. Si cette distinction apparaît assez claire, il faut cependant reconnaitre qu'il n´est pas simple de décider ce qui est sémantiquement transparent, car un certain arbitraire est toujours à l'œuvre. Pensons aux suites *colère noire et peur bleue*, qui peuvent être considérées comme des collocations en raison de l'autonomie sémantique de la base : elles restent cependant peu transparentes pour un locuteur étranger, et d'ailleurs *colère noire* semble l'être plus que *peur bleue* (car *noir* a généralement une connotation négative). Ces deux exemples nous montrent que la transparence peut être un facteur relatif, selon qu'on l'envisage dans la perspective de la langue maternelle ou de la langue seconde. Dans la première perspective, est opaque tout ce qui relève d´un procès de translation du sens comme

la métaphore ou la métonymie ; dans la seconde perspective, est opaque tout ce qui diffère de la structure correspondante dans la langue maternelle.

5) L'absence ou les restrictions de commutabilité syntaxique

Ce dernier critère, est lui aussi assez problématique. Il nous aide à distinguer les collocations des expressions libres d'un côté, et les collocations des expressions figées de l'autre, mais il ne peut pas être appliqué de façon systématique. Considérons d'abord la différence entre collocations et expressions libres. Comme le font remarquer Grossmann et Tutin (2003), les restrictions syntaxiques spécifiques des collocations par rapport aux associations libres ne sont pas systématiques. Voici quelques exemples de restrictions :

- L'absence du déterminant dans les constructions à verbe support

On dit *avoir faim* mais *essuyer un échec*, *perpétrer un délit.*

- Peu de flexibilité au niveau de la distribution syntaxique[39].

Si on compare *peur bleue* et *peur immense,* on remarquera la moindre flexibilité syntaxique de la collocation *peur bleue*, qui ne peut pas se transformer en **La peur de Jean a été bleue*, tandis qu'il est possible de dire *La peur de Jean a été immense.* D´ailleurs, dans ce cas également, la règle n'est pas systématique, car *steak bleu* est plus flexible que *peur bleue* et permet une variation syntaxique dans la phrase *Léo aime le steak quand il est bleu.*

- La présence d'alternances syntaxiques

Comparons *mener une attaque* et *raconter une attaque*, respectivement à verbe support et à verbe plein : on découvre une contrainte sur les déterminants pour **Luc mène cette attaque contre la citadelle* (tandis que *Luc raconte cette attaque contre la citadelle* est correct).

Considérons, enfin, la différence entre collocations et expressions figées. Dans les expressions figées, la restriction dans la commutabilité syntaxique n'est pas forcement présente : la phrase idiomatique *poser un lapin*, par exemple, admet l'insertion d'un modifieur (dans *poser un gros lapin*), même si elle est complétement figée. La même remarque vaut pour l'expression figée *jeter la lumière sur quelque*

[39] « Les collocatifs épithètes qui ont un fonctionnement prédicatif ne peuvent pas tous apparaître comme attributs, alors que dans les associations libres, ce contexte est pratiquement toujours possible avec les adjectifs ayant cette fonction sémantique » (Gross, 1988, cité par les auteurs).

*chose*, qui apparaît cependant d'interprétation plus facile.

### *3.2.3 Conclusions : quelle approche pour l'étude des collocations ?*

Comme nous l'avons expliqué, les nombreuses théories sur les collocations que nous venons de décrire se sont développées au sein de deux grands courants de recherche, l'approche statistique et l'approche syntaxique. Après ce long excursus des théories développées au sein de ces deux approches, essayons de « faire le point » en résumant en quoi consiste leur divergence et proposons une solution d'équilibre.

Dans leur analyse, Granger et Paquot (2008) expliquent que, bien que la diversité des approches soit une richesse, elle a contribué à la construction d'un concept d'unité phraséologique très flou.

L'approche phraséologique s'est plutôt intéressée à l'individuation de critères linguistiques (syntaxiques et sémantiques) pouvant distinguer les différents types d'unités phraséologiques. Comme le remarquent Frath et Gledhill (2006), cette approche explique les assemblages collocationnels en terme d'opérateurs/fonctions (à la Mel'čuk) ou de relations entre les parties de l'assemblage (qu'il s'agisse de relations sémantiques telle que la composition, comme dans *stone furniture,* ou qu'il s'agisse de relations prédicatives telle que la cause, comme dans en *malarial mosquitos)*. Dans cette optique nous trouvons par exemple Hausmann (1989), selon qui les composants de la collocation doivent entretenir une relation syntaxique spécifique (dont le rapport « article + substantif » est exclu par exemple) ; et Howarth (1996), qui définit les collocations restreintes comme :

« Restricted collocations are combinations in which one component is used in its literal meaning, while the other is used in a specialised sense. The specialised meaning of one element can be figurative, delexical or in some way technical»[40] (p. 47).

[40] « Les collocations restreintes sont des combinaisons dont un composant est utilisé dans son sens littéral, tandis que l'autre est utilisé dans un sens spécialisé. Le sens spécialisé de l'un des composants peut être figuré, delexicalisé ou en quelque sorte technique » (notre traduction).

Dans cette approche, les unités plus transparentes et libres apparaissent comme les moins prototypiques parmi les unités phraséologiques, tandis que les phrases idiomatiques seraient les plus prototypiques.

Par contre, l'approche statistique a été étroitement liée aux domaines de la lexicologie et de la lexicographie, et a proposé une analyse inductive à partir d'un repérage des cooccurrences lexicales dans les corpus. Un des représentants les plus importants de cette approche est Sinclair (1991, p. 71), selon qui la collocation est « [...] a tendency for words to occur together »[41]. Cette approche a également inclus parmi les unités phraséologiques, des associations moins figées mais qui constituent un choix fréquent de la part du locuteur. Ce sont des associations qui ne sont pas prises en considération par l'approche phraséologique, bien qu'elles soient plus fréquentes que les phrases idiomatiques ou les proverbes (Moon, 1998). Comme l'expliquent Frath et Gledhill (2005), la démarche statistique s'est développée autour d'outils toujours plus performants offerts par l'étude des corpus électroniques, et a renforcé la vision novatrice de l'usage langagier selon laquelle la phrase est une unité lexico-grammaticale (et non une structure syntaxique à remplir avec des sens isolés). Cependant, bien que l'utilité des outils informatiques et des mesures statistiques pour le traitement des grands corpus soit évidente, nombreux sont ceux qui restent encore méfiants envers l´approche statistique.

Enfin, nous voudrions préciser que les deux approches ont connu en leur sein quelques désaccords (Nesselhauf, 2005). Dans l'approche statistique, certains considèrent comme collocations tout type de cooccurrence fréquente, d'autres considèrent comme collocations seulement les coccurrences les plus fréquentes, d'autres encore s'éloignent encore un peu plus de l'interprétation firthienne originaire en acceptant les associations fréquentes où la relation grammaticale entre les composants est étroite. Le fait que les statisticiens privilégient la fréquence est beaucoup critiqué par l'approche phraséologique. Du côté de l'approche phraséologique également, la distinction entre les différentes unités non pragmatiques (les unités pragmatiques étant plutôt laissées à l'écart) se révèle souvent incohérente en ce qui concerne les critères utilisés : Hausmannn (1989)

[41] « [...] une tendance des mots à se présenter ensemble » (notre traduction).

utilise le critère de la commutabilité (le fait que des substitutions lexicales soient permises à l'intérieur de la collocation) pour distinguer les collocations des associations libres, et le critère de la transparence (le fait qu'un composant de la collocation ou que la collocation dans son ensemble garde un sens littéral) pour opérer une distinction entre collocations et phrases idiomatiques ; Mel'čuk (1998) utilise la commutabilité comme première distinction, et les deux critères ensemble pour la seconde distinction. En outre, certaines classifications ne différencient pas « collocations restreintes » et « collocations ouvertes ». Elles ne font pas non plus de différence entre « phrases idiomatiques imagées » et « phrases idiomatiques pures », comme le fait Cowie (1981). Cependant, lui non plus ne considère pas cette classification de manière rigide.

En résumé, le principal point de contraste entre l´approche statistique et l´approche phraséologique concerne les critères clés utilisés pour le repérage des collocations, respectivement la fréquence d'occurrence des composants de l'association *versus* l'existence de relations formelles syntaxiques et/ou sémantiques précises entre les composants de l'association. Le courant statistique considère la fréquence comme l'élément principal pour le repérage des collocations, le courant phraséologique la considère comme un critère secondaire par rapport aux relations d'ordres syntaxiques et/ou sémantiques qui ont lieu entre les composants de la collocation.

Cette première observation étant faite, les deux approches, bien qu'elles s'éloignent beaucoup l'une de l'autre sur certains points clés, ont en leur sein plusieurs positions moyennes qui préfèrent une méthode complémentaire.

D'un côté, en faveur de l´approche statistique, deux considérations principales peuvent être avancées. En premier lieu, la fréquence ne peut pas être ignorée dans les études sur le lexique, car elle est un critère objectif de sélection des « paquets » lexicaux présents dans l'input. En second lieu, pour récuser les accusations portées contre la validité du calcul statistique, nous précisons qu'à l´heure actuelle aucun représentant de l´approche phraséologique ne tombe dans le piège de considérer la fréquence comme une valeur discriminatoire et absolue. Les études basées sur corpus essaient de relativiser les erreurs de comptage de fréquence en utilisant des mesures associatives, et distinguent les phénomènes syntagmatiques réguliers des

phénomènes syntagmatiques restreints ou arbitraires. Comme le fait remarquer Hoey (2005, p. 3), la collocation est le lien qu'un item lexical instaure avec d'autres items qui apparaissent dans son meme context plus fréquemment que la simple probabilité le prédirait. Selon l'auteur, il ne faudrait pas donner la même importance à des cooccurrences non significatives du type *the student* (où *the* est en mot fréquent de manière absolue) et à des structures fréquentes et significatives de type lexical telles que les collocations.

D'un autre côté, en faveur de l´approche syntaxique, on remarquera que le recours à des critères linguistiques/fonctionnels (syntaxiques ou sémantiques) plutôt que statistiques/textuels rend mieux compte de la complexité du vaste domaine de la phraséologie, et se révèle indispensable pour des analyses descriptives plus fines ainsi qu'en TAL.

D'ailleurs, les approches mixtes, qui essaient de trouver un équilibre entre la statistique et l'analyse formelle, sont de plus en plus répandues.

Une fois ces précisions apportées, la description de deux théories que nous considérons comme étant très significatives, la typologie de Grossmann et Tutin (2003) d'un côté, et la théorie de Lo Cascio (1997) de l'autre, a ouvert une réflexion sur la fonctionnalité de la définition de collocation : les deux approches représentent deux perspectives différentes, mais elles sont identiquement valides au vu de l'objectif qu'elles se proposent d'atteindre par l'analyse du phénomène de collocation. Dans un but descriptif ou pour des applications en TAL, une caractérisation très détaillée qui distingue les types d'unités phraséologiques existantes et qui repose sur des critères formels précis apparaît comme étant la plus fonctionnelle. Complémentairement, dans des buts didactiques ou lexicographiques, une conception plus large de la collocation semble le choix idéal. En effet, pour un apprenant étranger, ce qui est ou non collocation dépend des divergences existantes avec sa langue maternelle et, pour cette raison, il devrait pouvoir accéder aux associations plus importantes sur le plan communicatif ou aux associations les plus fréquentes d'un mot donné (son univers contextuel). La même remarque vaut dans un but lexicographique : comme l'explique Mel'čuk (1998, p. 24), un bon dictionnaire devrait inclure tous les phrasèmes, parce que le phrasème est par définition non

compositionnel et il ne peut pas être produit à travers les règles les plus communes de la syntaxe. Les phrasèmes sont utilisés et mémorisés comme de véritables blocs.

Dans le chapitre 4 nous formulerons une définition de travail de la collocation, qui aidera à distinguer les collocations des autres types d'unités phraséologiques.

# CHAPITRE 4
# Langage préfabriqué et collocations : une définition de travail

Dans ce chapitre, au paragraphe 4.1 nous présenterons la vision novatrice de l'usage langagier ainsi que deux concepts importants qui lui appartiennent, les concepts de *lexical priming* (4.1.1) et de *lexique-grammaire* (4.1.2). Au paragraphe 4.2, nous présenterons une définition de travail de la collocation.

## 4.1 Une vision novatrice de l'usage langagier

Une vision novatrice de l'usage langagier envisage le langage comme un phénomène intrinsèquement social. Cette vision réaffirme l'ancien principe de l'économie linguistique selon lequel le locuteur se servirait, pour apprendre une langue, d'unités préfabriquées, prêtes à l'usage, fréquentes et typiques. La grammaire générative avait affirmé l'importance de la créativité linguistique au détriment des théories behaviouristes, et l'étude des séquences phraséologiques avait donc reçu très peu d'attention. Cependant, depuis quelques décennies, les théories cognitivistes ont mis en cause le système abstrait des génératives. Le cognitivisme minimise la distinction nette entre syntaxe et lexique, et réaffirme l'importance de l'automatisation et de la réutilisation de blocs lexicaux fonctionnels dans l'apprentissage linguistique. C'est l'usage, plutôt que l'abstraction, qui explique le fonctionnement de la langue. Legallois (2005) cite deux modèles anglo-saxons qui se sont imposés dans le courant cognitiviste : la « Grammaire de Construction » (représentée par exemple par Fillmore –Fillmore et al., 1988 - ; Goldberg, 1995 ; Langacker, 1987 ; Lakoff, 1987 ; Croft et Cruse, 2004) selon laquelle les unités linguistiques sont symboliques et les niveaux syntaxique et sémantique se confondent ; et la « Grammaire des Patterns » (représentée par les linguistes anglais

Francis et Huston, 2000) qui continue les travaux de l'école contextualiste et étudie « les patrons distributionnels qu'intègrent certaines classes de mots sémantiquement homogénéisés par le pattern » (p. 114). Selon l'auteur (à qui nous renvoyons pour approfondissement), ces deux modèles reconnaissent le rôle primaire de la phraséologie pour l'apprentissage linguistique et permettent de parler d'un véritable « tournant phraséologique de la grammaire » (p. 109).

Dans ce courant de réaction au générationalisme, les théories psycholinguistiques jouent, elles aussi, un rôle important : elles décrivent le développement linguistique comme un procès qui va des séquences conventionnelles et stéréotypées aux séquences créatives. Selon Ellis (2002), la séquence acquisitionnelle la plus naturelle a la structure « formula > low-scope pattern > creative constructions » : l'acquisition est un procès qui va des unités non analysées aux constructions produites de façon créative en passant par l'analyse des structures de la langue. L'émergence d'un système créatif serait liée à la décomposition des unités préfabriquées. Wray et Perkins (2000, p.1) partagent la même opinion. Ils considèrent l'unité phraséologique comme une séquence préfabriquée, mémorisée et restituée au moment de l'usage.

Il s'agit donc d'une nouvelle vision langagière, dont les concepts de *lexical priming* et de lexique-grammaire font partie intégrante.

### *4.1.1 Lexical-priming et apprentissage par blocs*

Dans le courant anglo-saxon, Hoey (2005) accorde une place centrale aux séquences figées dans l'apprentissage linguistique. Selon l'auteur, il est possible de produire deux phrases correctes grammaticalement, mais seule l'une des deux est naturelle, tandis que l'autre apparaît bizarre ou créative. En effet, les locuteurs natifs connaissent la façon dont les mots s'associent naturellement et produisent des associations privilégiées et standardisées. Hoey utilise le terme de *priming* pour indiquer la façon dont un mot implique un autre et pour expliquer l'existence des collocations. Selon l'auteur tout mot est mémorisé dans une collocation. Il explique que les collocations, en tant qu'associations fréquentes, sont avant tout un phénomène psychologique car on apprend un mot après l'avoir rencontré en contexte, en compagnie de ses collocatifs les plus fréquents. Dans cette optique, la

conception traditionnelle de la grammaire comme systématique et du lexique comme non systématique est fausse, car la grammaire n'est que l'« output » des séquences conventionnelles. En outre, la grammaire d'un locuteur est différente de la grammaire d'un autre locuteur car leur expérience du langage routinier et de ses formules typiques n'est pas la même. Le *priming* n'est pas un facteur permanent des associations de mots parce qu'il est « nourri » à chaque répétition des paquets lexicaux. Il se modifie dans le temps lorsqu'un « drift in the priming » a lieu : il se renforce ou il s'affaiblit selon que l'association entre deux mots est perçue ou non comme typique, familière ou fréquente. Cela concerne plusieurs membres d'une communauté linguistique en même temps, générant un changement temporaire ou permanent dans la langue.

Dans le même courant, Carter et Schmitt (2004) expliquent que les unités phraséologiques, en particuliers les unités les plus figées et les plus opaques, sont mémorisées dans le lexique mental comme des entrées unitaires. Cette vision holistique est confirmée par le fait qu'en production ces séquences sont prononcées plus vite que leurs correspondants créatifs, et qu'en réception, elles sont lues plus vite. Selon les auteurs, il est probable que ces séquences ne soient pas apprises d'un coup, mais au fur et à mesure que le locuteur comble ses défauts de connaissance. Cette considération apparaît tout à fait légitime si on considère que le lexique est appris de façon incrémentale : plus les rencontres avec une séquence sont nombreuses, plus s'accroît la connaissance que le locuteur en a.

### *4.1.2 Le lexique-grammaire*

Comme l'expliquent Altenberg et Granger (2002), dans les approches récentes sur le lexique, ce qui auparavant était perçu comme un phénomène syntaxique est généralement vu comme la projection de propriétés lexicales. L'idée à la base de ces nouvelles approches est qu'il existe un lexique-grammaire au sein duquel il n'y a pas de hiérarchie entre lexique et grammaire, les deux niveaux interagissant sur l'axe syntagmatique sous l'effet de l'usage des unités préconstruites.

Le concept de lexique-grammaire remonte à Maurice Gross (1975). La grammaire transformationnelle de Gross décrit systématiquement le fonctionnement

syntaxique et sémantique des catégories grammaticales du verbe, du nom et de l'adjectif en français, soulignant l'importance de concevoir lexique et grammaire comme des niveaux en étroite corrélation.

Comme l'explique Jackendoff (2002, p. 124) dans *Foundations of language*, la sémantique ne doit pas être pas considérée comme dérivant de la syntaxe, mais comme un système génératif indépendant qui se lie à la syntaxe à travers une interface. La sémantique produit les idées que la langue actualise ; la syntaxe et la phonologie sont les instruments à travers lesquels les idées sont converties en phrases réelles.

Sinclair (1991) est un autre important défendeur de l'étude de la langue en contexte. Il explique que, dans une étude de corpus, une concordance exhaustive devrait prendre en compte tous les *tokens* (formes) d'un même lemme et analyser attentivement leur sens en contexte. Il prend comme exemple le lemme *decline* en anglais : il étudie son comportement dans une étude de corpus et il remarque qu'il se présente 136 fois comme verbe et 109 fois comme nom. Cependant, suite à une observation plus fine, il remarque que *declining* n'est pas toujours un verbe parce qu'il est plus fréquemment utilisé comme modifieur. De son analyse, il résulte que *decline* a un usage plutôt nominal, *declining* adjectival et *declines* et *declined* verbal. En outre, *decline* signifie « refuser » seulement dans la forme *declined* (dans la majorité des cas comme « past simple » ou participe passé, et très rarement comme un verbe indéfini ou comme argument du verbe). Cet exemple montre l'existence d'un lien très fort entre le sens et la structure syntaxique qui le réalise : les sens sont souvent associés à des structures syntaxiques spécifiques.

La notion de lexique-grammaire est étroitement liée à la notion de productivité. Comme l'explique Gonzalez Rey (2002), il existe dans la production linguistique des structures unitaires prêtes à l'usage :

> […] il existe des schémas ou moules phraséologiques dotés d'éléments simples, d'ordre relationnel ou catégoriel, suivis de cases vides à combler par des lexèmes différents » (p. 58).

Comme le font remarquer Schmitt et al. (2004), les structures prêtes à l'usage dont les locuteurs disposent peuvent être plus ou moins flexibles. Considérons une structure du type *Watch out !* et une structure du type « ___ *think nothing of* ___ » dans la phrase *I think nothing of spending all my money when travelling* : les deux structures sont utiles, mais les dernières sont productives et peuvent être utilisées comme modèles pour produire d'autres sens en utilisant d'autres mots.

## 4.2 Définition de travail de la collocation

La collocation est une séquence polylexicale qui actualise un mot dans une unité sémantico-syntaxique typique, et qui se caractérise par le sémantisme transparent de la base et le sémantisme restreint du collocatif.

Cette définition met en évidence que :

1) La collocation est conçue comme une séquence polylexicale car elle résulte de l'association entre deux ou plusieurs éléments lexicaux[42] (ce qui exclut les rapports de colligation). Autrement dit, la collocation n'est pas une association dont l'un des composants est un élément grammatical : une expression du type *du coup* qui associe un élément grammatical et un élément lexical ne peut pas être considérée comme une collocation, mais plutôt comme une locution au sens strict.
2) La collocation s'actualise dans une unité sémantico-syntaxique typique[43], c'est-à-dire qu'elle est représentative de la façon dont un concept est mis en contexte par les locuteurs d'une langue. Comme l'explique bien Sinclair (1991, p. 44), à un sens correspond une structure, ainsi le sens influence la structure et vice-versa. Les sens s'actualisent dans des patterns syntaxiques spécifiques

[42] Les patterns syntaxiques dans lesquels les collocations apparaissent sont de nature variée et ne peuvent pas être interprétés selon la vision classique de la théorie des constituants. En effet, bien que de différentes classifications de patterns syntaxiques aient été proposées par de nombreux auteurs, en premier par Hausmann, elles souffrent des exceptions. Les cas irreguliers sont nombreux, par exemple *ivre mort* (cf. Grossmann et Tutin, 2003 pour un approfondissement et une liste des cas syntaxiques réguliers et irréguliers qui peuvent s'ajouter à la classification traditionnelle d'Hausmann).

[43] Pour « typique » on peut aussi entendre « fréquente », car l'association entre une base et son collocatif est en tout cas plus significative que le simple hasard.

et dans des contextes d'occurrence typiques. En anglais, par exemple, le verbe *to decline* dans le sens de « refuser » se présente toujours au passé composé comme dans l'exemple *He declined his invitation* (Sinclair, 1991).[44]

3) La collocation se caractérise par le sémantisme transparent de la base, car la base garde toujours son sens littéral, tandis que le sens du collocatif peut être transparent ou opaque. Considérons les collocations *commettre un crime* et *caresser l'espoir* : dans les deux cas, le substantif en fonction de base est transparent ; le verbe collocatif, par contre, apparaît plus transparent dans la première association que dans la seconde. Cela explique pourquoi la collocation reste plus ou moins compréhensible au décodage pour un locuteur étranger.
4) La collocation se caractérise par le sémantisme restreint du collocatif par rapport à sa base. Le collocatif peut avoir à la fois un sens littéral ou non littéral, mais il a toujours un sémantisme restreint : son association avec la base est exclusive ou quasi-exclusive dans un sens spécifique (le « meaning by collocation » firthien). Autrement dit, la commutabilité synonymique du collocatif est exclusive ou limitée[45] : en anglais, il est possible de dire *lose control*, mais il n'est pas possible de dire *miss control*. Cela dérive du fait que la collocation est soumise à des contraintes de nature sémantique imposées par l'arbitraire de l'usage linguistique.

Pour conclure, une importante considération à faire concerne les contraintes agissant au niveau syntaxique (par exemple le blocage de certaines propriétés transformationnelles) : elles ne sont pas toujours en actions et ne définissent pas un type d'unité phraséologique par rapport à l'autre car une forte variabilité est présente. Par exemple, les phrases idiomatiques peuvent avoir une certaine flexibilité syntaxique, bien que ce cas ne soit pas très fréquent. Elles doivent donc être considérées comme un élément possible mais non systématique.

[44] Nous souhaitons préciser que le fait que la collocation s'actualise dans une structure sémantico-syntaxique spécifique implique la présence de contraintes agissant à ces deux niveaux.

[45] Parfois, même si des synonymes sont admis, ils ne sont pas très fréquents en association avec cette base-là et l'association ne correspond pas à la façon la plus naturelle d'exprimer un concept (cf. l'exemple *part de gâteau* à page 81).

*4.2.1 La transparence de la base et le sémantisme restreint du collocatif*

Nous considérons la coprésence des points 3) et 4) comme un trait qui caractérise la collocation par rapport aux autres unités polylexicales. En effet, la transparence de la base est un trait en commun avec les combinaisons libres, tandis que le sémantisme restreint du collocatif est un trait que la collocation a en commun avec les phrases idiomatiques. Cela veut dire que seules les collocations regroupent ces deux traits.

La transparence de la base est un trait plutôt simple à reconnaitre. La base garde toujours son sens littéral, tandis que le collocatif peut être transparent ou opaque, selon l'association dont il fait partie : dans *peur bleue,* le collocatif est plus imagé que dans *suivre un conseil,* qui à son tour est plus opaque que *commettre un crime*, où *commettre* garde son sens prototypique et n'acquièrt pas de nouveau sens en association. Quand aussi la base perd son sens littéral, on a affaire à des phrases idiomatiques, comme dans l'exemple italien *perdere la testa per qualcuno* (traduit littéralement par « perdre la tête pour quelqu'un »), qui signifie « tomber fou amoureux », et où ni le sens de *perdre* ni le sens de *tête* sont gardés[46].

Par contre, le sémantisme restreint du collocatif est parfois un trait difficile à établir : un test de commutabilité synonymique permettrait d'exclure les associations où d'autres synonymes corrects sur le plan sémantico-conceptuel ne sont pas cepedant admis. On pourrait remettre en question la validité du test synonymique du fait que la synonymie absolue est très rare. Tout au contraire, ce dernier point milite en sa faveur, étant donné que la raison pour laquelle la synonymie absolue est très rare est notamment l'actualisation du sens en contexte. L'usage et le voisinage combinatoire d'un mot contribuent à définir son sens et imposent des contraintes sur l'utilisation d'autres candidats, ce dont dépend la rareté de la synonymie absolue. Voilà donc la perspective dans laquelle nous considérons la question : les liens collocationnels fournissent une explication de la rareté de la synonymie absolue.

[46] De nombreux auteurs affirment que le sémantisme du collocatif s'actualise en association avec sa base et que le collocatif acquiert un sens spécifique, nouveau et non littéral en cooccurrence (le célèbre exemple de Hausmann (1989) de *célibataire endurci* illustre ce sens). Cependant, comme nous l'avons expliqué, cela n'est pas toujours le cas.

Pour examiner comment agit le sémantisme restreint du collocatif, nous allons présenter quelques exemples :

a) *commit murder* (Benson, 1986b) > le seul collocatif qui génère le même sens pour la séquence polylexicale anglaise *commit murder* est *perpetrate* (p. 253). Mais *perpetrate* est moins fréquent, n'est pas typique et ne « parlerait » pas probablement à un locuteur natif de l'anglais. La même remarque vaut pour le français *commettre un meurtre* et *perpétrer un meurtre*, sauf que *perpétrer* n'est pas autant rare que son correspondant anglais.
b) *lose control* > collocation anglaise correspondant au français « perdre le contrôle », elle n'admet pas que le collocatif *lose* soit remplacé par son synonyme *miss.*
c) *eine Entscheidung treffen* > cette collocation allemande, qui signifie *« prendre une décision »* et qui est traduite littéralement par **rencontrer une décision,* n'admet pas la substitution du collocatif *treffen* par un synoynme ayant le même sens, par exemple *sehen.*
d) *part de gâteau* > un synonyme tels que *ration* n'est pas admis et d'autres synonymes (par exemple *portion*) sont possibles mais non fréquents dans le discours oral et écrit des natifs.
e) *mariage de raison* > *de raison* peut être substitué par *d'intérêt*, mais d'autres synonymes tels que *de convenance* ne sont pas possibles.
f) *strong coffee* > et non **powerful coffee.*
g) *caresser l'espoir* > l'espoir ne peut pas être touché, on ne peut que le caresser.
*h)* *peur bleue* > seulement *bleu,* et non d'autres couleurs, exprime le sens de « grande » en association avec *peur.*

## 4.3 Collocations et autres séquences lexicales

La définition que nous venons de formuler inscrit les collocations à mi-chemin entre combinaisons libres et phrases idiomatiques car les combinaisons libres n'ont aucune contrainte sémantique tandis que les phrases idiomatiques sont conçues comme des unités complètement figées dont les composants ont le même statut sémantique. En

ce sens, la collocation serait une unité semi-figée. A ce propos, il faudra ajouter quelques mots sur la nature des unités lexicales autres que les collocations.

### *4.3.1 Les phrases idiomatiques*

Les phrases idiomatiques telles que *poser un lapin* sont des unités dont les composants perdent leur sens littéral et génèrent, en association, un nouveau sens global. Cependant, elles peuvent apparaître plus ou moins transparentes selon qu'elles résultent d'un processus de translation ou de métaphorisation d'interprétation facile ou au contraire obscure (comparons *être aux anges* et *poser un lapin*, la première interprétable contrairement à la seconde). A la différence des collocations, elles sont donc non-compositionnelles. Dans les collocations, la base garde toujours son sens littéral tandis que le collocatif peut garder ou non son sens littéral (comparer *peur bleue* et *commettre un meurtre*). En outre, dans les collocations, le réfèrent correspond toujours à l'objet désigné (une *peur bleue* reste une *peur),* ce qui n'est pas le cas dans les phrases idiomatiques (dans *perdre la tête*, *pied noir* et *tirer son chapeau,* on n'a plus affaire à une « tête », ni à un « pied », ni à un « chapeau »). Enfin, les phrases idiomatiques sont généralement plus figées que les collocations car elles sont conçues comme des blocs lexicaux, bien qu'une certaine flexibilité syntaxique soit parfois admise : par exemple, la phrase idiomatique *tomber dans les pommes*, qui veut dire « s'évanouir », ne permet pas l'utilisation d'un autre déterminant ou de la forme au singulier, mais elle permet l'insertion de l'adverbe *subitement*).

### *4.3.2 Les combinaisons libres*

Dans les combinaisons libres, à la différence de ce qui se passe dans les phrases idiomatiques, les composants de l'association gardent tous les deux leur sens littéral et leur acception prototypique. Par exemple dans l'expression libre *gagner de l'argent*, le collocatif *gagner* signifie « obtenir », tandis que dans la collocation *gagner du poids*, il prend le sens d'« ajouter». Un autre exemple est l'expression libre *toucher une feuille*, où le collocatif *toucher* signifie « mettre la main sur quelque

chose », tandis que dans *toucher un chèque*, le collocatif prend le sens d'« encaisser » (exemple issu de Larivière, 1998). Cependant, comme nous l'avons remarqué précédemment, les collocations peuvent elles aussi garder leur sens littéral, comme dans *commit a murder*. La vraie différence entre combinaisons libres et collocations est la présence d'un sémantisme restreint dans le cas des collocations. De toute façon, la distinction apparaît assez délicate, d'où l'existence de diverses théories à ce propos :

1) La théorie des « classes d'objets » de Gross (1994)

Cette théorie semble pouvoir aider à définir les combinaisons libres, étant donné que dans ce type d'associations la base s'associe à un groupe limité de classes sémantiques de compléments, les « classes d'objets » notamment. Si le collocatif qui se présente en association avec une base n'est pas inclus dans sa classe d'objets, il s'agit d'une unité figée. Qu'est-ce que l'on peut « prendre » ? Une classe d'objets bien définis tels qu'une fourchette, des clés, etc. ; par contre, on ne peut pas *prendre une douche* dans l'acception standard d'« attraper ». Ainsi, *regarder les étoiles* est une expression libre mais *prendre une douche* ne l'est pas, car *douche* ne rentre pas dans la « classe d'objets » standard de *prendre* et d'autres synonymes tels que *faire* ne sont pas admis.

2) Le concept de « contraintes de sélection » de Katz et Fodor (1964, cités par Nesselhauf, 2005)

Cette théorie du courant génératif est également utile pour définir les combinaisons libres. Ce concept exprime la présence de contraintes sémantiques par l'utilisation de traits sémantiques. Par exemple, la contrainte de sélection qui opère sur *tuer* établit l'impossibilité de dire *tuer une chaise* car ce verbe ne se combine qu'avec des objets qui ont le trait sémantique [+ animé]. Ce phénomène doit être distingué de la « contrainte collocationnelle » (« collocational restrictions » selon Herbst, 1996 et Cruse, 1986) qui opère sur **café puissant,* qui ne dépend pas du fait que le sens de *puissant* est éloigné de celui de *fort,* mais du fait que l'expression correcte dans l'usage est *café fort.* Les « contraintes de sélection », qui dépendent du sens du mot, produisent les combinaisons libres, tandis que les « contraintes collocationnelles », qui dépendent de l'usage, produisent les collocations.

*4.3.3 Les constructions à verbe support et les affinités lexicales univoques*

Enfin, il faudra ajouter quelques mots sur des catégories particulières de collocations, les « constructions à verbe support » et les « affinités lexicales univoques » :

1) Constructions à verbe support

Terme introduit par M. Gross (1981), la construction à verbe support est un type particulier d'unité polylexicale qui, dans la terminologie anglaise, correspond à « stretched verb construction » (Allerton, 2002), « light verb construction » (O. Jespersen, 1942), « delexical verbs » (Collins Cobuild, 1992) ou « support verbs » (Heid, 1994).

Comme l'explique Ježek (2005), ce type de construction se compose généralement d'un verbe et d'un nom (qui est souvent précédé d'un article et rarement d'une préposition), par exemple *prendere una decisione* en italien, *prendre une décision* en français et *make a decision* en anglais. Ci-dessous nous présentons les traits principaux qui caractérisent les constructions à verbe support :

a) Les constructions à verbe support subissent des contraintes liées à l'usage.
b) Les composants sont généralement autonomes et peuvent subir des transformations, comme dans l'exemple *prendre la/une/beaucoup de décisions*. Cela n'est pas le cas si le nom n'est pas référentiel, comme dans *get some sleep,* où les composants ne sont pas autonomes syntaxiquement (Heid, 1994).
c) Les constructions à verbe support sont des prédicats analytiques pour lesquels il est généralement possible de trouver un correspondant synthétique. L'association entre le nom *appel* et le verbe support *faire* équivaut au verbe dont le nom dérive *appeler*.
d) Le verbe est « vide » ou delexicalisé et le nom supporté est prédicatif. Le verbe ne sert qu'à apporter des informations sur les marques de mode, de temps, d'aspect et de personne ; en outre, son sens est générique à cause des contraintes imposées par l'aspect, ce qui fait qu'*être* indique un état dans *être en doute*. Le nom garde son sens littéral et détermine le sens du verbe collocatif, comme dans *prendre une décision,* où *prendre* apparaît « vide » et

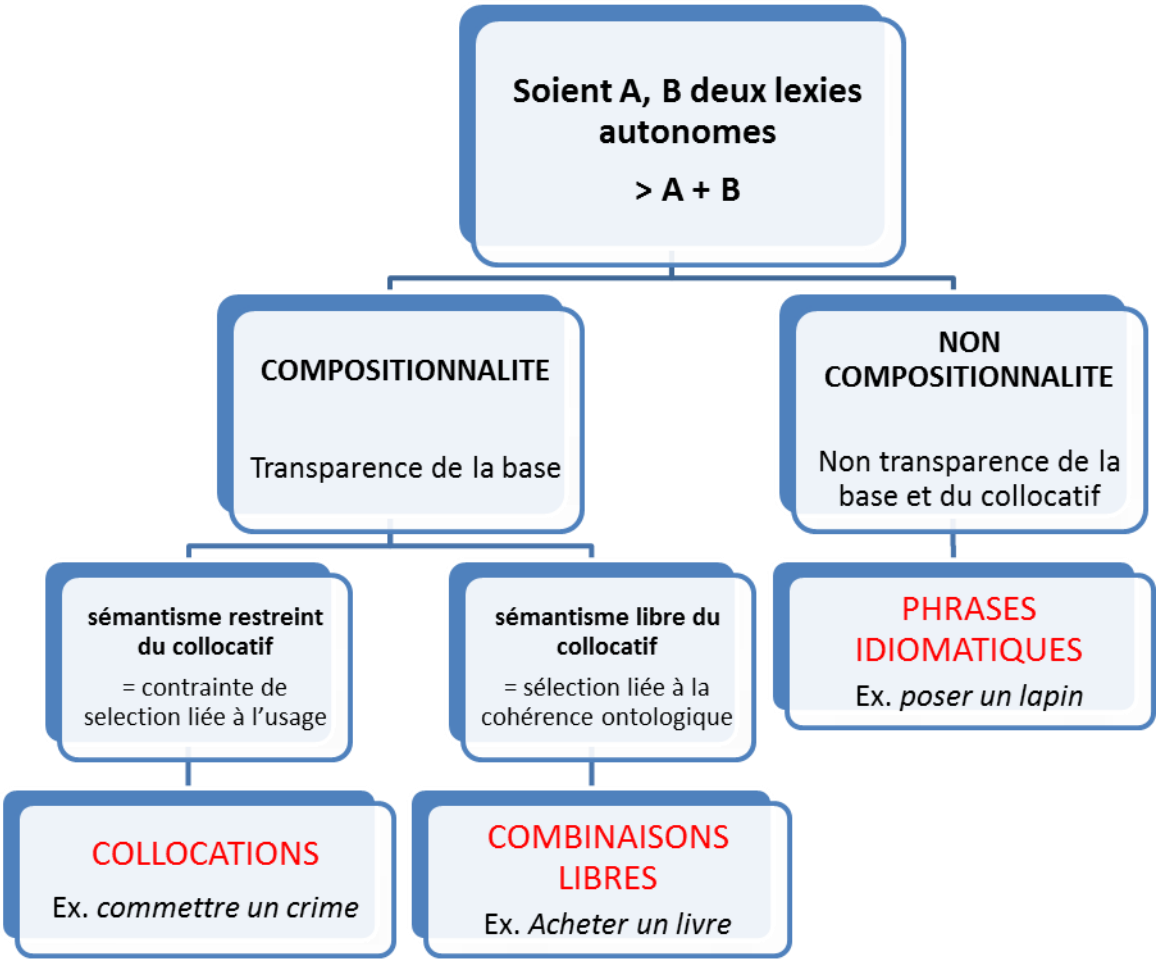
produit, en association avec *décision,* le sens de « décider ». Le nom exprime donc quasi totalement le sens de l'association.

Comme l'explique Nesselhauf (2005), il y a un certain désaccord dans la façon d'entendre ces constructions : certains considèrent seulement des constructions incluant des verbes tels que *faire, prendre, avoir* qui ont un article indéfini qui précède un nom d'événement qui a la même forme que le verbe dont l'expression est synonymique ; d'autres acceptent aussi des verbes tels que *run* dans *run a risk* (*prendre/courir un risque*) ; d'autres encore considèrent également des associations verbe-adjectif du type *be critical* (*être critique*). Mais tous s'accordent sur le fait que les constructions à verbe support sont un type de collocation restreinte.

2) Affinités lexicales univoques

Les affinités lexicales univoques (du collocatif par rapport à la base) sont des associations telles que *nez aquilin*, *l'âne brait* et *année bissextile* qui représentent des cas limites, et qui sont considérés par Grossmann et Tutin (2003) comme des collocations proches des expressions libres. Il n'y a que l´âne qui peut braire : *braire*, *aquilin* et *bissextile* n'existent qu'en association avec leurs bases. Etant donné que le collocatif ne se définit qu'en relation avec sa base et que l'association est compositionnelle et motivée, on pourrait affirmer qu'il s'agit d'une véritable expression libre. Cependant, les affinités lexicales univoques sont soumises à des contraintes agissant sur la commutabilité du collocatif.

En conclusion, le schéma à la page suivante pourra nous aider à différer entre expressions libres, expressions idiomatiques et collocations. La condition de départ est que A et B soient deux lexies autonomes, ce qui exclut les unités conceptuelles telles que les noms composés.




Après cette révision des critères adoptés pour définir les collocations, et après avoir formulé une définition de la collocation qui puisse être mise au service de la recherche développée dans le présent travail, nous présenterons dans le chapitre suivant le corpus ainsi que l'outil développé pour l'extraction des collocations fondamentales.

# CHAPITRE 5
# Le corpus et l'outil d'extraction

Dans ce chapitre, nous expliquerons, tout d'abord, quels critères ont été adoptés pour le choix des mots pivots à partir desquels repérer les collocations fondamentales (au paragraphe 5.1). Ensuite, nous décrirons le corpus choisi et l'outil d'extraction utilisé (respectivement aux paragraphes 5.2 et 5.3). Nous terminerons ce chapitre avec une description des mesures statistiques utilisées (au paragraphe 5.4).

## 5.1 Le choix des mots pivots

Le choix des mots pivots est à l'origine de la constitution de l'échantillon des collocations fondamentales. Le choix de repérer les collocations à partir de leurs bases est opéré dans la plupart des ouvrages lexicographiques, étant donné que, comme l'explique Hausmann (1979), l'utilisateur d'un dictionnaire cherche son collocatif à partir d'une base connue.

Le groupe de mots pivots qui représentent la « tête » ou « base » des collocations est constitué de dix substantifs : *colloque, conférence, congrès, conversation, débat, fête, interview, rencontre, réunion, séminaire.* Ces dix substantifs représentent des «événements sociaux » selon le *Thésaurus Larousse* de Péchoin (1991), et ils sont fondamentaux, c'est-à-dire fréquents ou disponibles selon le *Dictionnaire Fondamental de la langue française* de Gougenheim (1971). Dans ce qui suit, nous expliquerons plus en détail les raisons de notre choix.

1) Ils sont des substantifs.

Le substantif représente généralement (voir Grossmann, Tutin, 2003 pour des exceptions) la tête lexicale dans le couple base-collocatif pour de nombreuses raisons. En premier lieu, le substantif a un caractère référentiel et dénotatif plus précis que les adjectifs et les verbes (au moins dans la langue générale). En outre,

comme le montrent les études en psycholinguistique sur l'accès au lexique mental, les noms sont accédés plus rapidement que les verbes (voir par exemple Ellis, 1997). En second lieu, comme l'explique Lo Cascio (2000), dans les phénomènes combinatoires du lexique, le substantif joue un rôle principal en tant que tête syntaxique et sémantique du syntagme. C'est généralement à partir du substantif que l'on choisit le verbe ou l'adjectif à combiner et, pour cette raison, toute théorie didactique pour l'apprentissage devrait partir du substantif. Enfin, sur un plan méthodologique, l'extraction des substantifs d'un corpus est beaucoup plus simple que l'extraction des verbes : la variation morphologique du substantif est plus petite que celle du verbe, car elle est restreinte au nombre, au genre et éventuellement aux procès de dérivation ou de composition. De toute façon, comme le souligne Lo Cascio (2000), le verbe joue un rôle tout aussi important, puisque l'utilisateur d'un dictionnaire recherche une collocation non seulement à partir des substantifs (entités), mais aussi des verbes (opérations). Les autres catégories telles que les adjectifs, les adverbes et les quantifieurs sont moins significatives. D'autres catégories telles que les prépositions, les articles ou les pronoms ne sont, quant à elles, pas du tout pertinentes en tant que bases. Leur valeur est purement syntaxique, ayant une importance sémantique seulement en tant que collocatifs mémorisés comme compléments de la base.

2) Ils représentent des « événements sociaux » d'après le *Thésaurus* de Péchoin (1991).

Dans ce travail on considère comme étant un « événement social » toute occasion de rencontre avec d'autres personnes[47] ayant lieu dans un processus temporel avec un début et une fin. Le choix d'un domaine sémantique spécifique se justifie par différentes raisons. En premier lieu, l'analyse de toutes les entrées (quelques milliers de mots) de la liste de référence du vocabulaire fondamental du français (Gougenheim, 1970) demanderait un travail énorme et équivaudrait à la rédaction d'un ouvrage lexicographique. En outre, le choix d'une base correspondante à une catégorie grammaticale sans un groupement dans un domaine sémantique aurait

[47] Des mots comme *voyage*, *promenade* et similaires sont par exemple exclus de notre échantillon car ils n'impliquent pas forcement l'échange avec d'autres personnes.

constitué un inventaire peu homogène et peu exploitable (les relations de synonymie sont mises en évidence si on choisit des mots sémantiquement proches). De plus, l'analyse des substantifs qui appartiennent à un domaine sémantique précis présente de l'intérêt pour les études de didactique, perspective toujours présente dans notre travail. Enfin, l'étude d'un domaine sémantique spécifique dévoile les aspects culturels privilégiés dans une langue, et par la comparaison interlinguistique elle montre l'importance de l'idiosyncrasie linguistique et les nuances de sens liées au passage d'une langue à l'autre. Pour conclure, nous précisons que l'échantillon que nous allons constituer n'a pas la prétention de représenter de façon exhaustive tous les « événements sociaux » ni de constituer un mini-lexique combinatoire, mais vise à constituer un échantillon plus ou moins de base, essentiel, représentatif du domaine en question.

3) Ils sont fondamentaux d'après le *Dictionnaire fondamental de la langue française* de Gougenheim (1971).

Nous avons choisi des mots fondamentaux comme bases des collocations fondamentales pour leur nombre élevé d'occurrences ou du fait de leur haute disponibilité. Sur cette prémisse, « + » désignant un mot fondamental et « – » un mot non fondamental, les collocations fondamentales ne seront pas, à notre avis, de type « – – ». En effet, une collocation composée de deux mots non fondamentaux aboutit le plus souvent à un domaine de discours spécialisé et n'est pas représentative du langage fondamental, et nous avons donc *a priori* décidé de l'exclure de notre inventaire (on peut comparer par exemple *recevoir le courrier*, du type « + + », et *accusé de réception*, du type « - - »).

### *5.1.1 Le Thésaurus de Péchoin et le Dictionnaire fondamental de Gougenheim*

De quelles sources les substantifs pivots fondamentaux ont-ils été extraits ? Quelles sont les ressources disponibles à fin de vérification de leur caractère fondamental ? Pour l'extraction, nous avons consulté le *Thésaurus Larousse* de Péchoin (1991) en l'explorant à partir des articles « Sociabilité », « Compagnie » et « Relation ». Nous n'avons gardé que les mots fondamentaux parmi ceux extraits du *Thésaurus*, nous avons donc consulté le *Dictionnaire Fondamental* de Gougenheim (1971).

Le *Thésaurus* permet « d'explorer à partir d'une idée l'univers des mots qui s'y rattachent » et « de trouver des idées à partir des mots liés à une notion » (1991, p. v). Le sommaire présente les trois grandes sections du *Thésaurus* : « Le monde », « L'homme », « La société », chacune d'entre elles divisée en sous-groupes. Nous avons choisi, dans la section « La société », la partie « Les relations sociales », la sous-catégorie « Rapports entre personnes», et l'étiquette sémantique « Relations ». Chaque étiquette contient les articles qui correspondent à la notion. Nous avons choisi, dans « Relations », les articles 581 « Sociabilité » et 583 « Compagnie ». Le *Thésaurus* est également accessible par l'index. Nous avons également décidé de consulter l'article 13 « Relation » (dans la section « L'homme ») pour la pertinence avec les « événements sociaux », puisque d'ailleurs les articles « Sociabilité », « Compagnie » et « Relation » renvoient souvent l'un à l'autre (dans chaque article, des renvois numérotés invitent le lecteur à consulter des notions similaires). Après avoir consulté les articles, nous avons finalement obtenu la liste des mots qui s'y rattachent[48] : de cette liste nous avons choisi les dix mots pivots constituant le socle de la présente recherche. Etant donné que le groupe de mots concernant les relations sociales est assez hétérogène (incluant des « actants sociaux » comme *ami*, *camarade*, etc., des endroits sociaux comme *collège*, etc.), nous avons concentré notre attention sur les « événements sociaux».

Enfin, nous avons décidé de garder seulement les mots fondamentaux, c'est-à-dire les mots qui se trouvent dans le *Dictionnaire Fondamental* de Gougenheim (1971)[49]. Ce dernier est principalement consulté pour garder seulement les mots fondamentaux parmi ceux extraits du *Thésaurus*. Nous aurions pu choisir d'autres listes de fréquence existantes, mais un choix s'imposait, et le *Dictionnaire fondamental* dans sa version revue et augmentée se présentait comme l'ouvrage le plus significatif. De plus, cet ouvrage explique et distingue les différents sens des mots, en privilégiant le sens le plus usuel.

[48] L'article se divise en paragraphes (comme dans une entrée lexicographique) correspondant aux familles de sens pour la notion traitée. Les paragraphes sont ordonnés par catégorie grammaticale.

[49] Nous gardons *colloque*, *séminaire* et *interview*, bien qu'ils ne se trouvent pas aussi dans le dictionnaire de Gougenheim car il s'agit de mots plus importants aujourd'hui qu'à l'époque de la constitution du *Dictionnaire Fondamental* et car ils sont très proches de *conférence* et *congrès*. Nous incluons *fête* à la place de *party* (repéré dans Péchoin, 1991) car *fête* est plus commun et est présent aussi dans la liste de Gougenheim (1971).

Le parcours qui mène du *Français Elémentaire* (publié pour la première fois en 1954) au *Dictionnaire Fondamental* (dernière édition de 1971) a été assez complexe, comme nous l'avons vu dans le chapitre 2. La dernière édition se présente en forme de dictionnaire : elle utilise un système de « définissants » pour définir les mots, distingue les différentes acceptions et enregistre les expressions et les proverbes.

## 5.2 Le corpus

Dans cette section, nous introduisons le concept de corpus en linguistique, avant de décrire le corpus utilisé dans la présente étude[50].

### *5.2.1 Introduction*

Le corpus est un recueil de textes organisé selon des principes spécifiques et constituant un échantillon de la langue qu'il représente. Une définition très célèbre est celle offerte par Sinclair (1996), l'un des fondateurs de la linguistique de corpus :

« A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language » (p. 4).[51]

S'il existe un certain accord sur la définition générique de corpus, on rencontre en revanche un désaccord sur les traits qui le caractérisent, puisque certains auteurs appellent corpus seulement une collection de données langagières où les conditions de représentativité[52] et d'échantillonnage[53] sont totalement remplies.

[50] Pour l'état de l'art de la linguistique de corpus et la compréhension des différents types de corpus existants (monolingues, bilingues ou multilingues, comparables ou parallèles -alignés ou non-, bruts ou traités (annotés, étiquetés et lemmatisés, arborés, marqués par tag sémantique, pragmatique, etc..), nous renvoyons à ouvrages de référence tels que McEnery et al. (2006), Sinclair (1991).

[51] « Une collection de données langagières sélectionnées et ordonnées selon des critères linguistiques précis pour être utilisées comme échantillon de la langue » (notre traduction).

[52] Le corpus représente un échantillon d'une population, c'est-à-dire de la langue qu'il se propose de représenter : par exemple des articles scientifiques, des reviews et des livres de psychologie représentant la langue de spécialité de ce domaine spécifique.

[53] L'échantillonnage permet d'équilibrer le corpus en termes de sections le composant pour ne pas causer des biais en termes de genres textuels.

Comme l'explique Sinclair (1991, p. 4), l'avènement de la linguistique de corpus a complètement révolutionné l'étude de la langue, en faisant émerger le contraste existant entre l'introspection des locuteurs et les usages réels. Selon l'auteur, l'intuition humaine sur le langage est très spécifique et ne correspond pas à la façon dont les locuteurs natifs utilisent la langue.

McEnery et al. (2006, p. 3) considèrent la linguistique de corpus plus comme une méthodologie que comme une discipline, puisqu'elle explique comment utiliser les corpus pour conduire des analyses linguistiques dans différents domaines tels que la phonétique, la syntaxe, la sémantique ou la pragmatique. En outre, les avantages de l'analyse informatisée des corpus par rapport à l'analyse traditionnelle sont nombreux. Il est possible, grâce aux nouvelles technologies, de produire des analyses beaucoup plus fines qu'à l'époque de l'analyse manuelle. Voici les principaux avantages apportés par la linguistique de corpus, qui permet de :

- traiter de volumineuses banques de données ;
- stocker, manipuler et analyser automatiquement les données ;
- mettre en évidence des phénomènes qui ressortent de l'usage.

Ce dernier point est aujourd'hui très bien défendu par ceux qui sont convaincus que la linguistique de corpus présente un miroir de la routine langagière (et donc de la façon dont le langage est structuré). Selon cette optique, la linguistique de corpus permet de mettre en évidence les formules et les paquets linguistiques (les « repeated events » de Stubbs, 2002) dont tous les locuteurs se servent au-delà de la variation individuelle.

Les bénéfices tirés de cette révolution sont nombreux et intéressent différents domaines de recherche, de la lexicographie à la traduction à l'extraction terminologique et à la didactique.

### *5.2.2 Critères de choix du corpus*

Dans notre étude, nous utilisons le corpus *frWaC* (Baroni et al., 2010), un corpus de plus d'un milliard de mots constitué de textes issus du Web à l'aide d'un crawler filtrant le contenu d'Internet sur la base d'une liste de mots clés de fréquence haute et moyenne. L'outil d'exploration et d'extraction du corpus est constitué de divers

scripts Perl réalisés par Olivier Kraif (2011), qui travaille sur le développement d'outils pour l'exploration des corpus depuis plus de 10 ans. Dans ce paragraphe, nous allons expliquer les raisons du choix du corpus *frWaC* dans notre recherche, avant de décrire le corpus en détail au paragraphe suivant.

Un corpus est choisi selon l'objectif de sa propre recherche et le corpus *frWaC* se présente, selon nous, comme un corpus adéquat pour l'étude des collocations fondamentales, en termes d'avantages et de désavantages, par rapport à d'autres corpus. Les deux critères les plus importants qui ont guidé notre choix sont la très grande taille du corpus et sa représentativité de la langue générale :

1) La très grande taille du corpus

Selon Sinclair (1991, p. 99), chaque locuteur a sa propre façon de s'exprimer, il est créatif, parfois recourt à des expédients communicatifs. Par conséquent, l'usage standard de la langue peut être uniquement retrouvé par l'analyse d'un grand recueil de données. En effet, notre intérêt spécifique est le noyau central de la langue car les collocations fondamentales sont en grande partie des unités fréquentes. Le corpus *frWaC* se compose de plus d'un milliard de mots : le fait de disposer d'une vaste banque de données assure la validité des résultats car si la fréquence d'occurrence dépasse un certain seuil statistique, le pourcentage de fiabilité est plus élevé. Au contraire, pour des phénomènes moins fréquents, il existe une plus grande variation et les informations issues de la statistique sont susceptibles d'amener à des conclusions erronées.

2) La représentativité de la langue générale

Les collocations fondamentales étant notre objet d'étude, il était nécessaire d'interroger un corpus représentatif de la langue générale. Habituellement, pour des études sur la langue générale, des corpus de presse tels que les articles du journal *Le Monde* ou des corpus de littérature sont choisis. Ces choix sont à notre avis discutables. A ce propos, quelques précisions sur la représentativité de la langue générale apparaissent nécessaires. La notion de « langue générale » n'est pas complètement saisissable car elle suppose que le corpus qui la représente soit le plus varié possible en termes de genres textuels qui le composent. La plupart des auteurs s'accordent sur ce dernier point. Selon McEnery et al. (2006), les corpus de langue générale représentent la langue dans sa totalité ou dans l'une de ses variétés

(par exemple le *British National Corpus -BNC-* qui représente l'anglais moderne) et doivent donc se composer d'un nombre le plus élevé possible de types de textes. Suivant le même principe, Sinclair (1991, p. 17) affirme qu'un bon corpus de langue générale doit représenter la masse de ses locuteurs et les « [...] repeated patterns at the expense of unique ones » (« les phénomènes fréquents au détriment de ceux qui sont singuliers »). Selon l'auteur, un corpus de langue générale est :

« [...] gathered from a variety of sources so that the individuality of a source is obscured, unless the researcher isolates a particular text »[54] (Sinclair, 1991, p. 17).

En outre, comparé à un corpus de langue de spécialité, il existe, dans un corpus de langue générale, plusieurs mots qui partagent un plus grand nombre de contextes car les différents domaines sémantiques ne sont pas très marqués. Cela signifie que dans un corpus de langue de spécialité les sens sont plus univoques. Au contraire, dans un corpus de langue générale il y a des sens plus centraux et plus fréquents ainsi que d'autres sens périphériques plus ou moins stables (en général les mots les plus fréquents sont aussi les plus polysémiques).

Il apparaît donc évident que constituer un corpus représentatif de la langue générale contemporaine est une tâche très ambitieuse qui demande une grande collection de textes de genres différents. Nous avons considéré le corpus *frWaC* comme représentatif de la langue générale pour deux raisons principales :

a) Bien qu'il ne constitue pas forcément un échantillon équilibré de la population, (étant moins contrôlé qu'un corpus traditionnel car il est issu du Web), il est vaste et hétérogène, et il héberge une population d' « écrivains » très variée en termes d'âge et de niveau d'éducation. Nous remarquons que le fait d'analyser le corpus à travers un outil d'extraction construit selon des paramètres précis, et combinant de différents critères de sélection (fréquence, dispersion, mesures associatives telles que l'IM) limite les désavantages dérivant d'un corpus issu du Web.

[54] « Assemblé à partir d'une grande variété de sources, si bien que l'individualité d'une source donnée est masquée à moins que le chercheur n'isole un texte précis » (notre traduction).

b) Il consiste différents types de texte, des forums à la presse en passant par l'encyclopédie et les romans, etc. (non orientation de genre).

Pour conclure, nous sommes convaincue que le corpus choisi est une ressource adéquate pour notre étude du fait de sa grande taille et de sa représentativité de la langue générale. D'autres « points forts » sont les suivants :

a) il est déjà traité : les textes ont été lemmatisés et annotés morpho-syntaxiquement en utilisant le *TreeTagger* ;
b) il est disponible aussi dans d'autres langues (parmi lesquelles l'italien) en tant que corpus comparable.

### *5.2.3 Le corpus : frWaC*

Comme l'expliquent Baroni et al. (2010), *frWaC* est un corpus très large (plus d'un milliard de mots) extrait du Web par des procédures automatiques. Le corpus, issu de la collaboration entre l'Université de Bologne, la maison *Larousse dictionnaire bilingues* et l'Université de Trente, s'inscrit dans un projet de recherche internationale appelé *WaCky* (*Web as Corpus kool ynitiative*) qui a réuni dans un consortium des chercheurs voulant explorer le Web comme source de données linguistiques (Baroni et al., 2009). L'équipe de chercheurs a produit, à partir de 2005, des corpus issus du Web par *crawling* et annotés par *TreeTagger* : *itWaC* et *deWaC*, respectivement pour l'italien et l'allemand (les plus grandes ressources existantes pour ces langues aujourd'hui selon les auteurs - p. 1 -), *ukWaC* pour l'anglais (le seul corpus annoté issu du Web existant à la connaissance des auteurs - p. 1 -) et *frWaC* pour le français. Les corpus sont gratuitement téléchargeables sur le site http://wacky.sslmit.unibo.it/doku.php?id=download sur autorisation des auteurs et envoi d'identifiant et mot de passe.

Le corpus *frWaC* a été exploré, conjointement à la version anglaise *ukWaC*, dans une étude pilote visant à en évaluer la qualité (Baroni et al., 2009). Malgré quelques limites dérivant de la constitution de *frWaC* et *ukWaC* (par exemple le fait que les contenus ne sont pas accessibles), l'étude a pu montrer l'utilité et la comparabilité des deux corpus comme source de données à des fins lexicographiques. L'étude a

passé en revue un dictionnaire anglais-français (seulement dans cette direction et non aussi du français vers l'anglais) dans les étapes suivantes :

- la comparaison, pour un échantillon de mots pivots, entre l'information collocationnelle repérée dans *ukWaC* et celle repérée dans le *British National Corpus ;*
- l'évaluation des collocations extraites de la part de lexicographes compétents ;
- la traduction des collocations choisies pour inclusion dans le dictionnaire en français à l'aide du corpus *frWaC* ;
- l'évaluation des traductions produites par un traducteur professionnel et natif du français.

Les auteurs expliquent que le Web est, dans certains cas, la ressource la plus adéquate pour le repérage de données linguistiques : c'est le cas pour les analyses qui nécessitent un corpus de grande taille, pour l'étude de phénomènes récents qui ne sont pas enregistrés dans les corpus traditionnels, ou encore pour l'étude de sous-domaines de spécialité ou de langues minoritaires. Cependant, la constitution de corpus parallèles issus du Web (un format de corpus important pour des recherches interlinguistiques mais qui n'est pas souvent accessible facilement et gratuitement) se présente encore comme une tache très problématique. Pour cette raison, les auteurs proposent d'exploiter le Web pour la constitution de corpus comparables plutôt que de corpus parallèles. Les corpus *ukWaC* et *frWaC* ne sont pas envisagés comme des corpus comparables, mais comme des corpus de référence similaires à des corpus traditionnels importants tels que le *British National Corpus* (en termes de variété de typologies textuelles et de sujets traités). Cependant, les auteurs montrent qu'ils peuvent être utilisés comme des corpus comparables, grâce à leur grande taille et au fait d'avoir été constitués avec la même procédure. Le corpus *frWaC* (et son contemporain *ukWac*) a été constitué selon la même procédure que celle suivie pour la constitution de *deWaC*, la version allemande, et *itWaC*, la version italienne (ces derniers étant les premières ressources produites dans le cadre du projet *WaCky*).

La constitution de *frWaC* s'est déroulée en deux étapes fondamentales, la sélection des mots clé (« seeds » en anglais) et le *crawling,* d'une part ; et d'autre part, le nettoyage et l'annotation. Détaillons ces deux phrases :

1) La sélection des mots clé (« seeds ») et le *crawling*

Le but des auteurs était de repérer des textes représentatifs de la langue française générale (et non du français du Web) et de différents genres. Le corpus inclut des textes pré-édités et en format électronique sur le Web (des sermons aux recettes, en passant par les manuels techniques et les récits, et idéalement aussi des transcriptions de la langue orale) ainsi que des textes typiques du Web (pages personnelles, blogs, messages de forums). Nous avons procédé à la sélection des mots-clés et au *crawling* de la façon suivante :

a) 1 800 mots clés pleins (« seeds ») ont été choisis comme input de la sélection des contenus, pour s'assurer que les URL extraits de Google correspondaient aux critères recherchés. Selon les auteurs, dans les recherches automatiques de Google, l'insertion de mots clés typiques de l'écrit produit comme résultat des documents appartenant à la « sphère publique » (par exemple des textes académiques), tandis que l'insertion de mots clés de base produit comme résultat plutôt des pages appartenant à la « sphère personnelle » (par ex. des blogs). Pour cette raison, deux types de recherches ont été exécutées : la première a utilisé 1 000 paires de mots pleins de fréquence moyenne extraits de textes du journal *Le Monde Diplomatique* publiés entre 1980 et 2000 ; la seconde a utilisé 769 paires de mots choisis à partir d'une liste de vocabulaire destinée à des enfants de 8 à 10 ans (http://o.bacquet.free.fr/index.html).
b) Les URL ainsi sélectionnés ont été soumis au *crawler Heritrix* (http://crawler.archive.org/), limitant la recherche au domaine .fr et aux URL html (des extensions telles que .wav ou .jpg étant exclues par exemple).

2) Le nettoyage et l'annotation

Les documents extraits à l'aide du *crawler* ont été nettoyés avant l'annotation. Différentes opérations ont été exécutées.

a) Les auteurs n'ont gardé que les documents de type « MIME »[55] text/html entre 5 et 200 kb pour éviter le silence dérivant de documents trop petits et vice-versa le bruit de documents trop longs).

[55] Le type « MIME » (« Multipurpose Internet Mail Extensions ») est un standard utilisé pour typer les documents attachés à un courrier ainsi que pour typer les documents transférés par le protocole HTTP (http://www.slideshare.net/metadonnee/questce-que-le-type-mime).

b) Les doublons ont été éliminés, selon une approche qui, tenant compte de l'énorme taille du corpus, privilégie la « précision » (proportion de documents pertinents trouvés parmi tous les documents rassemblés par la recherche) plutôt que le « rappel » (proportion de documents pertinents trouvés parmi tous les documents pertinents dans le corpus).

c) Les documents ont été privés de leur code (html et javascript) et des formules standards liées à la navigation et à la mise en page (« boilerplate text » en anglais).

d) Des filtres linguistiques assez simples ont été utilisés afin d'exclure les textes écrits dans d'autres langues, les pages pornographiques générées automatiquement ainsi que les pages ayant totalement ou partiellement le même contenu (« near-duplicate documents »).

e) Les textes ont été lemmatisés et annotés morpho-syntaxiquement avec *TreeTagger*, un outil développé par Helmut Schmid à l'*ICLUS*, *Institute for Computational Linguistics* de l'Université de Stuttgart (http://www.ims.uni-stuttgart.de/projekte/corplex/*TreeTagger*/). Ces dernières actions ont enrichi le corpus d'informations linguistiques supplémentaires.

A la fin de toutes ces phases, *frWaC* avait les caractéristiques suivantes :

- 1 769 paires de mots clés
- 6 166 URLS
- 470 GB de données crawlées
- 9 GB de donnés sélectionnées
- 2.2 M de documents obtenus
- 27 GB des données après annotation
- 1 027 246 563 de tokens au total
- 3 987 891 de types au total

Le corpus *frWaC* a été gratuitement téléchargé du site http://wacky.sslmit.unibo.it/doku.php?id=download sur autorisation des auteurs et envoi d'identifiant et de mot de passe. Il se compose de 18 fichiers html, analysés et traités à l'aide de l'outil d'extraction décrit dans la section suivante.

## 5.3 L'outil d'extraction

### *5.3.1 Les paramètres généraux*

Le corpus *frWaC* a été exploré à l'aide de scripts Perl développés par Olivier Kraif (2011). L'exécution des scripts a produit en sortie deux fichiers au format .txt qui représentent l'output de l´interrogation :

1) une liste de fréquence accompagnée de mesures statistiques

Elle affiche la fréquence des cooccurrences, leur dispersion et leur degré d'association selon des mesures spécifiques telles que l'information mutuelle.

2) un concordancier

Le concordancier affiche les lignes de concordances pour chaque cooccurrence du mot pivot : les sources Web d'où elles sont extraites, c'est-à-dire l'URL, sont aussi indiquées. Nous lisons les concordances pour dériver les patterns sémantiques et syntaxiques caractérisant les cooccurrences extraites.

Les paramètres des scripts Perl qui ont été implémentés pour l'extraction des collocations fondamentales sont les suivants :

```
my $pivot="fête"
# Le mot pivot est indiqué dans sa forme base.
my $cat="NOM"
# Si une valeur est fixée on tient compte de la catégorie grammaticale.
my $collocatif=""
# Si une valeur est fixée, on obtient seulement les contextes où $pivot cooccurre avec ce collocatif.
my $recordPosition=0
# Si la valeur 1est fixée, la position du collocatif (avant ou après) est prise en compte.
my $lemme=1
# Si la valeur 1est fixée, les lemmes sont pris en compte.
```

```
my $caseSentitive=1
# Si la valeur 1est fixée, la casse est prise en compte.
my $collocSpan=4
# La taille de l'empan est de 4 cooccurrents à gauche et 4 à droite.
my $kwicSpan=50
# 50 est le nombre de caractères à afficher à gauche et à droite du pivot dans la sortie KWIC.
my $nbColloc=2 000
# Si une valeur est fixée, on établit le nombre de collocatifs à retenir parmi les plus fréquents
my $limit=0
# Si une valeur est fixé, on établit le nombre maximal de phrases traitées. Si la valeur est =0, on ne pose pas de limite.
```

Considérons-les en détail.

**my** $pivot="fête"

Ce paramètre établit le mot pivot de la cooccurrence. A chaque exécution du script nous insérons un des dix mots pivots précédemment choisis, dans leur forme base.

**my** $cat="NOM"

Ce paramètre utilise les tags d'étiquetage du corpus et permet de désambigüiser des cas d'homographie entre deux mots qui correspondent à deux catégories grammaticales différentes, par exemple *porte* (nom singulier) qui est différent de *porte* (verbe, 3ème personne du singulier). Nous opérons ici une désambigüisation grammaticale pour éviter que les fréquences indiquées ne soient fausses à cause de l'association des classes grammaticales homographes. L'ambiguïté syntaxique est un phénomène très répandu dans les langues, d'où la nécessité de ces procédés discriminatoires.

**my** $collocatif="faire"

Ce paramètre permet de conduire des statistiques concernant les formes spécifiques du collocatif. Si une valeur est fixée, par exemple le collocatif *faire* en cooccurrence avec le mot pivot *fête*, le concordancier n'affichera que les contextes où *fête* cooccurre avec *faire*. En association avec $lemme=1 (le paramètre qui prend en compte les formes dans lesquelles le mot pivot se présente), ce paramètre permet d'afficher des statistiques concernant la fréquence de cooccurrence des différentes formes du collocatif (*fais, fait, faisons,* etc.). En outre, ce paramètre sert à les regrouper pour le lemme du pivot ainsi que pour chacune de ses formes, comme dans l'exemple suivant (les chiffres sont donnés à titre indicatif) :

1. Répartition des formes fléchies des collocatifs du pivot *fête*
Collocatifs : *faire* (x4) ; *fait* (x2) ; *fais* (x1) ; *font* (x1)

2. Répartition des formes fléchies des collocatifs pour les différentes formes fléchies du pivot *fête*
Forme *fête* => 7 occurrences dans la collocation : *fête/faire* (x4) ; *fête/fais* (x1) ; *fête/fait* (x1) ; *fête/font* (x1)
Forme *fêtes* => 1 occurrences dans la collocation : *fêtes/fait* (x1)

Ce paramètre permet d'étudier l'impact des différentes réalisations syntaxiques et de comprendre quel pattern morphosyntaxique est plus fréquent pour une cooccurrence donnée. Nous nous intéressons à la fois à la fréquence globale d'une collocation prise sous forme canonique et aux fréquences de ses différentes réalisations morphosyntaxiques

**my** $recordPosition=1

Ce paramètre, si la valeur 1 est fixée, prend en compte la position du collocatif et différencie dans le calcul statistique les formes selon leur position par rapport à leur base (avant ou après). Il est utilisé si on remarque que le changement de position du collocatif provoque un changement de sens ou un changement stylistique significatif. Par exemple, la dislocation à gauche, typique de la langue orale, change l'ordre des termes de la phrase par l'antéposition du complément (notre mot pivot). Ce procédé met en valeur une information et crée aussi un effet de surprise chez l'interlocuteur : *cette amende, je l'ai déjà payée !* apporte un effet stylistique particulier par rapport à *j'ai déjà payé cette amende.*

**my** $lemme=1

Ce paramètre permet de repérer non seulement le lemme (le mot vedette), mais aussi ses différentes formes. Si la valeur 1 est fixée, les statistiques sont calculées en tenant compte des différentes réalisations de *faire* telles que *faisons, fait,* etc. Si la valeur 0 est fixée, on ne recherche que la forme orthographique correspondant à $pivot. L'analyse des formes est défendue par des auteurs comme Stubbs (2002), selon qui l'analyse des corpus met en évidence le fait que l'unité de sens est plus large que le lemme. En effet, les dictionnaires utilisent les lemmes comme mots vedettes pour une exigence de la pratique lexicographique, mais les unités du texte sont les formes.

**my** $caseSentitive=1

Ce paramètre est similaire au paramètre $cat, à la différence qu'il opère une désambigüisation au niveau de l'orthographe, plus spécifiquement sur la capitalisation des mots. Si la valeur 1 est fixée, toutes les formes dont la graphie est identique à celle indiquée pour $pivot seront sélectionnées. En résumé, ce paramètre prend en compte uniquement l'orthographe exacte du mot. Si le pivot est *porte*, alors *Porte* sera considérée comme une autre forme et ne sera pas comptabilisée

(sauf si elle est en majuscule au début de la phrase). Si la valeur 0 est fixée, alors aucune différence entre majuscule et minuscule ne sera faite (les mots *PORTE*, *porte* et *Porte* seront considérés comme représentants du même pivot). Il est évident que certaines occurrences « échappent » au calcul car elles sont en majuscules. Un prétraitement du corpus pourrait bien sûr être utile pour normaliser finement les critères de recherche des mots, mais étant donné que le corpus est très grand, une légère augmentation du silence n'est pas vraiment gênante.

**my** $collocSpan=4

Ce paramètre établit la taille de l'empan, à droite et à gauche du mot pivot, sur le modèle de Sinclair (1991), selon qui cette largeur contextuelle permet de repérer les collocatifs significatifs. Par la valeur 4 nous établissons que notre concordance affiche 4 collocatifs à droite de la base et 4 collocatifs à gauche de la base. Nous précisons que la ponctuation est comptée dans l'empan. Parfois des contextes plus larges sont nécessaires pour que le collocatif soit repéré, sinon le risque est de perdre des résultats significatifs. Dans la phrase suivante, par exemple, le collocatif *mettre en vente* en association avec le pivot *appartement* est perdu à cause d'une erreur dans le comptage, car il se trouve en cinquième position à droite (+5) dans la phrase suivante : *J'ai décidé d'acheter l'appartement que Jacques vient de mettre en vente, celui de la rue Jacques.* Le début de l'empan est *décidé* (le quatrième mot avant le pivot, -4) tandis que la préposition *de* est la fin de l'empan (le quatrième mot après le pivot, +4). Cependant, notre corpus est très grand et la perte de quelques résultats significatifs est secondaire : généralement, nous préférons éviter le bruit dérivant de concordances trop longues.

**my** $kwicSpan=50

Ce paramètre établit le nombre de caractères à afficher à gauche et à droite du pivot dans la sortie *KWIC* (*key word in context*), le format le plus commun utilisé pour afficher la concordance ayant le mot pivot au centre et le contexte autour. Nous décidons d'accéder à un contexte de 50 caractères (parfois les 9 mots de

l'empan de cooccurrence ne sont pas suffisants à comprendre le contenu d'une phrase). La concordance est définie par Sinclair (1991, p. 32) comme un index, c'est-à-dire une collection des occurrences d'une forme lexicale, chacune dans son contexte. En résumé, si l'empan est la fenêtre graphique de 4 mots à gauche et de 4 mots à droite choisie comme contexte minimal pour repérer des collocatifs significatifs, alors la concordance est le contexte d'occurrence le plus large (de 50 mots) affiché pour le mot pivot repéré. La concordance permet de visualiser le texte d'occurrence du mot pivot (par ordre alphabétique, par collocatif ou par n'importe quel autre mot) et d'étudier ses variations. Il serait idéal de pouvoir analyser chaque concordance pour analyser en détail le sens en contexte d'un mot, mais il s'agit d'un travail très minutieux et impossible pour de gros résultats, d'où la pratique d'ordonner les concordances par ordre alphabétique ou par collocatif, selon la nécessité propre à l'analyse considérée.

**my** $nbColloc=2 000

Ce paramètre établit le nombre de collocatifs retenus parmi les plus fréquents. L'exécution de tests sur les 2 000 collocatifs extraits a suggéré que ce seuil statistique est largement suffisant pour qu'aucune cooccurrence significative n'échappe au calcul et pour couvrir la marge de bruit (les signes de ponctuation, les catégories vides, les fausses cooccurrences, etc.). Il se réduit après le premier filtrage automatique (l'élimination des catégories vides) à environ 1 500 collocatifs et diminue ultérieurement après le filtrage manuel des sorties (l'élimination d'entités nommées, bogues, etc. Nous renvoyons au chapitre suivant pour approfondissement). Pour le mot *fête,* par exemple, les tests exécutés ont montré qu'au fur et à mesure que la fréquence se réduit, le nombre des cooccurrences considérées comme étant acceptables diminue progressivement. Considérons les 1 100 cooccurrents qui restent après le premier filtrage automatique pour *fête* :

- tranche de 0 à 100 > 73 cooccurrents acceptables
- tranche de 100 à 200 > 60 cooccurrents acceptables
- tranche de 200 à 300 > 44 cooccurrents acceptables
- tranche de 300 à 400 > 37 cooccurrents acceptables

- tranche de 400 à 500 > 33 cooccurrents acceptables
- tranche de 500 à 600 > 33 cooccurrents acceptables
- tranche de 600 à 700 > 18 cooccurrents acceptables
- tranche de 700 à 800 > 19 cooccurrents acceptables
- tranche de 800 à 900 > 19 cooccurrents acceptables
- tranche de 900 à 1 000 > 15 cooccurrents acceptables
- tranche de 1 000 à 1 100 > 16 cooccurrents acceptables[56]

**my** $limit=0

Ce paramètre établit le nombre maximal de phrases traitées et permet donc d'interrompre le script dès que nous avons un certain nombre de contextes (par exemple **my** $limit=1000 permet d'avoir 1 000 contextes). Si une valeur est fixée, la recherche sera donc moins exhaustive, puisqu'une partie du corpus ne sera pas explorée. Dans cette étude, aucune valeur limite n'a été fixée.

En conclusion, tous ces paramètres ont permis de réduire le bruit et d'affiner la recherche selon des critères établis *a priori*. En outre, au moment de l'analyse des résultats nous nous sommes rendu compte que de petits ajustements étaient encore nécessaires : par exemple éliminer parmi les résultats affichés les signes de ponctuation, améliorer la lemmatisation et procéder à la normalisation. Au départ, la lemmatisation ne fonctionnait pas car la normalisation des majuscules (par ex. en début de phrase) et des majuscules emphatiques (*Ministère* et *ministère*) n'était pas correctement exécutée par *TreeTagger,* et des mêmes mots étaient enregistrés dans deux catégories distinctes (et avec deux fréquences distinctes). Il a donc fallu ajouter un deuxième mécanisme de lemmatisation forcée.

[56] Pour le déroulement de ces tests on a mis en place une évaluation subjective des résultats et calculé le nombre de cooccurrents acceptables tous les 100 résultats. La valeur 1 a été attribuée à toute cooccurrence correcte au niveau conceptuel et sémantico-syntaxique : en effet l'extraction des collocations au sens plus strict est renvoyé au moment du traitement des sorties (voir chapitre 6), tandis qu'à cette étape-ci le but est simplement d'établir que tous les candidats à une relation de cooccurrence soient inclus dans le seuil établi (qu'il s'agisse de combinaisons libres ou de locutions idiomatiques). Par exemple pour le mot pivot *fête*, nous considérons comme correct *sens de la fête*, plutôt libre, ainsi que *fête foraine*, plutôt restreinte. On reporte la phase de filtrage linguistique au moment où l'échantillon devient plus réduit.

## 5.4 Les mesures statistiques

Comme nous l'avons expliqué, l'exécution des scripts Perl génère en sortie un tableau affichant des mesures statistiques (et un concordancier pour l'étude du contexte de cooccurrence des associations). Considérons par exemple le mot pivot *rencontre* et ses cinq premiers cooccurrents extraits. L'exécution des scripts Perl sur l'ensemble des 18 fichiers .xml du corpus génère un document affichant les statistiques suivantes (**tableau I**) :

**Tableau I** – Pivot *rencontre* : extrait de la sortie affichant les statistiques

| **Collocatif** | **f** | **f1** | **f2** | **N** | **logLike** | **t-score** | **z-score** | **IM** | **domaines** |
|---|---|---|---|---|---|---|---|---|---|
| être | 18786 | 227529 | 24391530 | 163381196 | 9282,108 | -82,376 | -110,77 | -0,59232 | 6599 |
| avoir | 14079 | 227529 | 14136196 | 163381196 | 1927,412 | -39,9651 | -47,2584 | -0,33525 | 4697 |
| lieu | 6839 | 227529 | 592657 | 163381196 | 17118,54 | 209,324 | 72,71799 | 2,114589 | 3495 |
| organiser | 6064 | 227529 | 373032 | 163381196 | 18932,41 | 243,261 | 71,20053 | 2,457269 | 2853 |
| faire | 5958 | 227529 | 4343761 | 163381196 | 1,422598 | -1,17306 | -1,18201 | -0,0152 | 2887 |
| aller | 5175 | 227529 | 1475200 | 163381196 | 3370,762 | 68,84857 | 43,3793 | 0,923854 | 2631 |

La première colonne affiche le cooccurrent du mot pivot lemmatisé (dans sa forme base, canonique), indiqué ici comme « collocatif ». Suivent « f », la fréquence absolue de la coccurrence ; « f1 », la fréquence de la base ; « f2 », la fréquence du collocatif ; « N », le nombre total de tous les couples de cooccurrents dans le corpus, y compris ceux de la base et du collocatif. Suivent les mesures statistiques associatives : l'information mutuelle spécifique[57] indiquée comme « IM », le log-likelihood ratio indiqué comme « loglike »[58], le « t-score »[59] et le « z-score »[60]. Enfin, les « domaines », c'est-à-dire le nombre de sources Web dans lesquelles notre cooccurrence est repérée. Cette mesure nous permet d'avoir une idée, bien qu'un peu imprécise, de la dispersion de nos cooccurrences. Dans ce qui suit, nous expliquerons plus en détail la valeur de trois mesures complémentaires : la fréquence, l'information mutuelle et la dispersion.

### *5.4.1 La fréquence (et notions relatives)*

Nous choisissons d'opérer en termes de fréquence absolue. La fréquence absolue mesure le nombre de fois qu'une cooccurrence est repérée dans le corpus et donne un poids aux associations les plus fréquentes de la production langagière : un locuteur français pourrait se tromper en produisant en anglais *big smoker* sous l'influence de sa langue maternelle, mais l'étude de la fréquence en corpus l'informerait que *heavy smoker* est l'association établie par l'usage, car elle apparaît

[57] Le terme « information mutuelle » recouvre en fait deux mesures distinctes : l'information mutuelle telle qu'elle est définie par la théorie de l'information, et puis l'IM spécifique (*Pointwise Mutual Information*), qui n'est qu'une composante de l'information mutuelle au sens premier, et qui est souvent, dans la littérature, confondue abusivement avec celle-ci. Nous avons désigné l'information mutuelle spécifique toujours par l'abréviation IM.

[58] Le log-likelihood est très similaire à l'information mutuelle issue de la théorie de l'information, car il repère les cooccurrences pour lesquelles l'hypothèse d'indépendance statistique est invraisemblable. Comme l'expliquent Manning et Schütze (1999), cette mesure compare deux hypothèses, l'hypothèse d'indépendance et l'hypothèse de dépendance des deux composants.

[59] Le t-score est très similaire à la fréquence car c'est une mesure qui favorise les cooccurrences les plus fréquentes. Comme l'explique Clear (1993), le t-score tient compte non de la force associative des mots, mais du nombre de fois que les composants sont observés ensemble.

[60] Le z-score est en principe une sorte de compromis entre ces deux tendances opposées et assure que l'association entre les deux mots ne soit pas due au hasard. Ce test tient en compte le fait que les composants s'associent plus fréquemment que le simple hasard.

plus souvent que d'autres expressions synonymiques qui ne correspondent pas à la façon la plus naturelle et « native » d'exprimer ce concept.

Pour calculer la fréquence absolue, Olivier Kraif (2011) applique un modèle « approché » d'empans sans chevauchement, qui découpe chaque phrase en empans fixes et contigus. Si, par exemple, une fenêtre est de largeur 5 (deux mots à gauche et deux mots à droite du mot pivot), dans une phrase qui contient 42 mots nous aurons 8,4 (42/5) empans. Nous faisons l'hypothèse que chaque cooccurrence se présente dans un empan séparé. Nous centrons les empans sur le pivot : les cooccurrences avec le pivot, à l'intérieur de la fenêtre, sont toujours comptées dans le même empan.

Ce qui permet de calculer ainsi :

N=somme de tous les empans

F1=nombre des empans où le mot pivot apparaît

F2=nombre des empans où le collocatif apparaît

F= nombre des empans ou F1 et F2 apparaissent ensemble, ce qui mesure la fréquence absolue de la cooccurrence.

Présentons une description plus technique de ce que nous venons d'expliquer. Si on raisonne en terme de tirages aléatoires (ou « observations ») :

- à l'unité u1, on associe la variable aléatoire binaire U1, désignant la présence (ou l'absence) de u1 lors d'une observation ;
- à l'unité u2, on associe la variable aléatoire binaire U2, désignant la présence (ou l'absence) de u2 lors d'une observation ;
- chacune de ces variables aléatoires est attachée à une loi de probabilité.

Ce que nous voulons mesurer, c'est la corrélation entre les deux lois de probabilités. Le corpus est considéré comme un échantillon, c'est-à-dire une série d'observations individuelles. A partir de cet échantillon, nous résumons les observations sous la forme d'un tableau de contingence qui permet d'estimer les lois de chaque variable prise indépendamment (lois marginales), ainsi que leur loi conjointe. Les mesures de corrélations sont toutes basées sur ce tableau de contingence, synthétisé par les quatre valeurs suivantes :

PP : nb d'observations avec u1 et u2

PA : nb d'observations avec u1 sans u2

AP : nb d'observations sans u1 avec u2

AA : nb d'observations sans u1 ni u2

Nous pouvons également synthétiser le tableau de contingence avec les valeurs suivantes :

PP : nb d'observations avec u1 et u2

TP1 : nb d'observations avec u1 (TP1=PP+PA)

TP2 : nb d'observations avec u2 (TP2=PP+AP)

N : nb total d'observations (N = PP+PA+AP+AA = TP1+TP2-PP+AA)

Ici, nous considérons qu'un empan de largeur « l » correspond à une observation, et nous faisons les hypothèses suivantes (et c'est là que se situe l'approximation) :

- toutes les formes (y compris le pivot) ne peuvent avoir plus d'une occurrence dans un même empan (si l'empan n'est pas trop large, ce qui est vrai la plupart du temps). Grâce à cette approximation, nous assimilons TP1 et TP2, le nombre d'empans où respectivement deux unités u1 et u2 apparaissent, aux fréquences d'occurrences f1 et f2 :

TP1 ≈ f1

TP2 ≈ f2

- les empans sont centrés sur les pivots (ce qui n'est vrai qu'en moyenne) : ce qui permet de considérer que tous les mots à -2 et à +2 du pivot sont des cooccurrents. Grâce à cette approximation, nous assimilons PP, le nombre d'empans où U1 et U2 apparaissent ensemble, à la fréquence de cooccurrence f12 de u2 dans le voisinage de u1 (le pivot) : PP ≈ f12

- les phrases, à l'intérieur desquelles nous calculons les cooccurrences, sont formées d'empans complets. En pratique, les phrases sont constituées d'empans et de fragments d'empans, puisqu'elles ont un nombre variable de mots, qui n'est pas un multiple de l :

N = nombre d'empans

N ≈ ∑(pour toutes les phrases P) longueur( P ) / l

Ce procédé permet de calculer f1, f2, et f12 « sur les mots », et pour calculer N nous ne tenons pas compte du nombre total de mots, mais du nombre total d'observations (i.e. d'empans). Ainsi, N dépend de la largeur de la fenêtre choisie. Il est normal de faire intervenir cette fenêtre dans le calcul, et les méthodes qui ne la prennent pas en compte sont des approximations encore plus grossières.

Néanmoins la simple fréquence d'une cooccurrence (et le t-score de façon similaire) ne garantit pas qu'il s'agisse d'une collocation au sens strict, car si d'un côté elle peut cacher des associations peu fréquentes mais significatives, comme par exemple *nez aquilin* ou *au fur et à mesure* (où *aquilin* et *fur* n'apparaissent que dans le contexte de leurs bases[61]), de l'autre côté elle peut faire croire qu'une simple combinaison libre ou colligation (une association ressortant de liens syntaxiques) comme *le nez* (déterminant + nom) est une collocation. En outre, la fréquence donne parfois comme résultat de fausses combinaisons, c'est-à-dire des mots qui ne sont pas du tout en relation.

### *5.4.2 L'information mutuelle*

Pour résoudre cette impasse, il est nécessaire de relativiser la valeur de la fréquence par le recours à des mesures telles que l'IM et le log-likelihood (g-score). Parmi celles-ci, la plus importante en opposition à la fréquence est l'IM. D'un côté, cette mesure permet de repérer les associations peu fréquentes mais dont les composants ont une association très forte en termes d'information partagée, comme dans les exemples précédents de *nez aquilin* et *au fur et à mesure*. Comme l'expliquent McEnery et al. (2006, p. 57) les collocations qui ont une valeur très élevée pour l'IM incluent des mots de basse fréquence (« collocations with high MI scores tend to include low-frequency words »). De l'autre côté, cette mesure permet d'exclure les associations très fréquentes mais non significatives signalant la présence d'une association sémantique très faible entre deux mots, comme dans l'exemple *le nez*, de fréquence très élevée. Pour ces raisons, l'IM est souvent considérée comme une

[61] Le collocatif *aquilin* s'associe aussi à *profil*.

mesure capable d'extraire les collocations de façon plus fiable que la seule mesure statistique des cooccurrences.

Il s'agit d'une mesure probabiliste dérivée de la théorie de l'information. Elle mesure l'information partagée par deux mots, c'est-à-dire l'information qu'un mot apporte sur un autre. Comme l'explique Bartsch (2004, p. 106), l'information mutuelle mesure la quantité d'information qu'un mot contient à partir de l'observation d'un autre mot.

Church & Hanks (1990, p. 24) citent Fano (1961) et donnent une définition plus technique de l'IM expliquant que si deux mots x et y ont une probabilité P(x) et P(y), alors leur information mutuelle, I(x, y), est définie comme :

« $I(x, y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$ »

Les auteurs expliquent que l'IM compare la probabilité d'observer x et y ensemble (« joint probability » ou probabilité conjointe) et la probabilité d'observer x et y par le simple jeu du hasard. Si l'association entre x et y est forte, alors la probabilité conjointe P(x,y) sera plus élevée que la valeur présupposant l'indépendance P(x) P(y) et le I(x,y) sera >> 0. Si l'association entre x et y n'est pas significative, alors la probabilité conjointe P(x,y) sera ~ 0. Enfin, si x et y sont dans une distribution complémentaire, alors P(x,y) sera inférieur à P(x) P(y), et par conséquent I(x,y) << 0. En résumé, plus forte est l'association, plus grande est l'IM. Selon les auteurs, les associations qui ont une information mutuelle > 3 sont « intéressantes », tandis que celles qui ont une information mutuelle < 3 ne le sont pas.

Pour résumer, l'information mutuelle est un indicateur de la force d'attraction entre les composants de la cooccurrence : une valeur très basse obtenue pour l'IM permet d'effacer les colligations, tandis qu'une valeur très haute obtenue pour l'IM suggère la présence d'une force associative entre le mot pivot et le collocatif.

*5.4.3 La dispersion*

La dispersion mesure la répartition d'une cooccurrence dans les parties dont le corpus se compose. Etant impossible de diviser notre corpus en sous-parties, nous avons décidé de calculer la dispersion de façon un peu grossière afin d'obtenir un signal d'alarme pour les combinaisons très fréquentes mais non significatives. Notre corpus est une collection de textes issus du Web. Nous avons donc calculé la dispersion (indiquée à la dernière colonne intitulée « domaines ») par le nombre de domaines différents dans lesquels une cooccurrence se présente. Par « domaine » nous désignons une source Internet, par exemple www.lemonde.fr, www.libe.fr, www.le-figaro.fr. Si le même phénomène est toujours observé sur un même site, la dispersion vaut 1 : cela signifie que les occurrences sont peu dispersées et ne se situent qu'au sein d'une source bien particulière. Par contre, un nombre élevé de domaines différents est une bonne indication de la généralité de l'observation statistique. Plus le nombre de domaines est élevé, plus le degré de dispersion de la cooccurrence dans le corpus est grand, selon l'hypothèse (et l'approximation) qu'un site Web traite plus ou moins du même sujet. La mesure de la dispersion s'est révélée très utile pour l'extraction des collocations fondamentales.

En conclusion, chaque mesure a sa spécificité et l'association de plusieurs mesures affine l'analyse. Par exemple, Bartsch (2004, p. 197-198) se sert de l'IM et du t-score ensemble car l'une sélectionne plutôt les mots pleins et l'autre les mots fonctionnels. Dubreuil (2008) est de la même opinion : le t-score est complémentaire à l'IM, étant donné qu'il permet de mettre en évidence, comme la fréquence absolue, beaucoup de colligations. Clear (1993, cité par Dubreuil, p. 282) affirme à propos du t-score que cette mesure repère de véritables collocations et permet au lexicographe de connaître les associations les plus typiques et les plus fréquentes dans l'usage. L'IM, par contre, est plus utile pour repérer des associations qui ne sont pas très fréquentes, mais dont les composants sont fortement associés, et qui forment des phrases idiomatiques, des proverbes, des phrases techniques et spécialisées.

Nous souhaitons préciser que les mesures d'associations que nous venons de citer sont des tests qui ne tiennent pas compte de la direction de la relation car elles n'indiquent pas si dans la cooccurrence un composant est plus significatif qu'un

autre. Considérons, par exemple, l'association *kith and kin* qui signifie « amis et parents » : *kith* joue le rôle de base parce qu'à la différence de *kin* ne se trouve pas dans d'autres contextes (exemple issu de Bartsch, 2004, p. 106).

Dans le chapitre suivant, nous présenterons la méthodologie suivie dans ce travail. Nous expliquerons en détail comment nous nous sommes servie des mesures statistiques que nous venons de décrire.

# CHAPITRE 6
# Méthodologie : la constitution de l'échantillon et le test soumis auprès de locuteurs natifs

Dans ce chapitre, nous montrerons tout d'abord la procédure suivie pour la constitution de l'échantillon des unités candidates au statut de collocations fondamentales (paragraphe 6.1). Puis nous décrirons le test soumis auprès de locuteurs natifs, à qui nous avons demandé de juger du caractère fondamental des associations extraites par la procédure automatique (paragraphe 6.2).

Dans le chapitre suivant, nous analyserons le score obtenu par les répondants dans le but d'évaluer l'échantillon extrait, et nous présenterons les résultats finaux de la recherche.

## 6.1 La constitution de l'échantillon

Comme nous l'avons expliqué auparavant, l'exécution de l'outil d'extraction a produit en sortie, d'une part, une liste de cooccurrents triée par fréquence d'association avec le mot pivot (la liste affiche aussi les valeurs des mesures associatives et de la dispersion) ; et d'autre part, une concordance. Ici, nous expliquerons comment nous avons traité la liste de cooccurrents dans le but de constituer l'échantillon des associations candidates au statut de collocations fondamentales :

- nous avons opéré un nettoyage à la fois manuel et automatique de la liste de fréquence, visant à éliminer le bruit des données extraites (paragraphe 6.1.1) ;
- nous avons choisi des valeurs de significativité pour les statistiques prises en compte (paragraphe 6.1.2) ;
- nous avons sélectionné les unités candidates au statut de collocations fondamentales (paragraphe 6.1.3).

### *6.1.1 Le nettoyage de la liste de fréquence*

Les sources majeures de bruit que nous avons éliminées automatiquement sont au nombre de deux : la présence parmi les cooccurrents de catégories vides ; et les fautes de lemmatisation. Nous les avons traitées de la façon suivante :

1) La suppression des catégories grammaticales vides

La suppression des catégories grammaticales autres que nom, verbe et adjectif a été permise grâce à l'utilisation d'une *stoplist* indiquant les mots grammaticaux les plus fréquents (ainsi que d'autres catégories telles que les pourcentages, les nombres, les signes de ponctuation, etc.). Les mots outils ne contribuent qu'à la création de liens de colligation, et pour cette raison nous avons voulu rétrécir l'analyse aux collocations lexicales et garder seulement les collocatifs correspondants aux catégories typiquement en association avec le nom (le mot pivot).[62]

2) La correction des erreurs de lemmatisation

La correction des erreurs de lemmatisation (qui ont pu persister à cause d'erreurs dues au lemmatisateur *TreeTagger*) a eu lieu grâce à un script de nettoyage permettant d'opérer des regroupements et de recalculer les valeurs des statistiques. Nous avons principalement lemmatisé les mots orthographiquement incorrects (la normalisation de l'accent, par exemple, a été faite pour *Noel* et *Noël*), les abréviations (par exemple *jours* et son abréviation *jrs* ont été regroupés sous le lemme *jour*) et quelques majuscules emphatiques que l'outil d'analyse ne traitait pas correctement (par exemple, *Saint* et *saint* – substantif et adjectif – étaient enregistrés séparément).

Le nettoyage manuel a également permis d'éliminer deux erreurs importantes : les collocatifs repérés seulement dans quelques sources (ayant une basse valeur de dispersion), et les cooccurrents produisant des entités nommées ou, de façon plus générale, des résultats non pertinents. Nous avons procédé de la façon suivante :

[62] Parmi les premiers résultats affichés par le calcul de la simple fréquence, le pivot *fête*, par exemple, se présente avec *la*, *à*, *pour*, car la simple fréquence ne tient pas compte de la force d'association entre les composants et sélectionne beaucoup de cooccurrents avec lesquels la base entretient un simple lien syntaxique ou même aucun.

3) La suppression des mots peu dispersés.

La dispersion mesure l'homogénéité de la diffusion d'une collocation dans le corpus. Le fait qu'une association ne soit fréquente que dans une section délimitée du corpus est révélateur d'un usage de cette association limité à des contextes restreints, et suggère son exclusion de l'inventaire fondamental.

La suppression des associations qui ne se présentent que dans quelques sources a eu lieu par élimination de toutes les cooccurrences ayant une valeur de dispersion égale ou inférieure à 50 sources. Ce seuil a été établi d'après le constat qu'avec un seuil de dispersion inférieur à 50 presque aucune cooccurrence n'est correcte ou ne peut être considérée comme étant acceptable. Généralement, l'élimination des cooccurrences ayant un indice de dispersion bas nous permet de réduire la liste des sorties d'environ 1 000 résultats, ce qui est un chiffre considérable.

Dans le **tableau II** (la dispersion est indiquée par l'étiquette « domaines »), nous montrons deux exemples très significatifs :

- L'association *fêtes panathénées,* à partir du mot pivot *fête,* est probablement repérée puisque les jours où cette association a été extraite du corpus du web correspondaient avec la célébration de ces fêtes, très discutées et commentées par la communauté Internet. La valeur très basse de la dispersion nous indique immédiatement qu'il ne s'agit pas d'une association fondamentale.
- De la même façon, nous avons remarqué que le cooccurrent *tarte* du pivot *fête* a été extrait par toutes les mesures utilisées sauf par la dispersion, dont la valeur 1 indique que *tarte* se trouve toujours dans la même source. L'association de *fête* et de *tarte* est due au fait que les deux mots partagent ponctuellement le même univers de discours et non à la présence d'un lien collocationnel.

**Tableau II** – Pivot *fête* : valeurs de dispersion de la cooccurrence avec *panathénées* et *tarte*

| Collocatif | f | logLike | t-score | z-score | IM | domaines |
|---|---|---|---|---|---|---|
| panathénées | 31 | 290,7203 | 89,87458 | 5,546559 | 5,570475 | 13 |
| tarte | 147 | 3558,637 | 5151,472 | 12,12429 | 12,10365 | 1 |

L'indice de la dispersion s'est donc révélé très utile pour réduire le bruit dérivant des biais du corpus (dûs à la structure du corpus, qui peut inclure des sites susceptibles de contenir des informations extrêmement répétitives) ainsi que la taille des données à analyser en sortie. Le plus étonnant est que la dispersion a souvent été le seul instrument ayant pu nous mettre en garde contre ces biais.

4) La suppression des cooccurrents qui produisent des entités nommées ou des résultats non pertinents

Cette action a nécessité la lecture presque intégrale des sorties, et a consisté en l'élimination d'emprunts (*aïd*), appellatifs (*Directeur*), titres de personnes (*Mr, Mme*), lieux géographiques (*Paris*), noms propres (*Jacques*), noms d'entreprises (*BBC*), mots étrangers (*webshopping*). Tous ces éléments produisent le plus souvent des entités nommées (des entités notamment moins sujettes aux variations que les collocations et abondamment traitées dans le domaine de l'extraction d'information). Si nous considérons de nouveau l'exemple de *fête*, nous remarquons que la suppression de *Beaujolais*, *Cinéma, Lumières, Mères, Epiphanie*, *Halloween, Toussaint,* permet de réduire de façon consistante le bruit dérivant des biais du corpus, celui-ci étant extrait du web. Il est vrai qu'il existe des fêtes très importantes et très représentatives du concept même de « fête », par exemple la fête de Noël. Il s'agit d'une association qui devrait sans doute être enseignée dans un cours de langue française destiné à des étudiants étrangers. Cependant, le but est de réduire le plus possible le bruit et de ne repérer que les cooccurrences les plus restreintes, qui sont probablement des collocations, et de dériver des implications didactiques qui les concernant. En résumé, les sources de bruit présentes dans notre corpus sont éliminées par des opérations à la fois automatiques et manuelles qui nous permettent de réduire de façon consistante la liste de fréquence en sortie. Des 2 000 cooccurrents initiaux, nous ne retenons pour certains mots pivots que 300 à 400 cooccurrents, qu'ensuite nous filtrons par la fréquence et les mesures associatives.

### *6.1.2 Le choix des seuils statistiques*

Après avoir nettoyé la liste des cooccurrents en sortie, il était nécessaire d'établir un seuil de significativité des statistiques prises en compte : la fréquence d'une part, et

les mesures associatives d'autre part (IM, t-score, z-score, log-likelihood). Pour toutes les mesures (sauf pour l'IM), nous avons établi le seuil de significativité aux 50 premiers résultats extraits. Expliquons la raison de ce choix.

1) Fréquence

Nous avons établi que les unités les plus fréquentes sont repérées parmi les 50 premiers cooccurrents, puisque :

a) Le seuil des 50 premiers cooccurrents ne devrait pas couvrir les associations techniques ou trop spécifiques, notamment celles privilégiées par des mesures comme l'IM et le log-likelihood.
b) Aux premiers rangs, les différentes mesures sont assez homogènes, car les cooccurrents extraits sont presque identiques. Cependant, nous précisons que cela est aussi dû à la taille très grande du corpus.
c) Au fur et à mesure que la fréquence se réduit, le nombre de cooccurrences considérées comme acceptables se réduit également au moyen de l'une des mesures choisies (et les cas d'accord entre plusieurs mesures deviennent très rares)

2) Mesures associatives

Pour les mesures associatives également (sauf pour l'IM), le seuil choisi est celui des 50 premiers résultats. Etablir une valeur-seuil fixe est assez complexe, sauf pour l'IM, pour laquelle les études indiquent une valeur de significativité assez stable correspondant à 3 (Church & Hanks, 1990), que nous avons retenue. Il a été donc nécessaire de trouver une solution pratique, qui est commune à plusieurs études. Elle consiste à sélectionner les premiers résultats les plus votés pour chaque mesure. Nous avons choisi les 50 premiers. Ce choix a été validé par une évaluation subjective qui a permis de constater l'absence de cooccurrents significatifs au-delà de ce seuil. Nous avons remarqué que ce seuil était parfois un peu trop large, mais nous avons préféré augmenter légèrement le bruit et ne pas avoir à exclure de coocurrents potentiellement significatifs.

### *6.1.3 La sélection des unités candidates au statut de collocations fondamentales*

La supposition à la base de notre travail est que les unités candidates au statut de collocations fondamentales, comme les mots fondamentaux, doivent pouvoir être repérées à l'aide de la fréquence et de la disponibilité[63], selon l'hypothèse que le caractère fondamental ne peut pas dépendre uniquement de la fréquence. Nos données en sortie ont été précédemment soumises à des opérations de nettoyage et ainsi réduites à quelques centaines. Nous avons voulu filtrer ces centaines de résultats et sélectionner, d'une part, les unités polylexicales les plus fréquentes, et d'autre part les unités polylexicales moins fréquentes mais pertinentes qui sont candidates au statut de collocations fondamentales (ce que nous allons vérifier par la suite interrogeant les locuteurs natifs). Dans les deux cas, les mesures associatives assurent la significativité de l'association repérée, c'est-à-dire la présence d'une collocation. Considérons le procédé de sélection des deux types d'unités plus en détail :

1) La sélection des unités polylexicales les plus fréquentes

Afin de relativiser les erreurs dérivant du simple calcul de fréquence telles que les liens de colligation et les fausses cooccurrences, nous avons choisi d'avoir recours aux mesures associatives. En outre, la fréquence en elle-même ne nous dit rien ou elle nous dit peu sur la présence d'un lien collocationnel. Nous avons donc sélectionné les 50 premiers cooccurrents les plus fréquents et nous les avons triés par ordre de significativité à l'aide de mesures associatives : l'information mutuelle, le log-likelihood, le t-score et le z-score (nous renvoyons au chapitre 5 pour une discussion sur la valeur de ces mesures). Chacune de ces mesures a sa propre tendance spécifique. Le t-score est similaire à la fréquence car il a tendance à favoriser les emplois les plus fréquents, et sa marge d'erreur est assez élevée en ce qui concerne les colligations et les fausses cooccurrences. Par contre, l'IM et le log-likelihood favorisent les emplois rares et les associations exclusives, du type *fêtes panathénées*. Enfin, le z-score semble une solution d'équilibre entre ces deux

[63] Les unités disponibles sont très fonctionnelles pour la communication bien que peu fréquentes dans le discours oral ou écrit.

opposés. La dispersion, test sur lequel nous nous appuyons constamment, sert à vérifier que les cooccurrences repérées sont bien réparties dans le corpus et qu'elles ne proviennent pas d'une même source du fait des biais du corpus.

Considérons les premiers 15 cooccurrents les plus fréquents de *rencontre* (**tableau III**) : selon la co-évaluation des mesures associatives (les mesures log-likelihood, t-score, z-score et Information mutuelle sont indiquées respectivement par 1., 2., 3. et 4.), certains d'entre eux sont très peu significatifs et d'autres ne le sont pas du tout.

**Tableau III** – Pivot *rencontre* : co-évaluation des cooccurrents triés par simple fréquence

| Collocatif | f | 1. logLike | 2. t-score | 3. z-score | 4. IM | co-évaluation |
|---|---|---|---|---|---|---|
| être | 18786 | 9282,108 | -82,376 | -110,77 | -0,59232 | 1 |
| avoir | 14079 | 1927,412 | -39,9651 | -47,2584 | -0,33525 | 1 |
| lieu | 6839 | 17118,54 | 209,324 | 72,71799 | 2,114589 | 123 |
| organiser | 6064 | 18932,41 | 243,261 | 71,20053 | 2,457269 | 123 |
| faire | 5958 | 1,422598 | -1,17306 | -1,18201 | -0,0152 | |
| aller | 5175 | 3370,762 | 68,84857 | 43,3793 | 0,923854 | 123 |
| échange | 3693 | 13015,48 | 215,8616 | 56,59305 | 2,677503 | 123 |
| site | 3069 | 534,3745 | 24,88989 | 19,91883 | 0,445592 | 3 |
| national | 2979 | 2377,104 | 59,54095 | 35,34579 | 1,04297 | 123 |
| international | 2777 | 4165,384 | 88,52255 | 41,25293 | 1,52707 | 123 |
| professionnel | 2515 | 1953,349 | 53,81157 | 32,19667 | 1,027251 | 123 |
| permettre | 2515 | 222,3738 | 15,66489 | 13,40823 | 0,311106 | |
| occasion | 2331 | 5197,569 | 111,0772 | 41,53101 | 1,96757 | 123 |
| nouveau | 2320 | 0,286924 | -0,53155 | -0,53449 | -0,01104 | |
| heure | 2315 | 27,78842 | -5,14812 | -5,4309 | -0,10695 | |

Les cooccurrents ayant une co-évaluation correspondant à « 1234 » sont repérés par toutes les mesures associatives. Parmi les 15 cooccurrents les plus fréquents du pivot *rencontre*, nous remarquons que cela n'est jamais le cas. Par contre, ils sont nombreux les cooccurrents repérés par 3 mesures associatives, par exemple *lieu* et *organiser*. Enfin, nous constatons que des cooccurrents génériques tels que *être*, *avoir* et *site* ne sont repérés que par une seule mesure, tandis que des cooccurrents tels que *heure* et *nouveau*, produisant des associations libres, ne sont repérés par aucune des mesures associatives (et nous les excluons).

Nous déduisons ainsi que les cooccurrences évaluées positivement par plusieurs mesures associatives produisent très probablement des collocations ; par contre, les cooccurrences sans aucun accord entre fréquence et mesures associatives ont peu de chance d'être des collocations.

2) La sélection des unités polylexicales moins fréquentes mais pertinentes

Comme nous l'avons précisé dans ce travail, bien que la fréquence joue un rôle important, elle seule n'est pas suffisante à définir ce qui est fondamental. Nous avons également pris en compte un critère comme la disponibilité car, comme l'explique Gougenheim (*Français élémentaire*, 1954, p. 10), « tandis que la fréquence correspond aux automatismes du langage, la disponibilité est liée à l'intérêt que présentent pour nous les notions exprimées par les mots ». Nous avons donc également repéré les unités non fréquentes mais pertinentes qui pourraient représenter une actualisation contextuelle fondamentale du mot pivot choisi, une partie essentielle de son univers de discours conceptuel. Nous avons extrait les associations qui se trouvaient au-delà du seuil des 50 unités les plus fréquentes, mais qui ont été repérées par au moins une des mesures associatives.

Le **tableau IV** montre les cooccurrences non incluses dans les premiers 50 résultats les plus fréquents, mais votées par au moins une des quatre mesures associatives, pour le mot pivot *colloque*.

**Tableau IV** – Pivot *colloque* : unités polylexicales moins fréquentes mais pertinentes

| collocatif | f | co-évaluation |
|---|---|---|
| compte-rendu | 255 | 1234 |
| clôture | 234 | 1234 |
| sénat | 218 | 1234 |
| interdisciplinaire | 183 | 124 |
| co-organisé | 78 | 124 |
| sciences | 124 | 124 |
| organisateur | 221 | 123 |
| symposium | 83 | 24 |
| contribution | 288 | 23 |
| salon | 282 | 23 |
| réunion | 376 | 13 |
| prochain | 374 | 3 |
| dérouler | 282 | 3 |
| atelier | 335 | 3 |
| pluridisciplinaire | 116 | 2 |
| singulier | 133 | 2 |

Pour résumer, nous avons tout d'abord repéré les unités polylexicales les plus fréquentes et celles moins fréquentes mais pertinentes, qui sont candidates au statut de collocations fondamentales. Dans les deux cas, l'analyse de la validité des cooccurrents effectuée par la consultation des concordances, nous a permis de réduire ultérieurement notre échantillon. Dans le chapitre 7, nous soumettrons les

associations sélectionnées au jugement des locuteurs natifs afin d'évaluer leur utilité communicative et d'exclure les unités fréquentes ou non fréquentes qui n'ont aucun intérêt pour les interviewés.

## 6.2 Le test soumis auprès de locuteurs natifs

Dans ce paragraphe, nous allons décrire le test soumis à quatre-vingt-dix locuteurs natifs du français dans le but de comparer l'échantillon constitué à l'aide de l'outil d'extraction et les jugements offerts par les natifs.

Le test demande aux répondants de marquer avec une croix les associations qui, à leur avis, peuvent être considérées comme étant « fondamentales ». Avant de commencer le test, ils sont invités à lire les instructions, qui leur fournissent une explication détaillée du concept d'« association fondamentale », avec des exemples extraits du corpus.

### *6.2.1 Les locuteurs natifs*

Nous avons interrogé quatre-vingt-dix locuteurs natifs à travers un questionnaire écrit dans le but de comprendre quelles associations étaient fondamentales selon leur intuition. Il s'agit de locuteurs volontaires, de nationalité française (pour la majorité), suisse, belge ou canadienne. Leur niveau d'instruction, leur âge et leur domaine professionnel sont plutôt variés. La plupart d'entre eux ont répondu à une annonce qui leur expliquait l'objectif et l'utilité du test et qui a été postée dans le réseau social web « couchsurfing » (www.couchsurfing.com).

Le test a été soumis à 3 groupes de 30 locuteurs chacun :

- le groupe des 3 mots pivots *conférence, débat, rencontre* aux 30 premiers locuteurs
- le groupe des 3 mots pivots *colloque, fête, réunion* aux 30 autres locuteurs
- le groupe restant des 4 mots pivots *congrès, séminaire, interview, conversation* aux 30 derniers locuteurs.

En résumé, un total de 10 mots pivots, distribués en trois groupes de 3, 3 et 4 mots chacun, a été réparti à 90 répondants, à leur tour divisés en 3 groupes de 30. Il est évident que le nombre d'unités polylexicales (les unités les plus fréquentes significativement et les unités moins fréquentes mais significatives et pertinentes extraites précédemment) soumises au jugement des locuteurs variait selon les résultats issus de l'extraction automatique : par exemple, pour la base lexicale *débat,* l'outil d'extraction a repéré 55 associations contre les 49 repérées pour *rencontre*.

### *6.2.2 Les instructions et le test*

Le concept de « fondamental » est un concept difficile à définir, bien qu'il existe (comme nous l'avons remarqué dans le chapitre 2) des critères pour le mesurer comme la fréquence, l'aspect concret, le référentiel, etc. Ce concept est encore plus difficile à expliquer à des locuteurs natifs qui ne sont pas forcément concernés par les matières linguistiques.

Le test envisagé pour la présente recherche a été conçu pour éviter de surcharger les locuteurs avec des explications techniques, et ainsi être le plus simple et intuitif possible. En effet, bien que le concept de « fondamental » puisse apparaître comme étant obscur, il a été demandé à chacun de reconnaître s'il existait des associations qui lui apparaissaient intuitivement plus familières que d'autres, afin qu'il puisse les identifier et tenter de les juger.

Nous avons expliqué le concept de « fondamental » en utilisant deux de nos dix pivots, *séminaire* et *fête*. Dans ce qui suit, le **tableau V** affiche les instructions (se servant d'exemples avec le mot *séminaire)* données aux locuteurs avant qu'ils exécutent le test. Nous avons bien entendu fait attention à ce que le locuteur qui lit les instructions avec des exemples extraits du corpus concernant *séminaire* ne soit pas appelé à juger des associations de ce même pivot. La même chose vaut pour *fête*.

**Tableau V** - Les instructions du test soumis auprès de locuteurs natifs

**POURQUOI CE TEST ?**

Dans la liste qui suit, sont présentés, pour un mot donné (qu'on appellera « base »), d'autres mots qui s'y associent en contexte : par exemple, la « base » *séminaire* peut s'associer avec les autres mots suivants (parmi de nombreux d'autres) :

**organiser** un *séminaire*
> Ex. L'association organise un séminaire annuel sur le racisme.

*séminaire* **de formation**
> Ex. L'AFNIC intervient régulièrement dans des séminaires de formation de Télécom Paris.

*séminaire* **résidentiel**
> Ex. De 1986 à 1990, le CEFAG se compose de cinq séminaires résidentiels.

Chacun d'entre vous, lorsqu'il entend un mot comme *séminaire* est capable de lui associer spontanément d'autres mots, parce que ses emplois sont fréquents ou importants dans la vie courante : parmi ces associations privilégiées, on trouve par exemple *organiser* un *séminaire*. En revanche, il vous viendrait probablement moins facilement à l'esprit une association comme *séminaire résidentiel*, parce que l'on n'en a besoin que rarement. Nous vous demandons de repérer les associations privilégiées et de laisser de côté les associations que vous jugerez moins fondamentales, du type *séminaire résidentiel*.

**COMMENT EXECUTER LE TEST ?**

Il vous est demandé de cocher avec un X, les mots associés permettant de construire des associations privilégiées ou fondamentales.

Le test ne prend que quelques minutes et doit suivre trois règles de base :

1. Réfléchir quelques secondes à chaque proposition de mot associé

2. Imaginer la façon dont la « base » et l'autre mot s'associent. Par exemple, à partir de *séminaire* et *recherche*, il est possible de former l'association *séminaire de recherche.*

3. Mettre une croix dans la colonne de droite chaque fois que vous jugerez que l'association entre les deux mots est « fondamentale » (c'est-à-dire immédiatement identifiable, parce que fréquente ou importante pour les besoins de la vie courante).

Bien que le fait de présenter un exemple puisse toujours constituer un biais en influençant les réponses des enquêtés, il nous ne semble pas que l'exemple fourni ait pu constituer un tel biais (amenant les locuteurs à éliminer automatiquement des associations rares ou spécialisées). En effet, nous avons fourni comme exemple une association très rare et très spécialisée comme *séminaire résidentiel* qui ne constitue certainement une association fondamentale.

Après la lecture des instructions, les locuteurs étaient censés commencer le test. Les associations fréquentes et celles moins fréquentes extraites à l'aide de l'outil d'exploration du corpus leur étaient présentées, et on leur demandait d'exprimer un jugement sur leur caractère fondamental. Le **tableau VI** présente le test concernant le pivot *conférence* et les réponses données par un locuteur.

**Tableau VI** – Pivot *conférence* : réponses au test d'un locuteur natif

| **BASE : conférence**<br>N.B. La touche * indique la position de la BASE, avant ou après le mot associé | |
|---|---|
| **MOTS ASSOCIES** | **FONDAMENTAL ?** |
| une * **annuelle** | |
| une * **de consensus** | |
| une * **épiscopale** | **x** |
| une * **inaugurale** | |
| une * **intergouvernementale** | |
| une * **internationale** | **x** |
| une * **introductive** | |
| une * **mensuelle** | |
| une * **ministérielle** | |
| une * **de presse** | **x** |
| la * **de rentrée** | |
| la * **s'intitule ...** | |
| une * **téléphonique** | **x** |
| une * **tripartite** | |
| **à l'occasion de la** * | |
| l'**animation** de la * | |
| **animer** une* | **x** |
| **assister à** une * | **x** |
| **au cours de la** * | |
| **avoir** une * | **x** |
| le **compte-rendu** de la * | |
| un **cycle** de *s | **x** |
| **donner** une * | |
| **être à** une * | |
| **faire** une * | |
| un **maître de** *s | **x** |
| **organiser** une * | **x** |
| **participer à** une * | **x** |
| le **président** de la * | |
| le **programme** de la * | |
| **prononcer** une* | |
| la **salle** de * | **x** |
| **tenir** une * | |
| la **tenue** de la * | |
| le **thème** de la * | |

Nous tenons à préciser que les associations sont présentées dans leur structure lexico-syntaxique d'occurrence la plus fréquente : par exemple, étant donné que *conférence* se présente toujours au pluriel en association avec *cycle*, la forme suivante a été choisie : *un cycle de conférences*. La même chose vaut pour l'association entre *conférence* et *s'intituler* : ce dernier se présentant le plus souvent dans notre corpus dans la forme verbale du présent, nous avons présenté aux répondants l'association *la conférence s'intitule*. Le même principe vaut pour l'alternance article défini/indéfini dans l'association de *animation* et de *conférence* : la forme la plus fréquente dans l'usage, *l'animation de la conférence*, a été choisie. Quant à l'ordre de présentation des associations, il a été choisi au hasard.

Dans le prochain chapitre, nous comparerons les collocations fondamentales repérées automatiquement et les collocations fondamentales sélectionnées par les locuteurs natifs. Notre but est de vérifier s'il existe une corrélation entre fréquence et score obtenu par les locuteurs, et de comprendre de quels facteurs autres que la fréquence peut dépendre l'attribution du caractère fondamental de la part des répondants. En outre, cela nous permettra d'évaluer la justesse de l'échantillon constitué (ainsi que celle de l'outil d'exploration et des mesures utilisées).

# Chapitre 7
# Evaluation et résultats finaux

Dans le chapitre précédent, nous avons décrit la procédure d'extraction automatique des unités polylexicales[64] candidates au statut de collocations fondamentales. Nous n'avons retenu, d'une part, que les unités les plus fréquentes, et d'autre part, les unités moins fréquentes qui pourraient être pertinentes du point de vue de l'acte de communication selon les locuteurs natifs. Pour ce faire, nous avons utilisé la fréquence, ainsi que le test de dispersion ainsi que des mesures associatives. Etant donné que les mesures associatives nous informent sur la force d'association entre les composants de l'unité, mais non sur l'utilité communicative de l'unité polylexicale, l'échantillon des associations extraites a ensuite été soumis au jugement des locuteurs natifs : le dépouillement des réponses des enquêtés qui ont complété leurs tests entièrement et lisiblement nous a permis de tirer des résultats significatifs, que nous présenterons dans ce chapitre.

## 7.1 Existence d'une corrélation positive non systématique entre fréquence et caractère fondamental attribué par les locuteurs

Le premier résultat de notre étude est la présence d'une corrélation faiblement positive, mais comme nous allons le voir, non systématique, entre fréquence des associations extraites du corpus et score obtenu par les locuteurs, à qui il était demandé de choisir les associations qui leur semblaient les plus essentielles (voir chapitre précédent). Cela veut dire que les associations les plus fréquentes ne sont pas *toujours* les plus sélectionnées. L'absence de systématicité dans la corrélation se

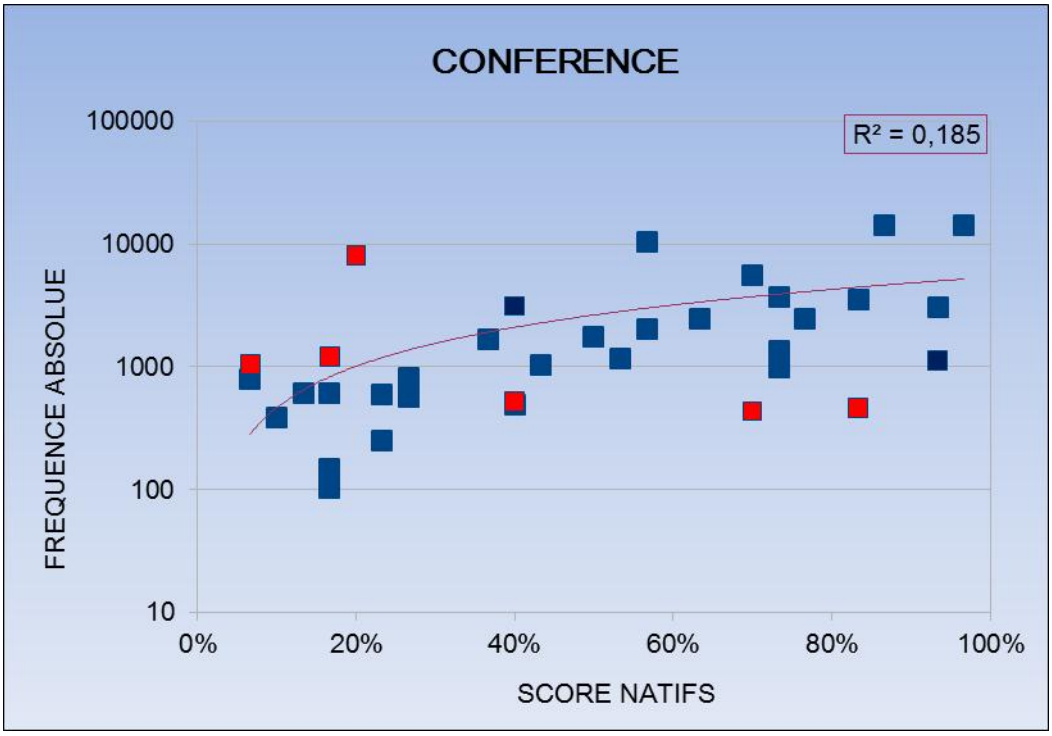
[64] Nous rappelons que nous avons souvent employé les termes « association » ou « unité polylexicale » à la place du terme « collocation », car les mesures associatives utilisées ne sont pas toujours capables de distinguer les véritables collocations des associations libres mais fréquentes. D'ailleurs, comme nous l'expliquerons dans le chapitre 8, une distinction rigide entre les deux types d'unité n'est pas toujours utile pour un traitement didactique.

justifie par deux raisons :

1) Il n'y a pas toujours de corrélation positive, car deux mots sur dix de notre échantillon présentent une corrélation négative.
2) Si une corrélation est présente, elle est affaiblie par la présence de points singuliers qui s'écartent de la norme.

Dans ce qui suit, nous proposons deux exemples de corrélation, la corrélation positive constatée pour le mot *conférence*, et la corrélation négative constatée pour le mot *colloque*. Nous montrerons, dans les deux cas, la présence de quelques associations qui représentent une déviation de la linéarité de la corrélation. Considérons, tout d'abord, le mot *conférence :* la **figure I** montre le graphe de corrélation.

**Figure I** – Pivot *conférence* : corrélation positive entre fréquence (échelle logarithmique) et score des natifs. Quelques points singuliers de fréquence élevée et score faible, et de fréquence basse et score élevé.




La droite de régression nous permet de mettre en rapport les deux variables, la fréquence des associations et le score obtenu par les locuteurs natifs. Nous

remarquons, pour ce mot, l'existence d'une corrélation faible, mais positive : au fur et à mesure que la fréquence de l'association augmente, le score attribué par les locuteurs à l'association croît également.

Cependant, comme la figure le met bien en évidence, quelques associations s'éloignent du « centre » de la distribution, car elles représentent une « déviation » dans l'évolution moyenne de la corrélation entre fréquence et caractère fondamental assigné par les interviewés. Autrement dit, les données ne se concentrent pas autour de la tendance centrale, mais sont très dispersées, car certaines associations sont plus à l'écart par rapport aux autres. Nous avons marqué en rouge ces associations.

De gauche à droite, les points singuliers les plus extrêmes, de haute fréquence et de score faible, sont *la conférence de rentrée* (1053, 7%), *le président de la conférence* (1203, 17%) et *avoir une conférence* (8069, 20%) ; tandis que des points singuliers de basse fréquence et de score élevé sont *une conférence intergouvernementale* (492, 40%), *une conférence téléphonique* (458, 83%) et *le compte-rendu de la conférence* (436, 70%). Il s'agit d'unités qui sont très fréquentes mais peu retenues par les locuteurs (par exemple, *avoir une conférence*), ou à l'inverse, d'unités peu fréquentes mais souvent retenues (par exemple, *une conférence téléphonique*).

Le coefficient de corrélation de Pearson permet de mesurer numériquement l'écart de la distribution des points par rapport à la droite de régression. Si le caractère fondamental des associations dépendait linéairement de la fréquence, on trouverait une valeur très élevée de ce coefficient. Mais dans nos observations nous avons toujours obtenu des valeurs inférieures à 0,5, avec un maximum de 0,43 pour *conférence*. Le coefficient de corrélation linéaire serait maximal et vaudrait 1 si l'une des deux variables était une fonction linéaire de l'autre (i.e. en représentant l'une en fonction de l'autre on obtiendrait une droite croissante ou décroissante). Dans le cas où il n'y aurait aucune corrélation, on obtiendrait la valeur 0. Dans notre exemple, si on avait une corrélation parfaite entre les deux variables, les associations jugées comme les plus fondamentales par les locuteurs natifs devraient *toujours* être les plus fréquentes, ce qui n'est pas le cas pour certaines associations.

Dans le **tableau VII** nous représentons les scores du coefficient de corrélation de Pearson obtenus par les dix mots pivots :

**Tableau VII** - Scores du coefficient de corrélation de Pearson pour les dix pivots

| PIVOTS | COEFFICIENT DE CORRELATION DE PEARSON |
|---|---|
| *Colloque* | -0,19 |
| *Conférence* | 0,43 |
| *Congrès* | 0,08 |
| *Conversation* | 0,27 |
| *Débat* | 0,30 |
| *Fête* | 0,16 |
| *Interview* | 0,05 |
| *Rencontre* | - 0,11 |
| *Réunion* | 0,29 |
| *Séminaire* | 0,29 |

Nous souhaitons souligner qu'à cause de la dispersion élevée des données, il n'est pas toujours simple de représenter ces points singuliers sur la même échelle (cf. figure I). D'ailleurs, nous n'avons marqué que les cas de déviation les plus nets. Ces cas sont généralement situés à l'extrémité supérieure gauche et à l'extrémité inférieure droite de notre figure. Les valeurs de fréquence absolue que nous avons mesurées dans notre corpus vont de 0 à 18 786 au maximum. Tous les mots analysés présentent la même distribution de données : on a, en général, un petit nombre d'associations de très haute fréquence, la plupart des autres associations étant regroupées en *clusters* avec des valeurs de fréquence assez proches les unes des autres. C'est pourquoi nous avons choisi une représentation logarithmique des données sur une échelle 10[65]. De cette façon, nous pouvons faire rentrer sur la même échelle les valeurs de haute fréquence et les autres associations. Cependant, si la représentation logarithmique nous aide à rapprocher les hautes fréquences des autres données, elle rapproche également, dans la figure, les données regroupées. Cela explique pourquoi quelques points singuliers peuvent apparaître aussi près de la droite et relativement centrés.

En dehors de cette vision géométrique de la corrélation, nous avons détaillé, dans le **tableau VIII**, les résultats de la figure I concernant *conférence* :

[65] Les graduations de l'axe des ordonnées correspondent respectivement au logarithme de 10, 100, 1 000, 10 000 et 100 000.

**Tableau VIII** – Pivot *conférence* : liste des associations triées par fréquence décroissante et quelques points singuliers (marqués en rouge).

| ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|
| une * **de presse** | 14207 | 97% |
| un **maître de** *s | 14106 | 87% |
| **être à** une * | 10418 | 57% |
| **avoir** une * | 8069 | 20% |
| **organiser** une * | 5589 | 70% |
| une * **internationale** | 3729 | 73% |
| **donner** une * | 3487 | 83% |
| un **cycle** de *s | 3111 | 40% |
| la **salle** de * | 3045 | 93% |
| le **thème** de la * | 2478 | 77% |
| **au cours de la** * | 2439 | 63% |
| **tenir** une * | 2045 | 57% |
| le **programme** de la * | 1749 | 50% |
| **faire** une * | 1674 | 37% |
| **animer** une* | 1331 | 73% |
| **participer à** une * | 1233 | 73% |
| le **président** de la * | 1203 | 17% |
| **à l'occasion de la** * | 1166 | 53% |
| **assister à** une * | 1130 | 93% |
| la * **de rentrée** | 1053 | 7% |
| une * **annuelle** | 1043 | 43% |
| l'**animation** de la * | 809 | 27% |
| une * **de consensus** | 790 | 7% |
| une * **épiscopale** | 609 | 17% |
| **prononcer** une* | 608 | 13% |
| la * **s'intitule ...** | 595 | 23% |
| une * **mensuelle** | 569 | 27% |
| une * **ministérielle** | 524 | 40% |
| une * **intergouvernementale** | 492 | 40% |
| une * **téléphonique** | 459 | 83% |
| le **compte-rendu** de la * | 436 | 70% |
| la **tenue** de la * | 382 | 10% |
| une * **inaugurale** | 248 | 23% |
| une * **tripartite** | 147 | 17% |
| une * **introductive** | 102 | 17% |

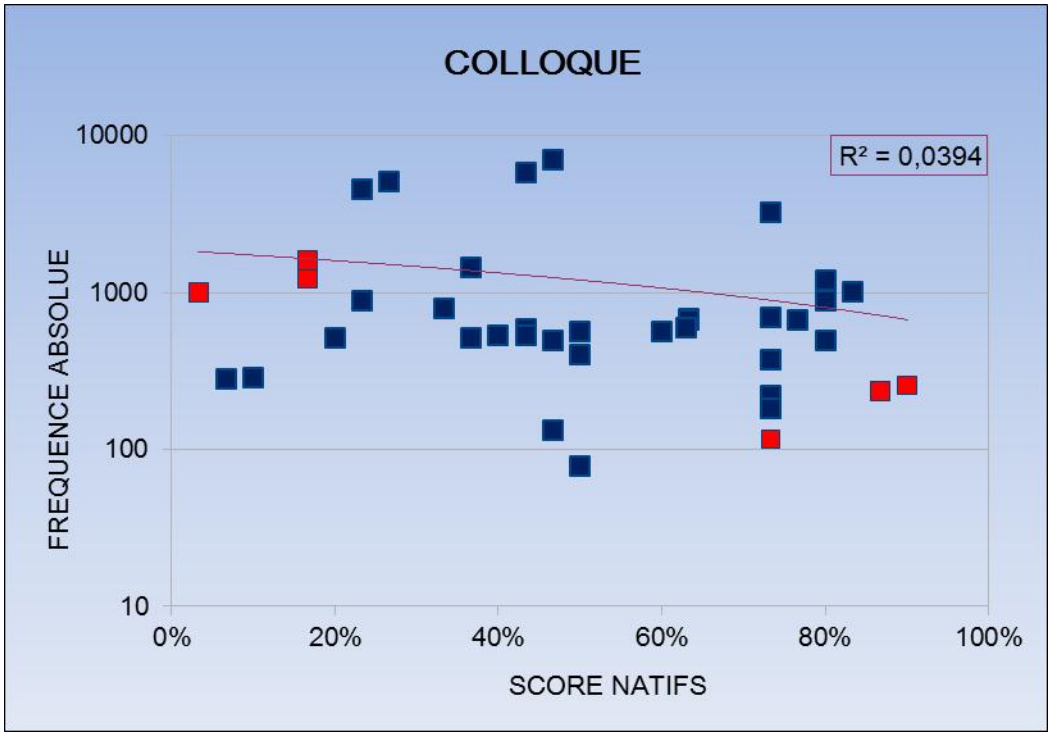
Dans le tableau que nous venons de présenter, nous avons explicité la liste des associations et les points qui présentent à la fois une fréquence élevée et un score faible (ou vice-versa) au vu de l'ensemble des autres associations. Ces points sont marqués en rouge. Le score en pourcentage obtenu par les répondants est également affiché. Quelques associations très fréquentes mais peu sélectionnées sont présentées dans la partie centrale supérieure de notre tableau ; à l'inverse, quelques associations souvent sélectionnées bien qu'elles aient une fréquence basse sont présentées dans la partie inférieure de notre tableau. Nous tenons à répéter que nous n'avons marqué que les points singuliers les plus nets, dont les valeurs de fréquence et de score sont très distantes.

Le fait que la fréquence ne puisse être considérée comme un facteur absolu, dont dépendrait le jugement des locuteurs de ce qui serait ou non fondamental, apparaît encore plus évident si nous analysons le mot *colloque* (**figure II** à la page suivante), car la corrélation entre les deux variables est faiblement négative.

Le mot *colloque*, de la même façon que le mot *conférence*, présente des associations très fréquentes mais peu sélectionnées, et à l'inverse des associations moins fréquentes mais souvent retenues. A la différence du cas précédent, ces points ne sont plus vraiment déviants par rapport à la droite de régression, car celle-ci est décroissante, la corrélation linéaire étant faiblement négative (cf. les points en rouge sur la figure I).

De gauche à droite, les points de haute fréquence et de score faible les plus nets sont : *un colloque scientifique* (1003, 3%), *le thème du colloque* (1215, 17%) et *les communications du colloque* (1605, 17%) ; tandis que les points de basse fréquence et de score élevé sont : *un colloque pluridisciplinaire* (116, 73%), *la clôture du colloque* (234, 87%) et *le compte-rendu du colloque* (255, 90%) :

**FIGURE II** – Pivot *colloque* : corrélation négative entre fréquence (échelle logarithmique) et score des natifs. Quelques points singuliers de fréquence élevée et score faible, et de fréquence basse et score élevé.




Voici également, pour le mot *colloque*, la liste de toutes les associations triées par fréquence (**tableau IX**) :

**Tableau IX –** Pivot *colloque* : liste des associations triées par fréquence décroissante et quelques points singuliers (marqués en rouge).

| ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|
| les **actes** du * | 7056 | 47% |
| **organiser** un * | 5772 | 43% |
| un * **international** | 5056 | 27% |
| **être** à un * | 4565 | 23% |
| **avoir** un * | 3220 | 73% |
| les **communications** du * | 1605 | 17% |
| l'**organisation** du * | 1451 | 37% |
| le **thème** du * | 1215 | 17% |
| **tenir** un * | 1188 | 80% |
| un * **national** | 1007 | 83% |
| un * **scientifique** | 1003 | 3% |
| le **programme** du * | 886 | 23% |
| une **journée** de * | 886 | 80% |
| **à l'occasion du** * | 790 | 33% |
| **participer** au * | 699 | 73% |
| **dans le cadre du** * | 686 | 63% |
| le * **a lieu …** | 673 | 63% |
| un * **consacré à …** | 664 | 77% |
| un * **d'étude** | 601 | 63% |
| **présenter** un * | 587 | 43% |
| **faire** un * | 562 | 50% |
| un * **annuel** | 559 | 60% |
| la **participation** au * | 535 | 43% |
| l'**ouverture** du * | 530 | 40% |
| un * **intitulé X** | 515 | 20% |
| les **participants** au * | 513 | 37% |
| l'**intervention** au * | 495 | 80% |
| **réunir** un * | 491 | 47% |
| la **présentation** du * | 399 | 50% |
| le **prochain** * | 374 | 73% |
| la **contribution** au * | 288 | 10% |
| le * **se déroule**… | 282 | 7% |
| le **compte-rendu** du * | 255 | 90% |
| la **clôture** du * | 234 | 87% |
| l'**organisateur** du * | 221 | 73% |
| un * **interdisciplinaire** | 183 | 73% |
| un * **singulier** | 133 | 47% |
| un * **pluridisciplinaire** | 116 | 73% |
| un * **co-organisé** | 78 | 50% |

Pour conclure, nous pouvons affirmer que, globalement, il existe une corrélation linéaire positive entre la fréquence et le caractère fondamental, mais que la présence de points singuliers qui s'écartent de la norme affaiblit cette corrélation et montre qu'elle n'est pas systématique. Les locuteurs natifs ont parfois ignoré certaines associations très fréquentes, tandis qu'ils ont d'autres fois retenu des associations plus rares. Cette observation confirme l'hypothèse selon laquelle ce qui est fondamental ne ressort pas uniquement du facteur de la fréquence. De quel facteur dépend alors l'attribution du caractère fondamental ? Nous tenterons de l'expliquer dans le prochain paragraphe, afin de comprendre ce qui oriente le choix des locuteurs quand ils excluent des unités très fréquentes, ou quand à l'inverse ils sélectionnent des unités peu fréquentes.

Nous renvoyons aux **ANNEXES A** pour les figures de corrélation concernant tous les pivots de notre échantillon.

## 7.2 Rôle du figement dans l'attribution du caractère fondamental

A travers une analyse plus attentive de nos données, nous avons remarqué la présence d'un trait qui revient dans tous les mots analysés, le rôle du figement dans l'attribution du caractère fondamental, au moins pour les points singuliers les plus nets. A notre avis, la fréquence est le facteur responsable de l'attribution du caractère fondamental aux associations de la part des locuteurs, mais de façon non systématique. En effet, dans des cas singuliers, le choix dépend du degré de figement : une association peu fréquente est quand même retenue lorsqu'elle comporte un certain degré de figement ; à l'inverse, une association très fréquente est exclue lorsqu'elle est libre. En résumé, si pour certaines associations, il n'y a pas de correspondance entre fréquence et attribution du caractère fondamental, alors la cause est à chercher dans le figement.

Avant de fournir un exemple, expliquons ce que nous entendons par figement et reprenons la définition de travail de collocation que nous avons présentée dans le chapitre 4, et qui nous aidera à établir le degré de figement d'une association. Le figement est un trait des unités polylexicales agissant au niveau sémantique et/ou syntaxique, que Gross (1996) explique de la façon suivante :

« Une séquence est figée du point de vue syntaxique quand elle refuse toutes les possibilités combinatoires ou transformationnelles qui caractérisent habituellement une suite de ce type. Elle est figée sémantiquement quand le sens est opaque ou non compositionnel, c'est-à-dire quand il ne peut pas être déduit du sens des éléments composants. Le figement peut être partiel si la contrainte qui pèse sur une séquence donnée n'est pas absolue, s'il existe des degrés de liberté » (p. 154).

Comme l'auteur le fait remarquer, il est plutôt rare qu'une suite lexicale figure comme une entrée lexicale indépendante et qu'elle soit complètement bloquée sur le plan sémantique et syntaxique. Une suite enregistrée comme étant une entrée à part dans les dictionnaires est par exemple *fait divers*, qui n'est ni transparente ni prédicative, et qui n'admet pas de nominalisation. La suite *fait historique*, par contre, n'est pas totalement figée, ni opaque. Gross (1996) affirme à ce propos que :

« [...] les suites totalement figées sont très minoritaires par rapport à celles qui ont des restrictions partielles » (p. 22).

Dans le chapitre 4, nous avons proposé une définition de travail de collocation qui, d'un côté, aide à distinguer les collocations des associations libres, et de l'autre, les collocations des phrases idiomatiques :

« La collocation est une séquence polylexicale qui actualise un mot dans une unité sémantico-syntaxique typique et qui se caractérise par le sémantisme transparent de la base et le sémantisme restreint du collocatif ».

La collocation se caractérise par le sémantisme transparent de la base : la base garde toujours son sens littéral. Le collocatif, par contre, peut avoir un sens transparent ou opaque, mais il a toujours un sémantisme restreint parce que l'association avec sa base est exclusive ou quasi-exclusive dans un sens spécifique (voir chapitre 4 pour des exemples), étant donné qu'elle est soumise à des contraintes de nature sémantique imposées par l'arbitraire de l'usage linguistique.

Comme nous l'avons expliqué auparavant, une importante considération à faire concerne les contraintes agissant au niveau syntaxique (par exemple le blocage de

certaines propriétés transformationnelles) : elles ne sont pas systématiques et ne définissent pas un type d'unité phraséologique par rapport à l'autre car une forte variabilité est présente. Par exemple, les phrases idiomatiques peuvent avoir une certaine flexibilité syntaxique, bien que ce cas ne soit pas très fréquent.

En résumé, nous avons considéré la coprésence des deux traits du sémantisme transparent de la base et du sémantisme restreint du collocatif comme caractérisant la collocation par rapport aux combinaisons libres et aux phrases idiomatiques. En effet, le sémantisme restreint du collocatif est un trait que la collocation a en commun avec les phrases idiomatiques, tandis que la transparence de la base est un trait en commun avec les combinaisons libres. Cela veut dire que seules les collocations regroupent ces deux traits.

Considérons de nouveau les associations fondamentales du mot *conférence* en nous servant d'exemples extraits du corpus. Nous allons montrer comment le figement de l'association influence le choix des locuteurs, à qui nous avons demandé de sélectionner les associations considérées comme fondamentales. Notamment, nous avons remarqué que :

1) Les associations très fréquentes mais avec un score faible (dans la partie centrale supérieure du tableau VIII) correspondent à des associations plutôt libres :

    a) *Avoir une conférence*

    Cette association est peu sélectionnée parce qu'il s'agit, visiblement, d'une association transparente et compositionnelle.

    Ex. : *Nous avons une conférence nationale sur l'activité et l'organisation des unions locales en novembre.*

    L'analyse approfondie des concordances nous a fait aussi remarquer que parfois les cooccurrences fréquentes sont fortuites : la valeur élevée de fréquence que nous avons repérée pour le cooccurrent *avoir* en

association avec le pivot *conférence* n'est pas réelle, puisqu'elle est affectée par l'usage très répandu de *avoir* en tant qu'auxiliaire dans des phrases du type :

Ex. : *Il a, le 26 juin, consacré sa conférence de presse au Nouveau Plan National Informatique.*

Ex. : *Avant de disparaître en 1999, il a présidé la conférence internationale en intelligence artificielle.*

b) *Le président de la conférence*

Cette association, comme la précédente, est transparente et compositionnelle.

Ex. : Le président de la conférence épiscopale, Mgr Alejandro Goic, a déclaré « nous ne pouvons vivre ancrés dans le passé ».

Dans ce cas également, la valeur de fréquence que nous avons repérée est artificiellement élevée, puisque l'analyse des concordances repère souvent l'association *la conférence du président* plutôt que *le président de la conférence*[66].

Ex. *Conférence de presse du Président de la République.*

*Ex. Cette position est commune à la FHF, aux conférences des présidents de CME et des directeurs de CHU.*

[66] Nous signalons ici encore la présence d'un certain bruit pour mettre en évidence la nécessité d'interroger les locuteurs natifs : ce qui nous permet de distinguer la fréquence repérée par la statistique (automatiquement) de la fréquence telle qu'elle est perçue des enquêtés (subjectivement).

En outre, l'association *le président de la conférence* ne semble pas avoir de statut en tant qu'unité puisqu'elle implique toujours une spécification de *conférence*.

Ex. *Monsieur le Président de la conférence des bâtonniers, Mesdames et Messieurs les bâtonniers.*

Ex. *Charges électives : Président de la conférence des directeurs d'IUFM de 1994 à 1998.*

c) *La conférence de rentrée*

Cette troisième association pourrait sembler, lors d'un premier examen, semi-figée.

Ex. *Et lors de sa conférence de presse de rentrée le 8 janvier, le chef de l'Etat avait assuré qu'« avec Carla, c'est du sérieux », assurant même aux journalistes qu'il y avait toutes les chances pour qu'ils « n'apprennent (un mariage) que le lendemain ».*

En effet, en français, les compléments déterminatifs du nom, lorsqu'ils sont relationnels (c'est-à-dire lorsqu'ils spécifient son extension) donnent souvent naissance à des collocations. Cependant, à travers un test de substitution synonymique, nous pouvons remarquer que l'association est, elle aussi, plutôt libre, car son sémantisme n'est pas restreint : *de rentrée* trouve des synonymes en *de début d'année*, *de reprise*, *inaugurale*. En outre, l'insertion d'autres éléments entre la base et le collocatif est possible, comme nous le montrons dans l'exemple suivant :

Ex. *Monsieur CURIEN, pressenti pour la conférence solennelle de rentrée novembre 2000, fait savoir avec regrets que son emploi du temps surchargé ne lui permet pas d'accepter cette invitation.*

2) Les associations peu fréquentes mais avec un score élevé (dans la partie inférieure du tableau VIII) correspondent à des associations plus figées.

a) *Une conférence téléphonique*

*Conférence téléphonique* est une association très utilisée aujourd'hui pour indiquer une « téléconférence », c'est-à-dire une conversation partagée par plusieurs personnes sur une même ligne de téléphone. A notre avis, il s'agit d'une unité qui comporte un degré de figement plus grand que celui présent dans la collocation : les deux composants, *conférence* et *téléphonique,* produisent une véritable unité de sens, très similaire à *vidéoconférence*, qui à son tour est synonyme de *visioconférence*, mot signalé par beaucoup de locuteurs natifs qui ont exécuté notre test.

b) *Le compte-rendu de la conférence*

Admettons que, dans ce cas, nous n'avons pas affaire au même type de figement que dans celui qui concerne *conférence téléphonique* et qu'il ne s'agit même pas d'une collocation au sens linguistique du terme, car le sens de l'association est compositionnel.

Ex. *Cet article fait partie du compte-rendu de la conférence sur la politique européenne en matière de propriété intellectuelle, qui s'est tenue le 10 mars 2005 à Copenhague, au Danemark.*

Cependant, nous pouvons remarquer que l'association est soumise à une restriction de sélection parce que *compte-rendu* est le seul terme approprié pour désigner le relevé de conclusions d'une conférence : *compte-rendu* ne se laisse pas remplacer par *récit* ou *exposé,* et limite sa substituabilité synonymique à *rapport.* Le sémantisme restreint du collocatif fait apparaître l'association comme une unité plutot figée et préfabriquée. Selon nous, seul ce trait nous aide à distinguer les collocations dont les

composants sont transparents et où aucune restriction syntaxique n'est présente, des combinaisons libres fréquentes mais qui ne sont pas soumises à des restrictions sémantiques.

c) *Une conférence intergouvernementale*

*L'association conférence intergouvernementale* semble également être une association libre en raison de sa transparence. En outre, l'écart entre la valeur de la fréquence et le score des natifs n'est pas aussi fort que celui dans *conférence téléphonique* et *compte-rendu de la conférence*. Cependant, nous remarquons qu'aucun synonyme ne remplace *intergouvernementale* et que les considérations faites pour *le compte-rendu de la conférence* valent également dans ce cas. Voici un exemple extrait du corpus :

Ex. *Dans la perspective de la conférence intergouvernementale qui a été chargée, en 1996, de la mise au point d'une politique de sécurité commune, il semble en effet opportun d'avoir une réflexion rétrospective.*

Cette même analyse, menée pour tous les autres pivots, a toujours confirmé l'existence de ce trait, même s'il opère de façon plus ou moins marquée selon l'unité polylexicale traitée**.**

## 7.3 Repérage des associations fondamentales

Parmi toutes les associations extraites automatiquement et soumises au jugement des locuteurs natifs, nous ne repérerons que celles qui finalement peuvent être considérées comme fondamentales (paragraphe 7.3.1) ; ensuite, nous évaluerons la valeur des mesures associatives pour l'extraction des collocations (paragraphe 7.3.2).

*7.3.1 La liste des associations fondamentales*

Comme nous l'avons remarqué, le facteur dont dépend en général l'attribution du caractère fondamental est la fréquence des associations. Cependant, le test soumis aux natifs a montré l'existence d'un facteur qui intervient dans le choix des locuteurs lorsque la fréquence ne joue pas de role : le degré de figement. Ce test nous a permis de repérer les associations fondamentales de chaque pivot : nous avons sélectionné seulement les associations, fréquentes ou non fréquentes, retenues par les enquêtés.

Dans le **tableau X**, nous présentons les associations fondamentales de *conférence* triées par fréquence. Nous surlignons **en gris** les associations considérées comme fondamentales. Si dans la section précédente nous n'avions marqué, à titre d'exemple, que les points singuliers les plus nets, ici nous tentons de marquer (toujours en rouge) tous les points qui semblent avoir un écart significatif entre fréquence et score obtenu par les locuteurs. Ensuite, nous considérons comme fondamentales les associations les plus fréquentes sauf les points de score faible, et seules les associations non fréquentes avec un score élevé. La ligne noire horizontale marque le seuil de significativité de la fréquence (désormais s.s.f.[67]) établi au moment de l'extraction. Pour *conférence,* le s.s.f. était 1 043. Ce seuil sépare les unités fréquentes (**F**) des unités non fréquentes mais pertinentes (éventuellement disponibles, indiquées avec **D**) que nous avions repérées comme candidates au statut de collocations fondamentales (cf. chapitre 6).

[67] Le s.s.f. varie selon le pivot traité, car chaque pivot a sa propre fréquence d'occurrence dans le corpus.

**Tableau X** – Pivot *conférence* : liste des associations fondamentales, surlignées en gris (seuil de significativité de la fréquence : 1043). En rouge, les points singuliers.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | une * **de presse** | 14207 | 97% |
| | un **maître de** *s | 14106 | 87% |
| | **être à** une * | 10418 | 57% |
| | **avoir** une * | 8069 | 20% |
| | **organiser** une * | 5589 | 70% |
| | une * **internationale** | 3729 | 73% |
| | **donner** une * | 3487 | 83% |
| | un **cycle** de *s | 3111 | 40% |
| | la **salle** de * | 3045 | 93% |
| | le **thème** de la * | 2478 | 77% |
| | **au cours de la** * | 2439 | 63% |
| | **tenir** une * | 2045 | 57% |
| | le **programme** de la * | 1749 | 50% |
| | **faire** une * | 1674 | 37% |
| | **animer** une* | 1331 | 73% |
| | **participer à** une * | 1233 | 73% |
| | le **président** de la * | 1203 | 17% |
| | **à l'occasion de la** * | 1166 | 53% |
| | **assister à** une * | 1130 | 93% |
| | la * **de rentrée** | 1053 | 7% |
| | une * **annuelle** | 1043 | 43% |
| D | l'**animation** de la * | 809 | 27% |
| | une * **de consensus** | 790 | 7% |
| | une * **épiscopale** | 609 | 17% |
| | **prononcer** une* | 608 | 13% |
| | la * **s'intitule ...** | 595 | 23% |
| | une * **mensuelle** | 569 | 27% |
| | une * **ministérielle** | 524 | 40% |
| | une * **intergouvernementale** | 492 | 40% |
| | une * **téléphonique** | 459 | 83% |
| | le **compte-rendu** de la * | 436 | 70% |
| | la **tenue** de la * | 382 | 10% |
| | une * **inaugurale** | 248 | 23% |
| | une * **tripartite** | 147 | 17% |
| | une * **introductive** | 102 | 17% |

Nous souhaitons préciser que la liste rédigée a une valeur indicative parce qu'il n'est pas simple de repérer un s. s. f. exact, ni de repérer les points singuliers (sauf quand l'écart entre fréquence et score est très large). Par exemple, l'association *faire une conférence,* sélectionnée par 37% des locuteurs et avec une fréquence de 1674 représente un cas intermédiaire : nous pourrions la considérer comme un point singulier pour le score faible, mais étant donné que la valeur de la fréquence dépasse le seuil de significativité, nous la considérons comme une association fondamentale, car nous adoptons une perspective didactique. En effet, dans des cas plus obscurs, le choix de ce qui est fondamental dépend du but que l'on poursuit.

Nous renvoyons aux **ANNEXES B** pour la liste des associations fondamentales, triées par fréquence, de tous les pivots (avec leurs s.s.f).

### *7.3.2 Collocations et mesures statistiques*

Comme nous l'avons déjà précisé, dans ce travail, nous avons souvent utilisé les termes « association » ou « unité polylexicale », plutôt que celui de « collocation », pour indiquer les associations extraites. En effet, nous avons constaté parmi elles la présence tant de véritables collocations que d'unités phraséologiques libres, ces dernières ayant la préférence du locuteur parce qu'elles sont perçues comme essentielles pour la communication. Etant donné que nous avons étudié un phénomène important et non marginal comme celui de la collocation fondamentale, et que nous avons combiné la fréquence aux mesures associatives, un tel résultat était attendu. D'ailleurs, la présence parmi les associations fondamentales de véritables collocations ainsi que d'associations libres possède un certain intérêt didactique (cf. chapitre suivant).

Nous tenons également à signaler, à l'inverse, l'absence d'unités très figées telles que les phrases idiomatiques, les locutions et les pragmatèmes. Cela s'explique, en premier lieu, par la nature du corpus, qui ne concerne pas la langue orale (et le langage informel) dont ces unités sont plus caractéristiques ; et en second lieu, par la méthode d'extraction, car la cohésion de ce type d'unités n'est pas toujours bipolaire comme dans les collocations, mais elle est liée à une séquence figée de plus de deux mots. C'est pourquoi ces mesures d'association ne les identifient pas. Leur absence

ne doit donc pas être interprétée comme ayant une importance moindre dans l'univers phraséologique fondamental. Tout le monde s'accordera à reconnaître que les pragmatèmes jouent au contraire un rôle fondamental dans les premières étapes de l'apprentissage d'une langue. Bien qu'ils soient parfois opaques et restreints syntaxiquement (considérons par exemple le pragmatème très fréquent *Ça va ?*), ils constituent le premier répertoire communicatif de tout locuteur.

Ce constat étant fait, considérons à présent la valeur des mesures associatives pour le repérage des collocations fondamentales. Pour ce faire, nous devons forcément prendre en compte la spécificité de notre corpus, un corpus issu du Web. Malgré le fait que notre corpus puisse être considéré comme adéquat dans le cadre de notre analyse en ce qui concerne sa taille et sa représentativité de la langue générale, nous admettons qu'il peut contenir des biais, à cause de la présence de pages Web qui se répètent et qui proposent les mêmes contenus. En effet, comme nous le savons, la communication sur le Web est très puissante, et elle trouve sa force dans le fait qu'un nombre énorme d'internautes partagent des informations et en discutent en même temps. Cette forme de communication tant efficace que redondante explique la présence, parmi les collocatifs extraits, d' « intrus » : ce sont en général des entités nommées, des noms propres, des noms géographiques, qui n'entrent pas dans notre définition des collocatifs. Pour atténuer ces biais, nous avons croisé la fréquence avec les mesures associatives et la valeur de dispersion.

Dans ce qui suit, nous allons essayer de comprendre comment ces différentes mesures ont pu contribuer au repérage des collocations fondamentales.

1) IM et t-score

Ces deux mesures nous semblent les moins utiles au repérage des collocations fondamentales. En effet, nous avons remarqué qu'elles repèrent davantage des associations très rares ou celles dont la dispersion possède une valeur basse, des *bogues* et des entités nommées.

Dans les **tableaux XI** et **XII** nous présentons les résultats extraits pour le pivot *rencontre* (avant le nettoyage des sorties). L'IM extrait un seul collocatif significatif, l'adjectif *nationale*, tandis que le t-score n'en extrait aucun. Les deux mesures extraient, par contre, des collocatifs tels que *jacques*, *gilles* et *goody* qui ne présentent aucun intérêt en tant que collocatifs de *rencontre* :

**Tableau XI** –Pivot *rencontre* : les 10 premiers résultats selon l'IM

| Collocatif | f | logLike | t-score | z-score | IM | domaines |
|---|---|---|---|---|---|---|
| blaesheim | 158 | 3678,084 | 4233,854 | 12,56969 | 11,63916 | 1 |
| goody | 161 | 3530,786 | 3050,605 | 12,68836 | 10,96482 | 1 |
| franche | 69 | 1491,893 | 1848,941 | 8,306456 | 10,81067 | 5 |
| jacques | 194 | 4169,54 | 3001,335 | 13,92809 | 10,74581 | 147 |
| bordeaux | 169 | 3585,579 | 2614,556 | 12,99968 | 10,60785 | 88 |
| averroès | 50 | 1048,87 | 1339,802 | 7,070871 | 10,48859 | 15 |
| udf-modem | 413 | 8593,453 | 3688,912 | 20,32178 | 10,40279 | 2 |
| méditerranée | 45 | 934,4993 | 1205,818 | 6,707996 | 10,38323 | 27 |
| Gilles | 88 | 1823,527 | 1667,388 | 9,380535 | 10,36075 | 44 |
| nationale | 44 | 911,7548 | 1179,021 | 6,63304 | 10,36075 | 13 |

**Tableau XII** – Pivot *rencontre* : les 10 premiers résultats selon le t-score

| Collocatif | f | logLike | t-score | z-score | IM | domaines |
|---|---|---|---|---|---|---|
| blaesheim | 158 | 3678,084 | 4233,854 | 12,56969 | 11,63916 | 1 |
| udf-modem | 413 | 8593,453 | 3688,912 | 20,32178 | 10,40279 | 2 |
| goody | 161 | 3530,786 | 3050,605 | 12,68836 | 10,96482 | 1 |
| jacques | 194 | 4169,54 | 3001,335 | 13,92809 | 10,74581 | 147 |
| paris | 910 | 16266,55 | 2629,164 | 30,16224 | 8,935661 | 457 |
| bordeaux | 169 | 3585,579 | 2614,556 | 12,99968 | 10,60785 | 88 |
| clae | 385 | 7268,501 | 2199,366 | 19,61986 | 9,438765 | 1 |
| franche | 69 | 1491,893 | 1848,941 | 8,306456 | 10,81067 | 5 |
| gilles | 88 | 1823,527 | 1667,388 | 9,380535 | 10,36075 | 44 |
| averroès | 50 | 1048,87 | 1339,802 | 7,070871 | 10,48859 | 15 |

2) Log-likelihood

Le Log-likelihood favorise les paires fréquentes et n'extrait pas les associations peu dispersées ; et il repère moins de *bogues* et d'entités nommées que les deux mesures précédentes. Cependant, cette mesure n'est pas non plus capable de réduire significativement le bruit. Ci-dessous, dans le **tableau XIII**, les dix premières associations repérées par le log-likelihood pour *rencontre* :

**Tableau XIII** – Pivot *rencontre* : les 10 premiers résultats selon le log-likelihood

| Collocatif | f | logLike | t-score | z-score | IM | domaines |
|---|---|---|---|---|---|---|
| organiser | 6064 | 18932,41 | 243,261 | 71,20053 | 2,457269 | 2853 |
| lieu | 6839 | 17118,54 | 209,324 | 72,71799 | 2,114589 | 3495 |
| paris | 910 | 16266,55 | 2629,164 | 30,16224 | 8,935661 | 457 |
| échange | 3693 | 13015,48 | 215,8616 | 56,59305 | 2,677503 | 2082 |
| françois | 1110 | 12306,93 | 709,2632 | 33,24347 | 6,120737 | 158 |
| occasionnel | 1593 | 12053,9 | 413,0217 | 39,54649 | 4,692046 | 76 |
| être | 18786 | 9282,108 | -82,376 | -110,77 | -0,59232 | 6599 |
| udf-modem | 413 | 8593,453 | 3688,912 | 20,32178 | 10,40279 | 2 |
| bayrou | 846 | 8185,637 | 479,605 | 28,97988 | 5,612722 | 31 |
| clae | 385 | 7268,501 | 2199,366 | 19,61986 | 9,438765 | 1 |

3) Fréquence et z-score

La fréquence et le z-score apparaissent par contre comme étant des mesures plus précises pour le repérage des collocations fondamentales. Pour *rencontre*, par exemple, les deux mesures repèrent une seule entité nommée, *premier* (qui dans l'analyse des concordances est le plus souvent suivie de *ministre)*, comme nous le présentons dans les **tableaux XIV** et **XV**:

**Tableau XIV** – Pivot *rencontre* : les 10 premiers résultats selon le z-score

| Collocatif | f | logLike | t-score | z-score | IM | domaines |
|---|---|---|---|---|---|---|
| lieu | 6839 | 17118,54 | 209,324 | 72,71799 | 2,114589 | 3495 |
| organiser | 6064 | 18932,41 | 243,261 | 71,20053 | 2,457269 | 2853 |
| échange | 3693 | 13015,48 | 215,8616 | 56,59305 | 2,677503 | 2082 |
| *premier* | *5125* | *3860,482* | *75,04225* | *45,39366* | *1,005358* | *2555* |
| aller | 5175 | 3370,762 | 68,84857 | 43,3793 | 0,923854 | 2631 |
| occasion | 2331 | 5197,569 | 111,0772 | 41,53101 | 1,96757 | 1387 |
| international | 2777 | 4165,384 | 88,52255 | 41,25293 | 1,52707 | 1119 |
| occasionnel | 1593 | 12053,9 | 413,0217 | 39,54649 | 4,692046 | 76 |
| annoncer | 1985 | 4227,518 | 98,64848 | 37,95722 | 1,910206 | 285 |
| national | 2979 | 2377,104 | 59,54095 | 35,34579 | 1,04297 | 804 |

**Tableau XV** – Pivot *rencontre* : les 10 premiers résultats selon la fréquence

| Collocatif | f | logLike | t-score | z-score | IM | domaines |
|---|---|---|---|---|---|---|
| être | 18786 | 9282,108 | 82,37597 | 110,7695 | 0,592316 | 6599 |
| avoir | 14079 | 1927,412 | 39,96513 | 47,25839 | 0,335246 | 4697 |
| lieu | 6839 | 17118,54 | 209,324 | 72,71799 | 2,114589 | 3495 |
| organiser | 6064 | 18932,41 | 243,261 | 71,20053 | 2,457269 | 2853 |
| faire | 5958 | 1,422598 | 1,173063 | 1,182011 | 0,015197 | 2887 |
| aller | 5175 | 3370,762 | 68,84857 | 43,3793 | 0,923854 | 2631 |
| *premier* | *5125* | *3860,482* | *75,04225* | *45,39366* | *1,005358* | *2555* |
| échange | 3693 | 13015,48 | 215,8616 | 56,59305 | 2,677503 | 2082 |
| site | 3069 | 534,3745 | 24,88989 | 19,91883 | 0,445592 | 895 |
| national | 2979 | 2377,104 | 59,54095 | 35,34579 | 1,04297 | 804 |

4) Dispersion

La dispersion, parfois négligée dans les études sur les collocations, se révèle être une mesure cruciale pour le repérage des collocations fondamentales. Etant donné que nous nous sommes intéressée à un phénomène central dans la langue, il était essentiel d'éliminer les fausses valeurs de fréquence. La dispersion mesure l'homogénéité de diffusion d'une collocation dans le corpus. Le fait qu'une collocation ne soit fréquente que dans une section du corpus est révélateur de son usage dans des contextes restreints, et suggère son exclusion de l'inventaire des collocations fondamentales. La suppression des associations présentes dans peu de sources a été permise grâce à l'élimination de toutes les cooccurrences ayant une valeur de dispersion égale ou inférieure à 50 sources. Ce seuil a été établi d'après le constat que, avec un seuil de dispersion inférieur à 50, presque aucune cooccurrence n'est correcte ou considérée comme acceptable. Parmi les opérations de réduction du bruit présent dans le corpus, celle-ci est la plus consistante car elle permet de réduire la liste des sorties d'environ 1 000 résultats, ce qui représente un chiffre conséquent. Considérons deux exemples très significatifs : l'association *fêtes panathénées,* à partir du mot pivot *fête,* a probablement été repérée (par la fréquence et surtout par les mesures associatives) puisque les jours où l'extraction du corpus du web a été effectuée correspondait avec la célébration de ces fêtes, qui étaient par conséquent très discutées et commentées par la communauté Internet. La valeur très basse de la dispersion nous indique immédiatement qu'il ne s'agit pas d'une association fondamentale. De la même façon, si nous considérons le mot pivot *fête*, nous pouvons remarquer que le cooccurrent *tarte* est sélectionné par toutes les mesures associatives, et cela avec des valeurs très élevées (l'IM, par exemple, vaut 12 : elle est quatre fois supérieure au seuil de significativité, qui vaut 3). Cependant, nous avons constaté, à travers l'analyse des concordances, que l'association de *fête* et de *tarte* était due au fait que les deux mots partageaient le même domaine sémantique (et les mêmes sites), et non pas à la présence d'un lien collocationnel. La mesure de la dispersion a confirmé notre observation car elle nous a indiqué que *tarte* n'était issu que d'une seule source. En particulier pour le pivot *fête*, la dispersion nous a permis d'éliminer les cooccurrences qui se référaient à de nombreuses fêtes différentes, par exemple *la fête du Beaujolais, les fêtes*

*panathénées,* etc., très commentées sur le web.

Pour conclure, les mesures associatives utilisées (excepté le z-score) constituent une arme à double tranchant quand il s'agit de mesurer le caractère fondamental d'une association : c'est pourquoi nous avons effectué une opération de nettoyage avant l'extraction et avons combiné différentes mesures. La fréquence, malgré ses nombreuses limites (dont nous avons largement discuté dans la présente recherche), nous semble être une mesure très utile pour l'extraction des collocations fondamentales. Enfin, la dispersion est sans doute la meilleure mesure pour éliminer le bruit dérivant des biais du corpus, et plus généralement pour l'étude de phénomènes non marginaux comme les collocations fondamentales ; elle signale, en effet, la présence de cooccurrences non pertinentes, car leur fréquence ou leur significativité (selon les mesures associatives) est calculée en prenant en compte les documents redondants du corpus.

Dans le prochain paragraphe, nous tenterons de montrer quelles peuvent être les contributions principales de notre recherche à la lumière des résultats obtenus.

## 7.4 Apports de l'analyse

Ici nous allons donner un aperçu critique des deux résultats principaux de notre étude, afin d'en éclairer rétrospectivement les apports.

1) Existence d'une corrélation positive non systématique entre fréquence et caractère fondamental attribué par les locuteurs.

Ce premier résultat de notre recherche nous a montré que la corrélation entre fréquence des associations et score obtenu par les locuteurs natifs n'était pas toujours présente, et si elle l'était, elle était faible et non systématique, à cause de la présence de cas particuliers qui s'écartent de la norme. Ces cas représentent des associations très fréquentes mais peu sélectionnées par les locuteurs, et inversement des associations plus rares mais souvent sélectionnées. Or, ce premier résultat semble déjà apporter une contribution non négligeable à l'étude des collocations fondamentales. En effet, il est vrai que les auteurs du *Français Fondamental* avaient, de façon très pertinente, compris que le caractère fondamental ne dépendait pas uniquement de la fréquence car il existait des mots fondamentaux dits

« disponibles » qui, bien que plus rares, étaient tout aussi importants dans la communication quotidienne. Mais il est également vrai qu'ils avaient pris pour acquis que les mots fréquents étaient des mots fondamentaux. Or, si cela peut être le cas pour les mots considérés comme éléments simples, cela peut être différent dans le cas d'unités polylexicales. En effet, nous avons constaté que parfois des unités polylexicales très fréquentes ne sont pas considérées comme fondamentales par les locuteurs.

2) Rôle du figement dans l'attribution du caractère fondamental.

Le second résultat de notre recherche montre que si la fréquence ne joue pas de rôle dans l'attribution du caractère fondamental aux unités polylexicales, il faut en chercher la raison dans leur figement. Cet aspect n'a pas non plus été exploré par les auteurs du *Français Fondamental* (Gougenheim et al., 1964), car ils n'ont pas réellement cherché à déterminer la cause du caractère fondamental attribué aux mots (comme cela a été fait, par exemple, par Thornton, Iacobini, Burani, 1997 pour le vocabulaire de base de la langue italienne). Voici donc ce qui pourrait être une contribution importante de notre travail : la découverte d'un facteur autre que la fréquence dont dépendrait l'attribution du caractère fondamental.

Dans le dernier chapitre nous tenterons de comprendre comment la notion de collocation fondamentale peut être exploitée en didactique.

# CHAPITRE 8
# Conclusions : contributions de l'étude et implications pour le FLE

Dans le cadre de cette thèse, nous nous sommes intéressée aux collocations fondamentales en essayant de les définir dans une étude de corpus qui prend en compte la fréquence et la dispersion, les mesures associatives, ainsi que la perception que les locuteurs natifs ont de la langue. Nous avons en outre dérivé des implications pour un traitement adéquat des collocations dans le cadre de la didactique de la langue étrangère.

Dans le premier chapitre, nous avons présenté la notion de vocabulaire fondamental et nous avons décrit les premières tentatives de production de listes de base, simplifiées ou réduites. Nous avons concentré notre attention sur l'ouvrage de référence du vocabulaire de base de la langue française, le *Français Elementaire* (1954), publié dans les années 50 du siècle dernier par Gougenheim et son équipe. Ensuite, nous avons souhaité élargir l'étude du vocabulaire de base à la dimension syntagmatique. Après avoir présenté un état de l'art raisonné et critique des études les plus significatives du fascinant domaine de la phraséologie et plus particulièrement de la collocation (dont nous avons donné une définition de travail), nous avons exposé le problème de la compréhension de la nature de la collocation fondamentale. Nous avons souligné la nécessité d'intégrer à l'étude purement statistique des unités polylexicales, la connaissance partagée que les locuteurs natifs en ont. Nous avons analysé, d'une part, le caractère fondamental issu de la statistique (la fréquence textuelle intégrée à la dispersion et aux mesures associatives) ; et d'autre part, le caractère fondamental issu de l'intuition des locuteurs natifs. Nous avons ainsi remarqué que les deux aspects ne divergent pas, mais représentent au contraire deux visions complémentaires : si la fréquence ne détermine pas le caractère essentiel de la langue, d'autres facteurs interviennent dans

le choix des locuteurs. Plus précisément, en premier lieu, nous avons découvert la présence d'une corrélation non systématique entre fréquence et caractère fondamental attribué par les locuteurs natifs aux associations. Cela est dû à la présence de cas particuliers qui s'écartent de la norme. En second lieu, nous avons compris que le figement de l'association peut jouer un rôle dans l'attribution du caractère fondamental : si la fréquence n'oriente pas les locuteurs dans leur sélection des unités polylexicales fondamentales, la raison est à chercher dans leur nature plus ou moins figée. Ce parcours nous a également amené à évaluer l'outil d'extraction développé et les mesures auxquelles nous avons eu recours. La plupart des unités polylexicales extraites automatiquement sont ont été reconnues par les enquêtés, car nous avons constaté une certaine correspondance entre fréquence des associations et sélection des locuteurs (notre division entre unités polylexicales fréquentes et non fréquentes suit la progression du score obtenu par les locuteurs). D'ailleurs, l'analyse des faits objectifs devrait pouvoir correspondre partiellement aux jugements des locuteurs natifs, bien que nombreux soient ceux qui ne sont pas d'accord avec cette affirmation. Sinclair (1991, p. 4), par exemple, explique que l'avènement de la linguistique de corpus a complètement révolutionné l'étude de la langue en faisant émerger le contraste existant entre l'introspection des locuteurs et les usages réels.

Nous ne pouvons que souscrire à une telle affirmation, mais à condition que l'on précise que, pour trouver une correspondance entre usages réels et introspection des locuteurs, la méthode d'investigation choisie a une importance cruciale et que les locuteurs, étant non-spécialistes, ne devraient pas être interrogés de façon explicite sur des phénomènes de nature strictement linguistique qui sont en dehors de leur compréhension. En complément à l'analyse statistique, le fait de faire passer des tests à des locuteurs natifs s'est révélé un bon choix : les données extraites par la fréquence étaient très dispersées, avec quelques associations très fréquentes, la plupart des associations ayant des valeurs de fréquence assez proches, formant un véritable *cluster*, comportant cependant quelques cas singuliers. Cette disposition des données n'aurait pas été facile à interpréter si nous n'avions pas pensé à interroger les natifs.

Après avoir évalué l'intérêt de notre recherche, considérons l'objectif de ce dernier chapitre :

- nous reviendrons à la notion de « collocation fondamentale » dans la perspective du FLE et nous formulerons quelques implications de caractère général (paragraphe 8.1)
- nous préciserons la structuration du domaine sémantique des « événements sociaux » à partir de l'analyse des collocations fondamentales repérées (paragraphe 8.2)
- nous tirerons des considérations finales (paragraphe 8.3).

## 8.1 La notion de collocation fondamentale dans la perspective du FLE : implications de caractère général

Nous voulons ici revenir brièvement sur l'importance des collocations fondamentales pour la didactique de la langue étrangère.

Pourquoi les collocations seraient-elles si importantes pour l'apprentissage linguistique, et par conséquent pour la didactique ? Comme l'explique Granger (2008), dans l'apprentissage d'une langue étrangère, la connaissance des usages stéréotypés de la langue est sans doute un obstacle majeur. Les mots s'associant entre eux selon des préférences combinatoires, les locuteurs s'expriment le plus souvent en morceaux de langue typiques. L'unité polylexicale connue comme « collocation » reflète la connaissance naturelle qu'un natif a de sa langue maternelle. Il est notamment connu que des traductions offertes par un locuteur étranger qui sembleraient formellement acceptables ne correspondent pas forcement au choix effectué par le locuteur natif, choix qui repose sur des critères idiosyncratiques liés à l'usage : même entre deux langues sœurs comme le français et l'italien, il peut y avoir de fortes différences dans les structures lexico-grammaticales choisies pour exprimer un concept donné, et le plus souvent l'apprenant produit des associations peu fréquentes ou fausses à cause de la transposition de sa langue maternelle (sur la traduction français-italien, cf. Velez, 2007, 2009). La connaissance des préférences combinatoires de la langue cible introduit l'apprenant dans la structure de la nouvelle langue et l'aide, malgré le caractère arbitraire de ce type d'association, à construire un modèle mental des associations les plus typiques.

Ce constat étant fait, quelle est l'utilité de la notion de « collocation fondamentale » dans la didactique du FLE ? Même si aujourd'hui l'enseignement du lexique en contexte est une démarche unanimement reconnue, les matériaux didactiques existants ne témoignent pas de cette prise de conscience, et il reste encore des doutes « opérationnels » sur le choix du lexique à enseigner. Le lexique étant un ensemble très vaste qui regroupe des millions de mots, une sélection des contenus s'impose donc. Notre étude se révèle, en ce sens, de forte utilité, car les collocations fondamentales sont les collocations les plus importantes pour la réussite de l'interaction quotidienne. D'après les résultats obtenus, il nous semble possible de présenter quelques principes de caractère général en vue d'intégrer les collocations fondamentales dans la didactique du FLE et afin de mieux structurer le choix du lexique en FLE. Il ne s'agit pas d'indiquer une méthodologie didactique précise, mais de présenter des critères-guide pour que l'enseignant puisse décider quelles réalisations contextuelles, et en particulier quelles collocations, privilégier pour un mot donné[68].

1) L'enseignant qui opère un choix du lexique à enseigner devrait prendre en compte l'importance de la fréquence, d'une part ; et d'autre part la présence de mécanismes d'analyse linguistique qui influencent les préférences des locuteurs dans la sélection des items à apprendre.

L'importance de la fréquence pour la sélection du vocabulaire est reconnue par la plupart des études sur le lexique. Comme nous l'avons également mis en évidence dans ce travail, la fréquence est liée à l'usage, elle facilite les procès de mémorisation et favorise des items cognitivement prioritaires. Sur la base de cette observation, l'enseignant ne devrait pas oublier, dans la construction d'un inventaire lexical, d'y inclure les mots les plus fréquents ainsi que leurs associations. Cependant, il ne faut pas tomber dans le piège de traduire le sens de « fondamental » *uniquement* par « fréquent », car cela signifierait fonder l'enseignement du vocabulaire seulement sur des facteurs quantitatifs. Dans notre étude syntagmatique, nous avons pu confirmer l'intuition gougenheimienne sur l'existence d'items plutôt rares qui sont à considérer

[68] Nous précisons que les implications formulées se réfèrent à cette étude spécifique, même si nous espérons qu'elles pourront être généralisées par des approfondissements futurs.

comme fondamentaux pour leur grande utilité communicative. Nous avons voulu prolonger son analyse et nous avons aussi cherché la cause de l'attribution du caractère fondamental de la part des locuteurs natifs. Nous avons découvert que lorsque la fréquence ne joue pas de rôle, c'est le figement de l'unité polylexicale qui intervient à sa place. Dans ce cas, des associations peu fréquentes mais plutôt figées sont retenues par les locuteurs, et à l'inverse, des associations très fréquentes mais libres ne présentent aucun intérêt pour les enquêtés. Cela suggère que le figement d'une unité peut faciliter ou entraver son apprentissage. Si l'unité figée est peu transparente, l'explication de la part de l'enseignant du mécanisme (de translation ou métaphorique) qui est à la base de l'association de deux ou plusieurs composants peut favoriser le mécanisme de mémorisation.

2) L'enseignant devrait présenter toutes les associations privilégiées d'un mot, sans opérer un filtrage sur la base d'une différenciation entre les types d'unités phraséologiques.

Bien que notre outil d'extraction ait été paramétré spécifiquement pour le repérage des collocations, l'échantillon que nous avons obtenu ne se limite pas aux collocations, mais inclut également de nombreuses associations libres, et même des locutions nominales du type « à l'occasion de », « au cours de ». Si cela peut représenter un point faible des mesures associatives censées repérer les collocations, ce recensement n'est cependant pas inutile pour une réflexion concernant la didactique. Pourquoi, bien que notre méthodologie ait visé à repérer uniquement les collocations, des associations libres ont-elles également été extraites ? Cela s'explique par le fait que, parfois, les composants des associations libres se sélectionnent réciproquement si souvent qu'ils semblent avoir le même statut que les collocations selon les mesures associatives (elles seraient cependant exclues malgré tout par l'approche phraséologique). D'un point de vue strictement linguistique, on pourrait objecter à leur étiquetage en tant que collocations, car aucune restriction sémantique et/ou syntaxique n'est présente. Cependant, d'un point de vue didactique, et même lexicographique, on ne pourra que reconnaitre l'importance de leur inclusion dans un inventaire lexical, car l'univers conceptuel de discours qui gravite autour d'un mot pivot donné (comme le suggère Lo Cascio, 2007) inclut tout type d'unité polylexicale, même les colligations. Ici, nous l'appelons « voisinage combinatoire ». Nous voulons

donc souligner l'importance de fonder la sélection du vocabulaire à enseigner sur l'usage, au-delà des différenciations phraséologiques possibles (associations libres, collocations, pragmatèmes etc.). Nous voulons aussi rappeler que si nous n'avions pas opéré un paramétrage spécifique pour le repérage des collocations (par exemple la création d'une *stoplist* éliminant, parmi les collocatifs, toutes les catégories vides), nous aurions même extrait des colligations, elles aussi importantes sur le plan didactique. Il en va ainsi par exemple des colligations *au chômage* et *du coup*. Dans *au chômage*, une préposition précise est sélectionnée en association avec *chômage*. *Du coup*, en tant que connecteur de conséquence très fréquent en français oral, est conçu comme une véritable unité. Pour conclure, nous croyons que, en classe de FLE et/ou dans le domaine lexicographique, une approche rigide n'est jamais synonyme de réussite. La priorité de l'enseignement est de satisfaire les besoins de communications immédiats des locuteurs étrangers. Un inventaire complet des associations fondamentales devrait être constitué sur la base du principe de l'utilité communicative.

Or, quelques précisions apparaissent nécessaires. En premier lieu, il faut remarquer que si l'on a affaire à des apprenants de niveau débutant, les unités les plus figées comme les phrases idiomatiques et les proverbes ne constituent pas un inventaire lexical adéquat (de survie). Comme on le sait, la nature de ce type d'unités fait qu'elles sont apprises lors de stades d'apprentissage successifs. D'ailleurs, pour des raisons déjà expliquées ici, elles n'ont pas été repérées par notre outil d'extraction. En second lieu, il ne faudrait pas oublier que dans la perspective de l'apprenant toute association qui n'a pas de correspondance avec sa langue maternelle est problématique. La collocation est un phénomène idiosyncratique et arbitraire : ce qui pour un locuteur natif est une association libre transparente, ne l'est parfois pas pour un locuteur étranger. Par exemple, l'association *faire un rêve,* transparente pour un locuteur français, ne l'est pas pour un locuteur anglais, car dans sa langue maternelle, le même sens peut s'actualiser dans une association de mots différente, *have a dream*. Or, une approche contrastive prenant en compte ces mécanismes agissant entre deux langues différentes serait une solution possible, voire idéale bien que pas toujours applicable, surtout si l'enseignant opère dans une classe mixte où interagissent des étudiants de différentes nationalités. Pour conclure,

si sur le plan descriptif l'exigence de désigner ces termes de façon précise est forte, dans le domaine de l'enseignement, par contre, la solution la plus pratique apparaît, répétons-le, être celle du « voisinage combinatoire », qui reconstruit les réalisations lexico-grammaticales les plus importantes.

3) L'enseignant devrait toujours tenir en compte que l'usage est marqué par une interdépendance entre lexique et grammaire.

L'enseignant devrait présenter une unité polylexicale dans ses contextes privilégiés dans l'usage, car cela permet d'insister sur les réalisations lexico-syntaxiques dans lesquelles cette unité se présente plus fréquemment. Nous allons nous attarder ici sur la fusion entre lexique et grammaire. La didactique du FFL semble peu orientée vers la reconnaissance de la priorité de l'usage sur la créativité langagière, et manque parfois de systématicité dans l'enseignement du lexique. A partir d'un mot donné, de nombreuses associations peuvent être produites (correspondant à divers signifiés et à diverses réalisations morpho-syntaxiques), mais lesquelles faudrait-il privilégier dans l'enseignement de la langue étrangère ? Est-il utile de faire mémoriser des associations aux apprenants avec toutes leurs réalisations possibles, dont ils ne se serviraient que peu, surtout s'ils ne sont qu'au début de leur parcours d'apprentissage ? La réponse est sans doute non, car comme l'explique Sinclair (1991, p. 5-6), ce qui ne correspond pas à l'usage réel n'est pas important car cela se réfère à des contextes non existants. Grammaire et lexique ne sont pas séparés, mais se fondent l'un dans l'autre, comme l'expliquent les souteneurs de l'existence d'un lexique-grammaire (par exemple, Willis, 1990). Le principe du lexique-grammaire est la clé de la didactique de la langue étrangère, et l'enseignant ne devrait pas s'attarder sur des structures qui ne correspondent pas à l'usage. L'usage garantit que ce qu'on enseigne correspond bien à ce qui est typiquement utilisé, partagé par la communauté des locuteurs natifs ; en outre, ce qui ressort de l'usage peut être renforcé par le feedback des locuteurs natifs lorsque l'apprenant s'expose à la langue cible. Comme le fait remarquer Sinclair (1991, p. 44), en anglais, par exemple, le verbe *to decline* dans le sens de « refuser » se présente toujours au passé composé comme dans l'exemple *He declined his invitation*. Il s'agit d'une structure lexico-grammaticale répandue dans l'usage et l'enseignant peut donc y insister en présentant le verbe *decline*, avec ce signifié, dans sa structure d'occurrence typique.

Comme l'explique l'auteur, à un sens correspond une structure, ainsi le sens influence la structure et vice-versa. Les sens s'actualisent dans des patterns syntaxiques spécifiques et dans des contextes d'occurrence typiques.

Dans le prochain paragraphe nous concentrerons notre attention sur le domaine sémantique traité dans ce travail.

## 8.2 Structuration du domaine sémantique des « événements sociaux »

Dans ce paragraphe, nous allons présenter la structuration du domaine sémantique qui a servi d'exemple dans notre recherche, le domaine des « événements sociaux ». Pour ce faire, nous allons traiter deux points : les traits généraux du domaine et les traits spécifiques aux pivots du domaine.

1) Traits généraux du domaine

Commençons par expliquer comment le repérage des collocations fondamentales permet de mieux cerner ce domaine et de représenter son noyau central.

Notre domaine était relativement hétérogène : en effet, *séminaire, colloque, congrès, conférence, réunion* et *rencontre* forment un groupe assez homogène, désignant des événements sociaux ayant lieu plutôt dans le domaine universitaire ; en revanche, *conversation*, *débat*, *interview* et *fête* sont moins susceptibles d'être regroupés. Cependant, il a été possible de repérer un noyau central de collocatifs communs à un nombre plus ou moins grand de mots pivots. Nous avons considéré ce noyau sémantique commun comme représentatif du domaine choisi. Nous le présentons ci-dessous dans le **tableau XVI** :

**Tableau XVI** : Collocatifs partagés par les mots pivots

| **COLLOCATIONS** | séminaire | colloque | congrès | conférence | réunion | rencontre | fête | débat | interview | conversation |
|---|---|---|---|---|---|---|---|---|---|---|
| participer | 90% | 73% | 87% | 73% | 23% | 50% | 60% | 80% | | 73% |
| avoir | 37% | 73% | 17% | 20% | | 7% | 7% | | 47% | 80% |
| organiser | 80% | 43% | 83% | 70% | 67% | 47% | 70% | 33% | | |
| être à | 73% | 23% | 63% | 57% | 87% | 7% | 17% | | | |
| faire | 30% | 50% | 20% | 37% | 20% | | 93% | | 57% | |
| annuel | 57% | 60% | 90 | 43% | 73% | 50% | 47% | | | |
| à l'occasion du/de la | 77% | 33% | 73% | 53% | | 60% | 47% | | | |
| compte-rendu | 53% | 90% | 50% | 70% | 83% | 27% | | | | |
| avoir lieu | 77% | 63% | 77% | | 30% | 90% | | 63% | | |
| national | 27% | 83% | 43% | | | 30% | 97% | 43% | | |
| thème | 77% | 17% | 70% | 77% | | 27% | | 63% | | |
| animer | 57% | | | 73% | 17% | | | 83% | | 53% |
| se dérouler | 87% | 7% | | | 43% | 50% | 50% | | | |
| prochain | 77% | 73% | 87% | | 23% | 80% | | | | |
| organisation | 73% | 37% | 83% | | 33% | 40% | | | | |
| tenir | 40% | 80% | 37% | 57% | | | | | | 43% |
| international | 57% | 27% | 70% | 73% | | 67% | | | | |
| salle | 50% | | | 93% | 80% | | 97% | | | |
| au cours du/de la | 77% | | | 63% | | | | | | 90% |
| participation | 67% | 43% | | | 20% | | | | | |
| programme | 77% | 23% | | 50% | | | | | | |
| participant | 77% | 37% | 83% | | | | | | | |
| mensuel | 40% | | | 27% | 40% | | | | | |
| journée | 67% | 80% | | | | 37% | | | | |

Plus précisément, les collocatifs que nous venons de présenter sont :

a) des verbes
   - transitifs ou intransitifs qui expriment la présence des actants à un événement social : *participer à, avoir, être à*
   - transitifs qui expriment l'action des actants sur un événement social : *organiser, faire, animer*, *tenir*
   - intransitifs qui expriment l'actualisation de l'événement social, le mot pivot étant en position de sujet : *avoir lieu, se dérouler*

b) des substantifs
   - indiquant les actants de l'événement social : *participant* ; ou résultant de l'action des actants : *organisation, participation*
   - indiquant des objets liés à l'événement social (préparation ou résultat) : *compte-rendu, programme*
   - indiquant l'objet, l'espace ou le temps de l'événement social: *à l'occasion de, thème*, *salle, au cours de, journée*

c) des adjectifs
   - relationnels ajoutant une spécification à l'événement social : *annuel, national, international, mensuel*
   - épithètes à valeur temporelle permettant d'apporter une précision : *prochain*

Cette analyse plus fine nous permet de mieux définir le concept d'événement social, que nous avons considéré comme toute occasion de rencontre avec d'autres personnes ayant lieu lors d'un processus temporel avec un début et une fin (voir chapitre 5). Nous avons également pu remarquer que les collocatifs partagés par les pivots du domaine sémantique traité sont peu spécifiques, plutôt transparents, et qu'ils s'associent généralement aux pivots selon les règles de la libre combinatoire lexicale. Enfin, nous avons constaté que *séminaire* est un « événement social » de caractère très générique par rapport aux autres pivots, avec lesquels il partage beaucoup de collocatifs.

2) Traits spécifiques aux pivots du domaine

En second lieu, nous avons pu remarquer la présence de collocatifs spécifiques à chaque mot pivot. A titre d'exemple, considérons quelques collocatifs qui peuvent

être considérés comme exclusifs aux pivots traités ou très fréquemment en association avec les pivots traités :

- *colloque : présenter, intervention, co-organisé*
- *conférence : donner, maitre, de presse*
- *congrès : adopte*[69]*, vote*[70]*, extraordinaire*
- *conversation : interrompre, au fil de, en pleine*
- *débat : relancer, cœur, télévisé*
- *fête : souhaiter, de fin d'année, joyeuses*
- *interview : accorder, intégralité, exclusive*
- *rencontre : venir à la, site de, amicale*
- *réunion : se rendre à, de section, collective*
- *séminaire*[71] *: de recherche, transversal*

Ce second type d'analyse nous a permis de repérer les collocatifs qui expriment les notions-clé liées aux pivots traités, et en même temps de les distinguer l'un de l'autre.

Nous souhaitons ici approfondir un point supplémentaire, la notion de polysémie, qui constitue un enjeu majeur de la sélection du lexique dans des buts didactiques. Dans certains cas, nous avons remarqué que le pivot a une acception plus essentielle que d'autres. Considérons par exemple *séminaire*. Ce pivot a une acception principale : tous les collocatifs repérés concernent le sens de « réunion pour étudier un problème ou un sujet » (*tenir un séminaire, séminaire de travail, séminaire de recherche*) ; les autres sens de *séminaire*, par exemple « l'établissement où l'on fait son éducation ecclésiastique » ou « le temps que l'on passe dans cet établissement », ne sont pas représentés. La même chose vaut pour *débat*, *colloque* et *réunion* : seulement le sens de « discussion autour d'un thème » (*animer le débat, débat télévisé ; la clôture du colloque, colloque international ; réunion collective, réunion préparatoire*) est représenté ; les autres acceptions, par exemple celle de « conflit intérieur », d' « entretien » et de « rassemblement d'éléments »

[69] III personne du singulier.

[70] Verbe et substantif.

[71] Pour le pivot *séminaire* nous n'avons repéré aucun collocatif verbal fondamental qui lui est exclusif.

respectivement, ne sont pas fondamentales. Le pivot *conversation* a, lui aussi, un sens primaire, celui d' « échange » (*entamer une conversation*), plus représenté que le sens d' « entretien » (*une conversation avec le jury*).

Dans d'autres cas, nous avons remarqué que les différentes acceptions d'un pivot sont toutes des notions importantes et fondamentales. Considérons par exemple le pivot *fête* : nous avons repéré *fête* tant dans l'acception de « solennité religieuse » ou de « cérémonie commémorative » (*souhaiter de bonnes fêtes, en période de fête, le comité des fêtes, fête religieuse, etc.*) que dans l'acception de « festin » ou de « bal » (*organiser une fête, fête costumée, une fête d'anniversaire, la salle des fêtes*). Ces différents sens sont également représentés. La même chose vaut pour *congrès :* le sens diplomatique et international (*le congrès adopte une réforme, congrès européen)* semble tout autant représenté que le sens académique (*congrès scientifique, le thème du congrès*). Enfin, le mot *interview* a, lui aussi, deux sens fondamentaux : il désigne l' « entretien » visant à interroger quelqu'un (*accorder une interview, donner une interview*) ; et l'« article issu de cet entretien » (*l'extrait d'une interview, publier une interview*).

Dans d'autres cas encore, nous avons remarqué qu'il existe deux acceptions majeures, l'une étant cependant plus représentée que l'autre, et une ou plusieurs acceptions mineures. Considérons par exemple le pivot *rencontre* : le sens d' « entrevue » ou de « conversation » (*une rencontre internationale, l'issue de la rencontre*) est légèrement plus représenté que le sens de « rencontrer quelqu'un » (*une rencontre inattendue, une rencontre fortuite*). A côté de ces deux sens primaires, il y a un sens secondaire : celui de « compétition sportive », moins représenté que les autres (*une rencontre sportive, disputer la rencontre*). La même chose vaut pour *conférence*. Le sens diplomatique et international (*conférence intergouvernementale, conférence ministérielle)* semble tout autant représenté que le sens académique (*assister à une conférence, le compte-rendu de la conférence)*. Le pivot *conférence* a une troisième acception, celle d'« exposé », qui est moins représentée mais tout aussi fondamentale (*maître de conférences*, *tenir une conférence*).

En conclusion, nous souhaitons préciser que dans une pédagogie sur « objectifs spécifiques », le choix des acceptions à enseigner devrait être effectué en fonction

des objectifs visés : dans une pédagogie orientée vers les métiers du sport, par exemple, il sera plus important de récupérer le sens « sportif » de rencontre.

## 8.3 Considérations finales

Les deux aspects que nous venons d'analyser, les traits généraux du domaine et les traits spécifiques aux pivots du domaine, représentent, d'un point de vue didactique, les notions les plus pertinentes pour les apprenants. Nous conclurons notre travail par une réflexion sur la façon dont le repérage des collocations fondamentales dans un domaine sémantique spécifique peut servir la didactique du FLE. Notamment, deux démarches pour l'étude du vocabulaire peuvent être envisagées : la démarche sémasiologique, de la forme au sens, qui vise à représenter toutes les informations concernant un terme ; et la démarche onomasiologique, du sens à la forme, qui vise à représenter les réalisations linguistiques d'un concept ou d'une idée. Les deux démarches, bien qu'elles peuvent sembler éloignées l'une de l'autre, nous apparaissent complémentaires. Dans le cas de l'approche sémasiologique, le repérage des collocations fondamentales permet de mieux hiérarchiser et de centrer l'apprentissage, car parmi toutes les associations possibles, seules les associations les plus pertinentes sont fournies aux apprenants ; complémentairement, dans l'approche onomasiologique, le repérage des collocations fondamentales permet, à partir d'une notion comme celle d' « événement festif », de fournir les associations les plus essentielles pour s'exprimer.

La didactique du FLE, dont l'une des priorités est la sélection du vocabulaire à enseigner, devrait selon nous exploiter la notion de collocation fondamentale, et prendre en compte dans son approche méthodologique les facteurs qui la sélectionnent. La production du matériel didactique devrait s'orienter vers cette direction. Nous espérons avoir pu démontrer qu'une approche holistique, qui intègre la statistique à l'analyse linguistique, est essentielle pour l'étude du lexique ainsi que pour contribuer à améliorer la didactique du FLE.

# BIBLIOGRAPHIE


- Allerton D., 2002, *Stretched Verb Constructions in English*, London, Routledge
- Altenberg B., Granger S., 2002, « Recent trends in cross-linguistic lexical studies*»,* dans Altenberg B., Granger S. (Ed.), *Lexis in contrast. Corpus-based approach*, Volume 7, Amsterdam et Philadelphia, John Benjamins Publishing Company, p. 3-48
- Bally C., 1951, *Traité de stylistique française*, Paris, Klincksieck
- Baroni M., Bernardini S., Ferraresi A., Zanchetta E., 2009, « The WaCky Wide Web : A Collection of Very Large Linguistically Processed Web-Crawled Corpora » dans *Language Resources and Evaluation,* 43 (3), p. 209-226
- Baroni M., Bernardini S., Ferraresi A., Picci G., 2010, « Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation » dans Xiao R. (Ed.), *Using Corpora in Contrastive and Translation Studies*, Newcastle - Cambridge, Scholars Publishing
- Bartsch S., 2004, *Structural and functional properties of collocations in English*, Thèse de doctorat, Tübingen, Gunter Narr Verlag
- Benigno V., 2007, « Il vocabolario di base : tratti costitutivi, rilevanza cognitiva e acquisizione in italiano L2 » dans Lo Cascio V. (Ed.), *Parole in rete : apprendimento e teoria nell'era elettronica*, Novara, Utet-Università, p. 151-174
- Benson M., 1985, « Collocations and Idioms » dans R. Ilson (Ed.), *Dictionaries, lexicography and language learning*, Oxford, Pergamon Press Ltd & The British Council, p. 61-68
- Benson M., Benson E., Ilson R. (Ed.), 1986a, *The BBI Dictionary of English : a guide to word combinations,* Amsterdam et Philadelphia, John Benjamins Publishing Company
- Benson M., Benson E., Ilson R., 1986b, *Lexicographic Description of English*, Amsterdam, John Benjamins Publishing Company
- Bolly C., 2005, « Au seuil du figement : séquences (semi) figées et constructions récurrentes avec le verbe à haute fréquence *prendre* en FL1 et FL2 » dans Bolly C., Klein J. R., Lamiroy B. (Ed.), *La phraséologie dans tous ses*

*états,* Actes du colloque Phraséologie 2005, 13-15 Octobre*,* Louvain-la-Neuve, Cahiers de l'Institut de Linguistique de Louvain, CILL 31. 2-4, p. 149-167

- Bortolini U., Tagliavini C., Zampolli A., 1971, *Lessico di frequenza della lingua italiana contemporanea – LIF -*, Milano, Garzanti
- Brent W., 2009, « Meaning-last vocabulary acquisition and collocational productivity » dans Fitzpatrick T., Barfield A. (Ed.), *Lexical Processing in Second Language Learners : Papers and Perspectives in Honour of Paul Meara (Second Language Acquisition),* Bristol, Multilingual Matters*,* p. 128-140
- Buchanan M. A., 1927, « A graded Spanish Word Book », dans *Publications of the American and Canadian Committees on Modern Languages*, vol. III, Toronto, University of Toronto Press
- Bybee J., Hopper P., 2001, *Frequency and the Emergence of Linguistic Structure*, Amsterdam et Philadelphia, John Benjamins Publishing Company
- Carter R., 1998, *Vocabulary. Applied Linguistic Perspectives,* 2e éd., London et New York, Routledge
- Carter R., Schmitt N., 2004, « Formulaic sequences in action. An introduction » dans Schmitt N. (Ed.), *Formulaic sequences. Acquisition, processing and use,* Amsterdam*,* John Benjamins Publishing Company, p. 1-22
- Catach N., 1984, *Les listes ortographiques de base du français (LOB). Les mots les plus fréquents et leurs formes fléchies les plus fréquentes*, Paris, Nathan-recherche
- Church K. W., Hanks P., 1990, « Word association norms, mutual information, and lexicography » dans *Computational Linguistics*, vol. 16, p. 22-29
- Clear J., 1993, « From Firth principles : computational tools for the study of collocation » dans Baker M., Francis G., Tognini-Bonelli E. (Ed.), *Text and technology,* Amsterdam, John Benjamins Publishing Company, p. 271-292
- *COBUILD English Language Dictionary,* 1987, London, Collins
- Cohen M., 1955, *Français élémentaire ? Non*, Paris, Editions sociales
- *Collins Cobuild,* 1992, London, HarperCollins Publishers
- Cowie A. P., 1981, « The treatment of collocations and idioms in learners' dictionaries » dans *Applied Linguistics 2,* p. 223-235
- Cowie A. P., Mackin R. (Ed.), 1983, *Oxford Dictionary of English Idioms*, Oxford,

Oxford University Press
- Cowie A. P., 1998, *Phraseology. Theory, Analysis, and Applications,* Oxford, Clarendon Press
- Cowie A. P., 1999, *English dictionaries for foreign learners : a history*, Clarendon Press, Oxford
- Croft W., Cruse D. A., 2004, *Cognitive linguistics*, Cambridge, Cambridge University Press
- Cruse D. A., 1986, *Lexical Semantics*, Cambridge, CUP
- D'Agostino E., 1998, « Il lessico di frequenza dell'italiano parlato e la didattica dell'italiano » dans *Quaderns d'Italià 3*, p. 9-28
- De Mauro T., 1980/2003, *Guida all'uso delle parole*, Roma, Editori Riuniti
- *Dictionnaire des fréquences. Vocabulaire littéraire des XIXe et XXe siècles*, 1971, Trésor de la langue française, Nancy, CNRS
- *Dictionnaire historique de la langue française*, 2000, Alan Rey (Ed.), Paris, Le Robert, 3 vol.
- Dubreuil E., 2008, « Collocations : définitions et problématiques », dans *Texto !*, janvier 2008, vol. XIII, n°1
- Ellis R., 1997, *Second language acquisition*, Oxford, Oxford University Press
- Ellis N. C., 2002, « Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition » dans *Studies in second language acquisition,* 24, p. 143-188
- Faucett L., Palmer H., Thorndike E. L., West, M., 1936*, Interim report on vocabulary selection for Teaching English as a Foreign language*, London, P.S. King and Son, Ltd
- Fillmore CJ., Kay P., O'Conner M.C., 1988, « Regularity and idiomaticity in grammatical constructions : the case of let alone. » dans *Language*, 64/3, p. 501-538
- Firth J. R., 1957, « A synopsis of linguistic theory, 1930-1955 » dans Firth J. R. et al., *Studies in linguistic analysis*, Special volume of the Philological Society, Oxford, Blackwell, p. 1–32
- Fontenelle T., 1998, « Discovering significant lexical functions in dictionary entries » dans Cowie A. P. (Ed.), *Phraseology. Theory, Analysis, and*

*Applications*, Oxford, Clarendon Press, p. 189-208

- Forsberg F., 2006, « Le langage préfabriqué en français parlé L2. Etude acquisitionnelle et comparative » dans *Forskningsrapporter, Cahiers de la Recherche 34,* Université de Stockholm
- Francis G., Huston S., 2000, *Pattern Grammar, a corpus-driven approach to the lexical grammar of English*, Amsterdam et Philadelphia, John Benjamins
- Frath P., Gledhill C., 2005, « Qu'est-ce qu'une unité phraséologique ? », dans Bolly C., Klein J.R., Lamiroy B. (Ed.), *La phraséologie dans tous ses états*. Actes du Colloques Phraséologie, 2005, Cahiers de l'Institut linguistique de Louvain, 31.2-4, p. 11-25
- Fuster Márquez M., Pennock Speck B., 2008, « The spoken core of british English : a dyachronic analysis based on the BNC » dans *Miscelánea : a journal of English and American studies,* vol. 37, p. 53-74
- Galisson R., 1971, *Inventaire thématique et syntagmatique du français fondamental*, Paris Hachette-Larousse, Coll. Le Français dans le Monde/BELC
- Galli de' Paratesi N., 1981, *Livello soglia per l'insegnamento dell'italiano come lingua straniera*, Strasburgo, Consiglio d'Europa
- Gläser R., 1988, « The Grading of Idiomacity as a Presupposition for a Taxonomy of Idioms » dans Hüllen, Werner & Schulze, Rainer (Ed.), *Understanding the Lexicon. Meaning, Sense and Work Knowledge in Lexical Semantics*, Tübingen, Max Niemeyer Verlag
- Goddard C., Wierzbicka A., 2007, « Semantic primes and cultural scripts in language learning and intercultural communication » dans Farzad S., Palmer G. B. (Ed.), *Applied Cultural Linguistics : Implications for second language learning and intercultural communication,* xiv, p. 105–124
- Goldberg A. E., 1995, *Constructions : A Construction Grammar Approach to Argument Structure*, Chicago, University of Chicago Press
- Gonzalez Rey I., 2002, *La phraséologie du français,* Toulouse, Presses universitaires du Mirail, Linguistique et didactique : Interlangues
- Gougenheim G., Michéa R., Rivenc P., Sauvageot A., 1956, *L'élaboration du français élémentaire. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base,* Paris, Didier

- Gougenheim G., 1958/1971, *Dictionnaire fondamental de la langue française,* Didier Edition internationale
- Gougenheim G., Michéa R., Rivenc P., Sauvageot A., 1964, *L'élaboration du français fondamental (1er degré). Nouvelle édition refondue et augmentée,* Paris, Didier
- Granger S., Paquot M., 2008, « Disentangling the phraseological web », dans Granger S., Meunier F. (Ed.), *Phraseology. An interdisciplinary perspective*, Amsterdam et Philadelphia, John Benjamins Publishing Company, p. 27-50
- Gross M., 1975, *Méthodes en syntaxe*, Paris, Hermann
- Gross G., 1981, « Les bases empiriques de la notion de prédicat sémantique », dans *Langages 73,* Paris, Larousse, p. 7-51
- Gross G., 1988, « Degré de figement des noms composés » dans *Langages* 90, p. 57-72
- Gross G., 1994, « Classes d'objets et description des verbes », dans *Langages,* n 115, Paris, Larousse, p. 15-30
- Gross G., 1996, *Les expressions figées en français. Nom composés et autres locutions*, Paris, Ophrys
- Grossmann F., Tutin A., 2002, « Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif », dans *Revue française de linguistique appliquée,* VII-1, p. 7-25
- Grossmann, F., & Tutin, A. (Ed.), 2003, *Les collocations : analyse et traitement*, Amsterdam, De Werelt
- Grossmann F., Paveau M-A., Petit G., 2005, *Didactique du lexique : langue, cognition, discours*, Grenoble
- Grossmann F., Plane S. (Ed.), 2008, *Lexique et production verbale : vers une meilleure intégration des apprentissages lexicaux*, Villeneuve d'Ascq, Presses Universitaires du Septentrion
- Halliday M. A. K., 1978, *Language as Social Semiotic. The social interpretation of language and meaning*, London, Arnold
- Halliday M. A. K., 1985, *An introduction to functional grammar*, 2e éd., London, Arnold
- Halliday M. A. K., 1996, « Lexis as a Linguistic Level » dans *Journal of*

*Linguistics 2(1)*, p. 57-67

- Hausmann F. J., 1979, « Un dictionnaire des collocations est-il possible ? » dans *Travaux de linguistique et de littérature XVII* (1), Strasbourg, p. 187-195
- Hausmann, F. J., 1984, « Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen franzosischer Wortverbindungen » dans *Praxis des neusprachlichen Unterrichts 31*, p. 395-406
- Hausmann F. J., 1989, « Le dictionnaire de collocations », dans Hausmann F. J., Reichmann O., Wiegand H. E., Zgusta L. (Ed.), *Wörterbücher : ein internationals Handbuch zur Lexicographie. Dictionaries. Dictionnaires*, Berlin/New-York, De Gruyter, p. 1010-1019
- Haygood J. D., 1937*, Le vocabulaire fondamental du français. Etude pratique sur l'enseignement des langues vivantes,* Genève, Librairie E. Droz
- Heid U., 1994, « On Ways Words Work Together - Topics in Lexical Combinatorics », dans *Euralex '94 Proceedings*
- Henmon V. A. C., 1924, *A French Word Book*, University of Wisconsin, Bureau of Educational Research, Bulletin n 3
- Herbst T., 1996, « What are collocations : sandy beaches or false teeth ? » dans *English Studies,* 77 (4), p. 379-393
- Hoey M., 2005, *Lexical priming. A new theory of words and language*, London, Rotledge
- Hornby A. *S.* (Ed.), 1974, *OALDCE - Oxford Advanced* Learner's *Dictionary of Current English,* Oxford, Oxford University Press
- Howarth P. A., 1996, « Phraseology in English academic writing : some implications for language learning and dictionary making », dans *Lexicographica Series Major 75,* Tübingen, Max Niemeyer
- Howarth P. A., Nesi H., 1996, *The teaching of collocations in EAP. Technical report*, Université de Leyde
- Jackendoff R., 2002, *Foundations of Language : Brain, Meaning, Grammar, Evolution*, Oxford, Oxford University Press
- Jespersen O., 1942, *A Modern English Grammar on Historical Principles. Part VI*, Copenhagen, Ejnar Munksgaard
- Ježek E. , 2005, *Lessico*, Bologna, il Mulino

- Jones S., Sinclair J., 1974, « English Lexical Collocations » dans *Cahiers de Lexicologie* 24, p. 15–61
- Juilland A., Brodin D., Davidovitch C., 1970, *Frequency dictionary of French words*, Paris, Mouton
- Juilland A., Traversa V., 1973, *Frequency Dictionary of Italian Words*, The Hague, Mouton
- Kaeding F. W., 1898, *Haufigkeitsworterbuch der Deutschen Sprache*, Steglitz bei Berlin
- Katz J. J., Fodor J. A., 1964, « The Structure of Language, Englewood Cliffs » dans Katz J. J., Fodor J. A. (Ed.), *The structure of Language. Readings in the Philosophy of Language*, NJ Prentice Hall, Englewood Cliffs, p. 479-518
- Knease T. C., 1933, *An Italian Word list from literary sources*, Toronto, The University of Toronto Press
- Kraif O., 2011, « Les concordances pour l'observation des corpus : utilité, outillage, utilisabilité », dans Jean Chuquet (sous la dir. de), *Le langage et ses niveaux d'analyse*, Presses universitaires de Rennes (PUR), chap. 4, p. 67-80
- Lakoff G., 1987, *Women, fire, and dangerous things – What categories reveal about the mind*, Chicago, Chicago University Press
- Langacker R., 1987, *Foundations of cognitive grammar, Volume 1 : Theoretical prerequisites*, Standford, Standford University Press
- Larivière L., 1998, « Valeur sémantique du verbe dans les collocations verbales spécialisées » dans *Diachronie et synchronie / Diachronics and Synchronics*, Association canadienne de traductologie, vol. 11, numéro 1, 1er semestre 1998, p. 173-197
- Laufer B., Nation P., 1995, « Vocabulary size and use : lexical richness in L2 written production » dans *Applied Linguistics 16*, p. 307-322
- *Le Français élémentaire*, 1954, Ministère de l'éducation nationale, Centre National de Documentation Pédagogique, Paris
- *Le français fondamental (1er degré)*, 1959, Ministère de l'éducation nationale, Direction de la coopération avec la communauté et l'étranger, Paris, Institut pédagogique national
- *Le français fondamental (2e degré)*, 1959, Ministère de l'éducation nationale,

Direction de la coopération avec la communauté et l'étranger, Paris, Institut pédagogique national

- Legallois D., 2005, « Du bon usage des expressions idiomatiques dans l'argumentation de deux modèles anglo-saxons : la Grammaire de Construction et la Grammaire des Patterns » dans Bolly C., Klein J. R., Lamiroy B. (Ed.), *La phraséologie dans tous ses états,* Actes du colloque Phraséologie 2005, 13-15 Octobre*,* Louvain-la-Neuve, Cahiers de l'Institut de Linguistique de Louvain, CILL 31. 2-4, p. 109-127
- Lewis M., 1993, *The lexical Approach*, London, Language Teaching Publications
- Lo Cascio V., 1997, « Semantica lessicale e i criteri di collocazione nei dizionari bilingui a stampa ed elettronici », dans De Mauro T., Lo Cascio V. (Ed.), *Lessico e grammatica*, Roma, Bulzoni, p. 63-88
- Lo Cascio V., 1998, « Compétence linguistique et Collocations », dans Collès L., Dufays J. L., Fabry G., Maeder C. (Ed.), *Didactique des langues romanes : le développement des compétences chez l'apprenant*, Actes du colloque de Louvain-la-Neuve janvier 2000, Bruxelles, De Boeck & Duculot, p. 349-359
- Lo Cascio V., 1999, « Standardisation and Collocations », dans Telen M., Lewandowska-Tomaszczyk B. (Ed.), *Translation and meaning- part 5*, Maastricht, Hoogschool Zuyd, p. 23-38
- Lo Cascio V., 2000, « La théorie des profils textuels et la compétence lexicale : les collocations » dans Collès L., Dufays J.-L., Fabry, G., Maeder C. (Ed.), *Didactique des langues romanes : le développement des compétences chez l'apprenant*, Actes du colloque de Louvain-la-Neuve janvier 2000, Bruxelles, De Boeck & Duculot, p. 349-359
- Lo Cascio V. (Ed.), 2007, « Il lessico nell'era digitale », dans *Parole in rete : apprendimento e teoria nell'era elettronica*, Novara, Utet-Università, p. 3-44
- Lo Cascio V., 2008, « Electronic Dictionaries and the Future of Translation », dans *Electronic Proceedings FIT*, Shanghai, World Congress on Translation
- *Longman dictionary of contemporary English*, 1978, London, Longman
- Manning C. D., Schütze H., 1999, *Foundations of statistical natural language processing*, Cambridge, MIT Press
- Mayer G., Reichenbach D., Ters F., 1969, *L'Echelle Dubois-Buyse d'orthographe*

*usuelle française*, Neuchâtel, Messeiller, éd. Critique

- McEnery T., Xiao R., Tono Y., 2006, *Corpus-based language studies. An advanced resource book*, London, Routledge
- Mel'čuk I. et al., 1984/1988, *Dictionnaire explicatif et combinatoire du français contemporain*, Montréal, Canada, Presses de l'Université de Montréal
- Mel'čuk I., 1998, « Collocations and Lexical Functions » dans Cowie A. P. (Ed.), *Phraseology. Theory, Analysis, and Applications*, Oxford, Clarendon Press, p. 23-53.
- Migliorini B., 1943, *Der grundlegende Wortschatz der Italienischen. Die 1500 wesentlichsten Worter*, Marburg, Elwert
- Moon R., 1998, *Fixed Expressions and Idioms in English : a corpus-based approach*, Oxford, Clarendon Press
- Morgan B. Q., 1928, « A German Frequency Word Book », dans *Publications of the American and Canadian Committees on Modern Languages*, vol. IX, New York, The Macmillan Co
- Nation I. S. P., 2001, *Teaching and explaining vocabulary*, Cambridge, Cambridge applied linguistics
- Nattinger J. R., DeCarrico J. S., 1992, *Lexical Phrases and Language Teaching*, Oxford, Oxford University Press
- Nerima L., Seretan V., Wehrli E., 2006, « Le problème des collocations en TAL », dans *Nouveaux cahiers de linguistique française 27*, p. 95-115
- Nesselhauf N., 2005, *Collocations in a learner corpus - Studies in corpus linguistics*, Amsterdam et Philadelphia, John Benjamins Publishing Company
- Ogden C. K ., 1930, *Basic English, a general introduction, with rules and grammar*, London, K. Paul
- Orliac B., 2004, *Automatisation du repérage et de l'encodage des collocations en langue de spécialité*, thèse de doctorat présentée à l'Université de Montréal, Montréal
- Osgood C. E., Suci C. J., Tannenbaum P. H., 1957, *The measurement of meaning*, Urbana, University of Illinois Press
- Palmer H. E., 1933, *Second interim report on English collocations*, Tokyo, Kaitakusha

- Palmer H. E., Hornby A. S., 1937, *Thousand-Word English*, London, George Harrap
- Palmer, 1938, *A Grammar of English Words*, London, Longman
- Paradis M., 2004, *A neurolinguistic theory of bilingualism*, Amsterdam, John Benjamins Publishing Company
- Péchoin D., 1991, *Thésaurus Larousse, Des mots aux idées, des idées aux mots*, Paris, Larousse
- Richards K., Savard J., 1970, *Les Indices d'utilité du vocabulaire fondamental français,* Québec, Les Presses de l'université de Laval
- Russo G. A., 1947, « A Combined Italian Word List » dans *The Modern Language Journal*, vol. 31, n. 4, p. 218-240
- Schmitt N., Grandage S., Adolphs S., 2004, « Are corpus-derived recurrent clusters psychologically valid ? » dans Schmitt N. (Ed.), *Formulaic Sequences,* Amsterdam et Philadelphia, John Benjamins Publishing Company, p. 127-151
- Sciarone A. G., 1977, *Vocabolario fondamentale della lingua italiana*, Bergamo, Minerva Italica
- Sinclair J., 1991, *Corpus, concordances, collocation*, Oxford, Oxford University Press
- Sinclair J., 1996, *Eagles preliminary recommendations on corpus typology*, EAG-TCWG-CTYP/P
- Sinclair J., 2000, « Lexical grammar », dans *Naujoji Metodologija 24*, p. 191–203 [http://donelaitis.vdu.lt/publikacijos/sinclair.pdf]
- Sinclair J., Jones S., Daley R., 2004, *English collocation studies : the OSTI report*, London et New York, Continuum
- Stubbs M., 1982, « Stir Until the Plot Thickens » dans Carter R. A., Burton D. (Ed.), *Literary Text and Language Study*, London, Arnold, p. 57-85
- Stubbs M., 1986, « Language development, lexical competence and nuclear vocabulary » dans *Educational Linguistics*, Oxford, Blackwell
- Stubbs M., 2002, *Words and phrases : corpus studies of lexical semantics*, Oxford, Blackwell Publishing
- Thompson M. E., 1927, *A study in Italian vocabulary frequency*, Thèse de master, Université de l'Iowa

- Thorndike E. L., 1921, *The Teacher's Word Book*, New York Teachers College, Columbia University
- Thornton A. M., Iacobini C., Burani C., 1997, *BDVDB – Una base di dati sul vdb della lingua italiana*, Roma, Bulzoni
- Tomasello M., 2003, *Constructing a language. Usage-Based Theory of Language Acquisition*, Cambridge, Massachusetts, London, Harvard University Press
- Van Roey J., 1990, *French-English contrastive lexicology : an introduction*, Louvain-la-Neuve, Peeters
- Vander Beke G. E., 1929, « French Word Book », dans *Publications of the American and Canadian Committees on Modern Languages*, vol. XV, New-York, The Macmillan Co
- Velez A., 2007, « Dictionnaires à portée de souris » dans *Glossari, dizionari, corpora : lessicologia e lessicografia delle lingue europee*, Gargnano del Garda (BS), 25-27 mai 2006, p. 247-274, Monza, Polimetrica International Publisher
- Velez A. (Ed.), 2009, *Acte du colloque : Giornate internazionali di studi sulla traduzione - Journées internationales d'études sur la traduction* , Cefalù 30-31 octobre et Ier novembre 2008, vol. 1, p. 1-358, Palermo, Herbita
- West M., 1953, *A General Service List of English Words*, London, Longman
- Williams G., 2003, « Les collocations et l'école contextualiste britannique », dans Grossmann F., Tutin A. (Ed.), *Les collocations : analyse et traitement*, Amsterdam, de Werelt, p. 33-44
- Willis D., 1990, *The Lexical Syllabus*, London et Glasgow, Collins ELT
- Wray A., Perkins M. R., 2000, « The functions of formulaic language : an integrated model », dans *Language and Communication*, 20, p. 1-28
- Wray A., 2002, *Formulaic language and the lexicon*, Cambridge, Cambridge University Press

# SITOGRAPHIE


- http://crawler.archive.org/
- http://o.bacquet.free.fr/index.html
- http://wacky.sslmit.unibo.it/doku.php?id=download
- http://www.ims.uni-stuttgart.de/projekte/corplex/*TreeTagger*/
- http://www.slideshare.net/metadonnee/questce-que-le-type-mime
- www.couchsurfing.com

# ANNEXES A

# Figures de corrélation entre fréquence et score des natifs

**FIGURE A.1 – Pivot *colloque*** : corrélation négative entre fréquence et score des natifs. Quelques points singuliers de fréquence élevée et score faible (*les communications du colloque* : 1605, 17% ; *le thème du colloque* : 1215, 17% ; *un colloque scientifique* : 1003, 3%), et de fréquence basse et score élevé (*le compte-rendu du colloque* : 255, 90% ; *la clôture du colloque* : 234, 87% ; *un colloque pluridisciplinaire* : 116, 73%).

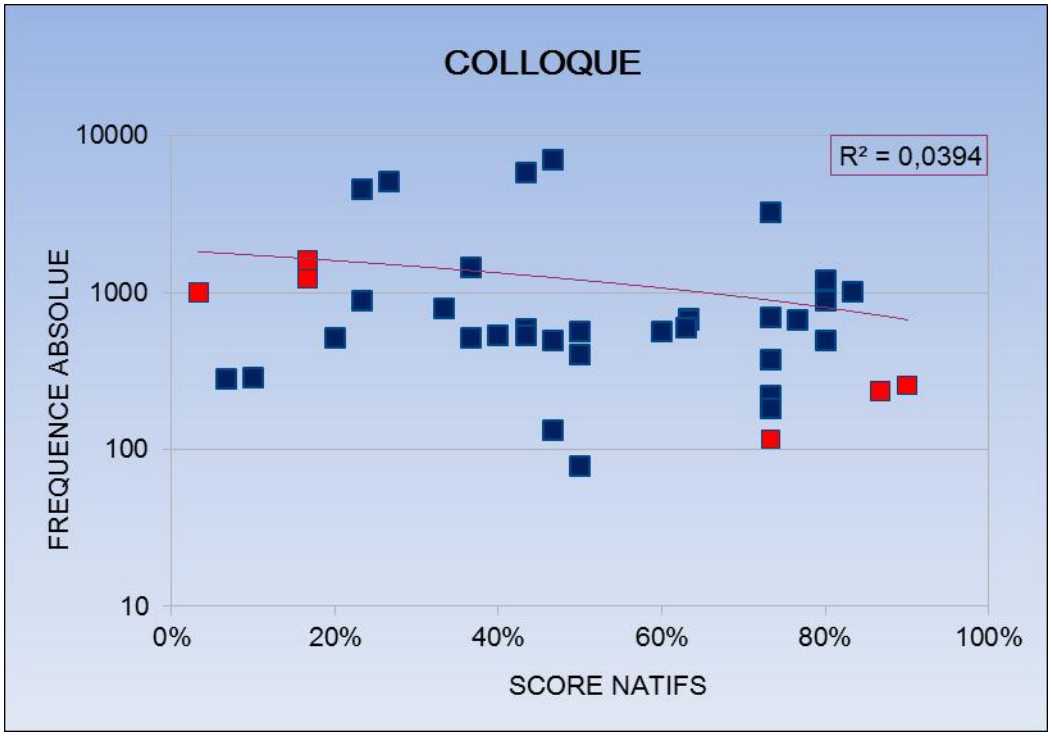


**FIGURE A.2 - Pivot *conférence*** : corrélation positive entre fréquence et score des natifs. Quelques points singuliers de fréquence élevée et score faible (*avoir une conférence* : 8069, 20% ; *le président de la conférence* : 1203, 17% ; *la conférence de rentrée* : 1053, 7%), et de fréquence basse et score élevé (*une conférence intergouvernementale* : 492, 40% ; *une conférence téléphonique* : 459, 83% ; *le compte-rendu de la conférence* : 436, 70%).

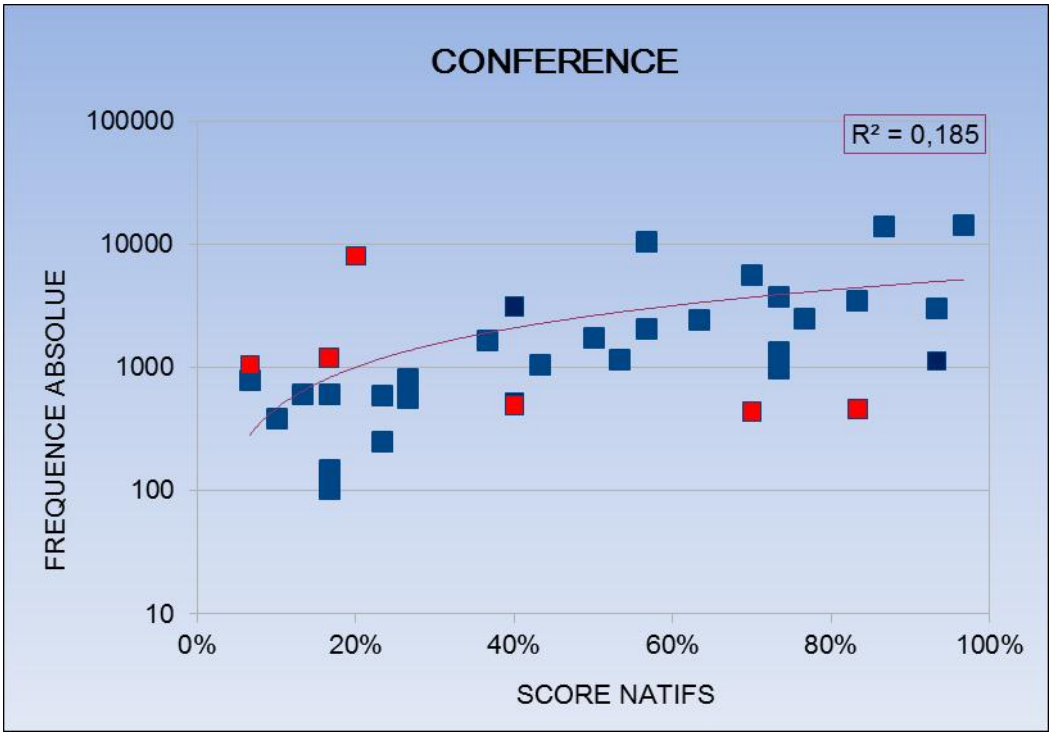


**FIGURE A.3 - Pivot *congrès*** : corrélation positive entre fréquence et score des natifs. Quelques points singuliers de fréquence élevée et score faible (*avoir un congrès*, 3670, 17% ; *réunir un congrès* : 1110, 23% ; *faire un congrès* : 644, 20%), et de fréquence basse et score élevé (*le congrès vote - une réforme, etc.* - : 371, 77% ; *les participants au congrès* : 276, 83% ; *l'organisateur du congrès* : 163, 67%).

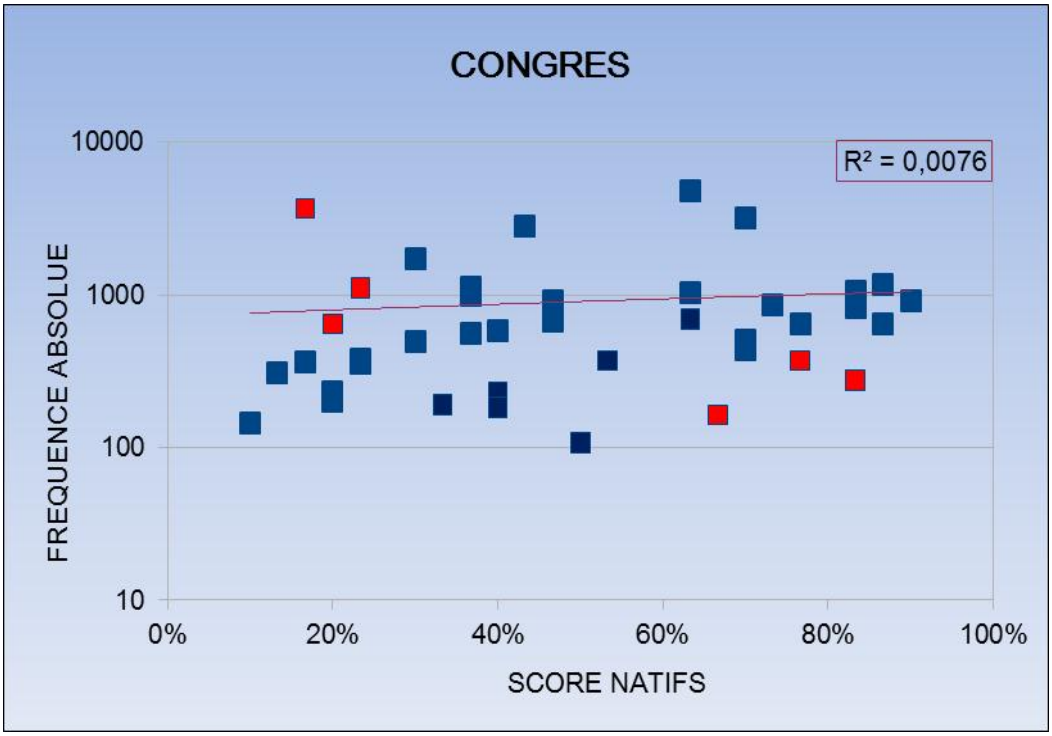


**FIGURE A.4 - Pivot *conversation*** : corrélation positive entre fréquence et score des natifs. Quelques points singuliers de fréquence élevée et score faible (*une conversation avec le jury* : 248, 23% ; *une conversation écrite* : 112, 7% ; *la conversation orale* : 98, 7%), et de fréquence basse et score élevé (*interrompre une conversation* : 93, 93% ; *en pleine conversation* : 87, 90% ; *une conversation intéressante* : 71, 97%).

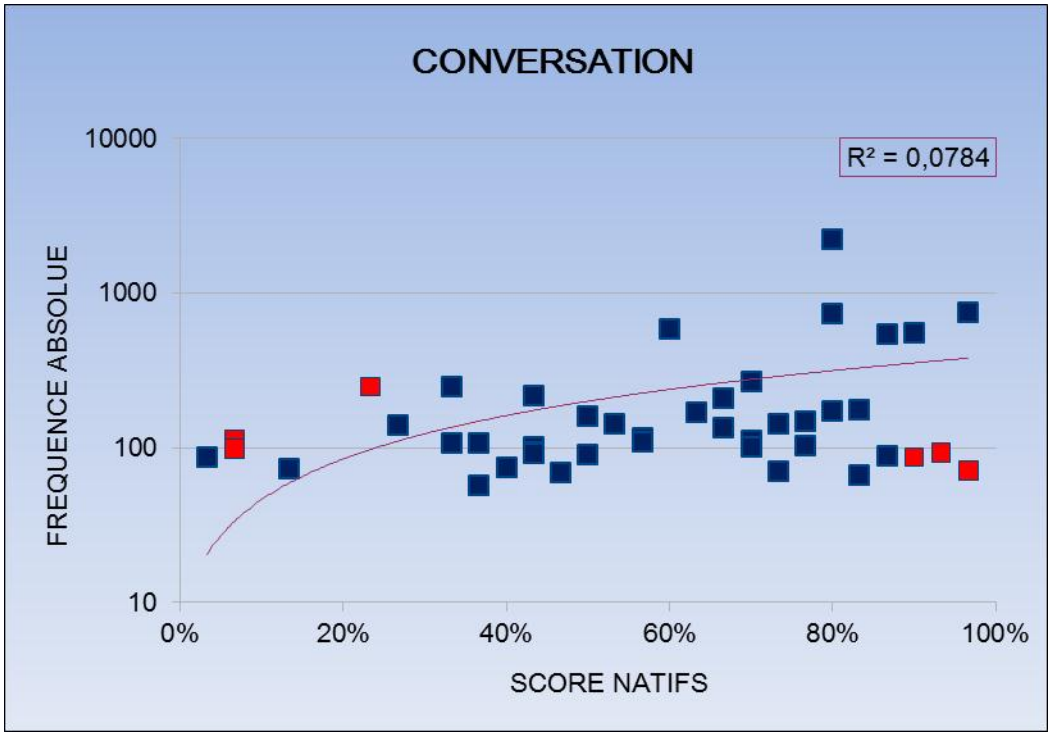


**FIGURE A.5 - Pivot *débat*** : corrélation positive entre fréquence et score des natifs. Quelques points singuliers de fréquence élevée et score faible (*un débat participatif* : 1706, 23% ; *un débat de réflexion* : 1420, 13% ; *un débat d'orientation* : 1361, 10%), et de fréquence basse et score élevé (*relancer le débat* : 1169, 87% ; *un débat télévisé* : 1101, 93% ; *alimenter le débat* : 899, 80%).

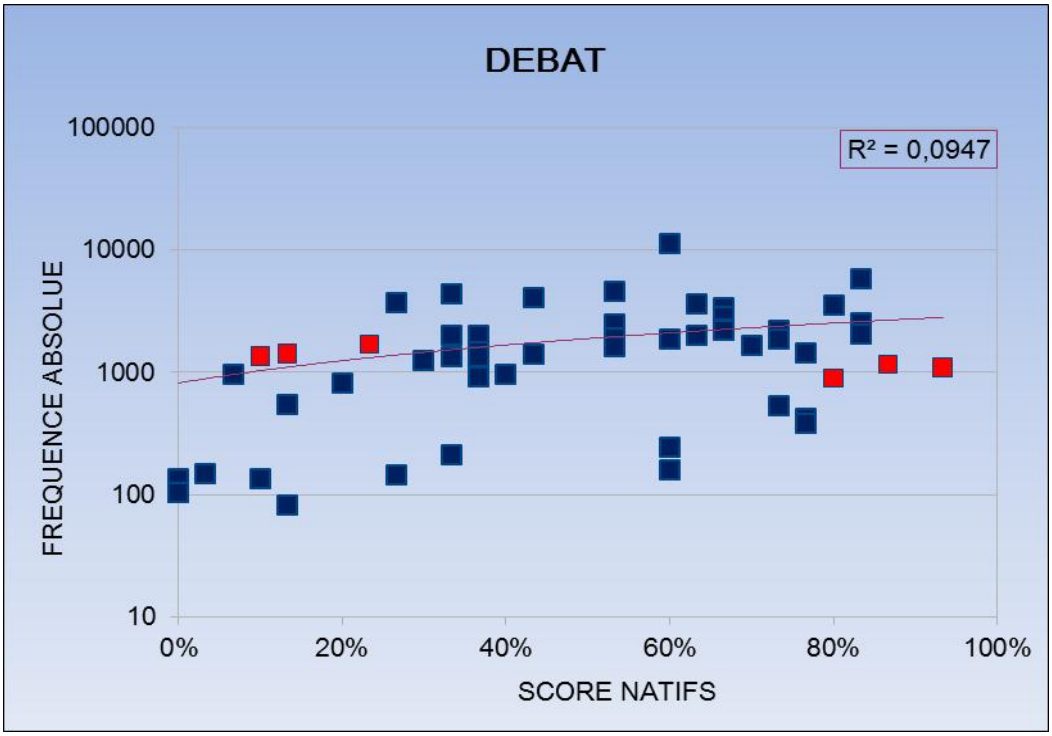


**FIGURE A.6 - *Pivot fête*** : corrélation positive entre fréquence et score des natifs. Quelques points singuliers de fréquence élevée et score faible (*avoir une fête* : 5291, 7% ; *une fête sainte* : 2281, 7% ; *l'édition d'une fête* : 1159, 3%), et de fréquence basse et score élevé (*l'ambiance de la fête* : 462, 77% ; *gâcher la fête* : 196, 83% ; *en habits de fête* : 135, 77%).

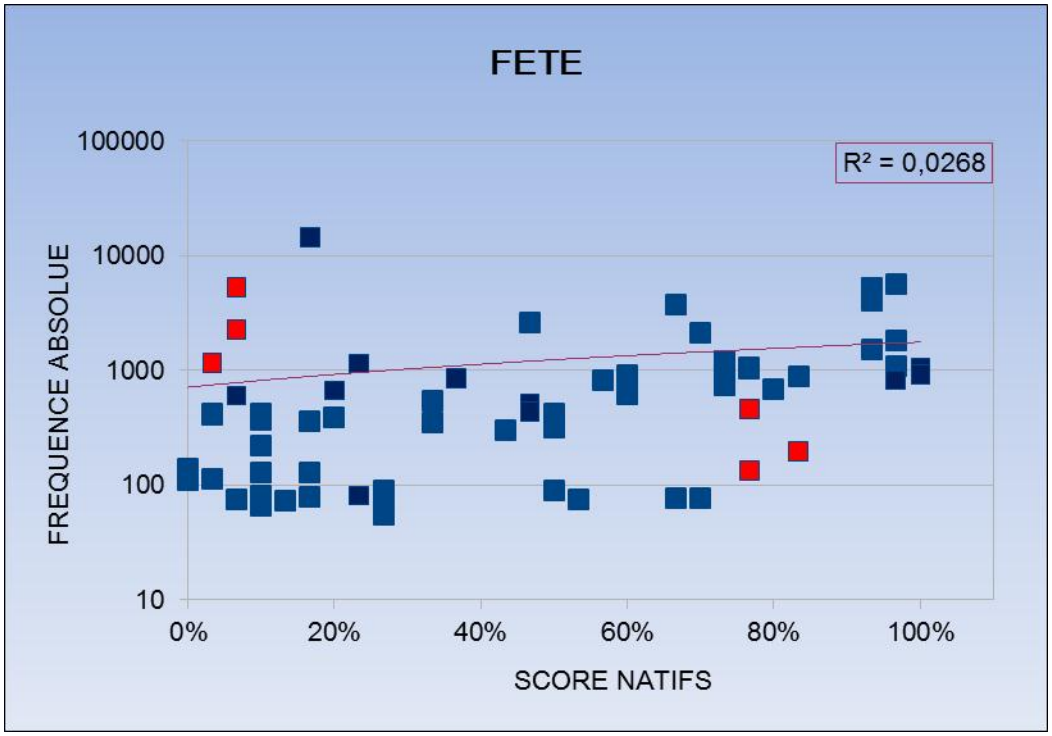


**FIGURE A.7 - Pivot *interview*** : corrélation positive entre fréquence et score des natifs. Quelques points singuliers de fréquence élevée et score faible (*une interview vidéo* : 1348, 13% ; *voir une interview* : 837, 20%), et de fréquence basse et score élevé (*une interview télévisée* : 163, 80% ; *une interview inédite* : 114, 80%).

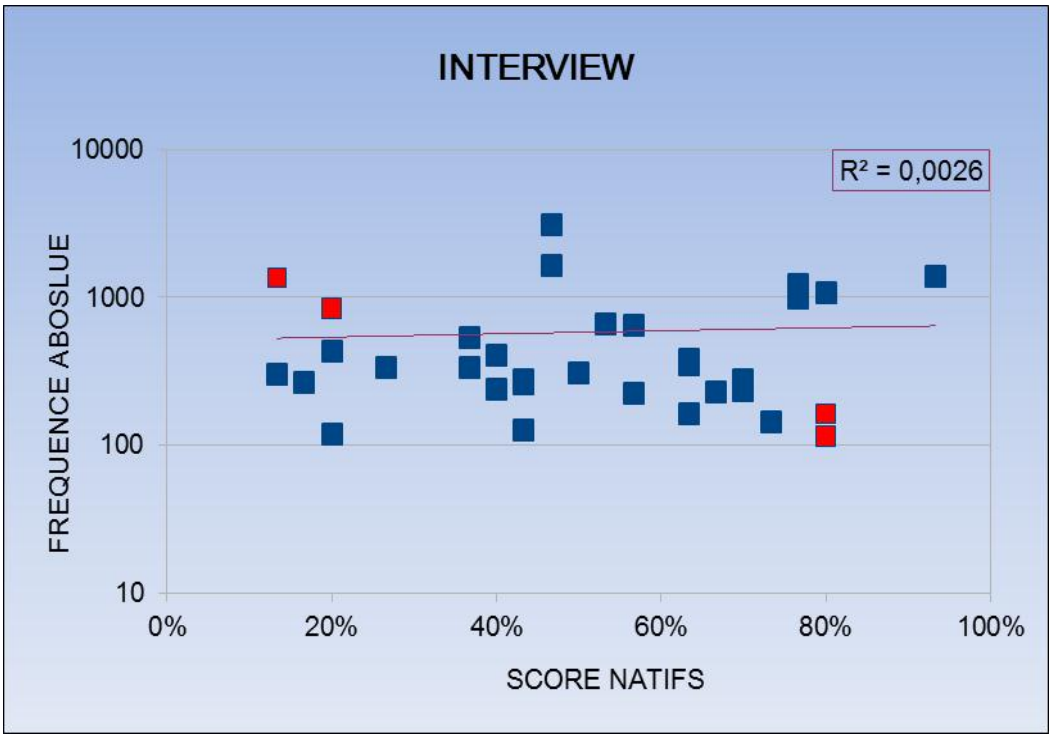


**FIGURE A.8 - Pivot *rencontre* :** corrélation négative entre fréquence et score des natifs. Quelques points singuliers de fréquence élevée et score faible (*être à une rencontre* : 18786, 7% ; *avoir une rencontre* : 14079, 7% ; *annoncer une rencontre* : 1985, 10%), et de fréquence basse et score élevé (*une rencontre amicale* : 908, 87% ; *une rencontre inattendue* : 291, 80% ; *une rencontre enrichissante* : 156, 73%).

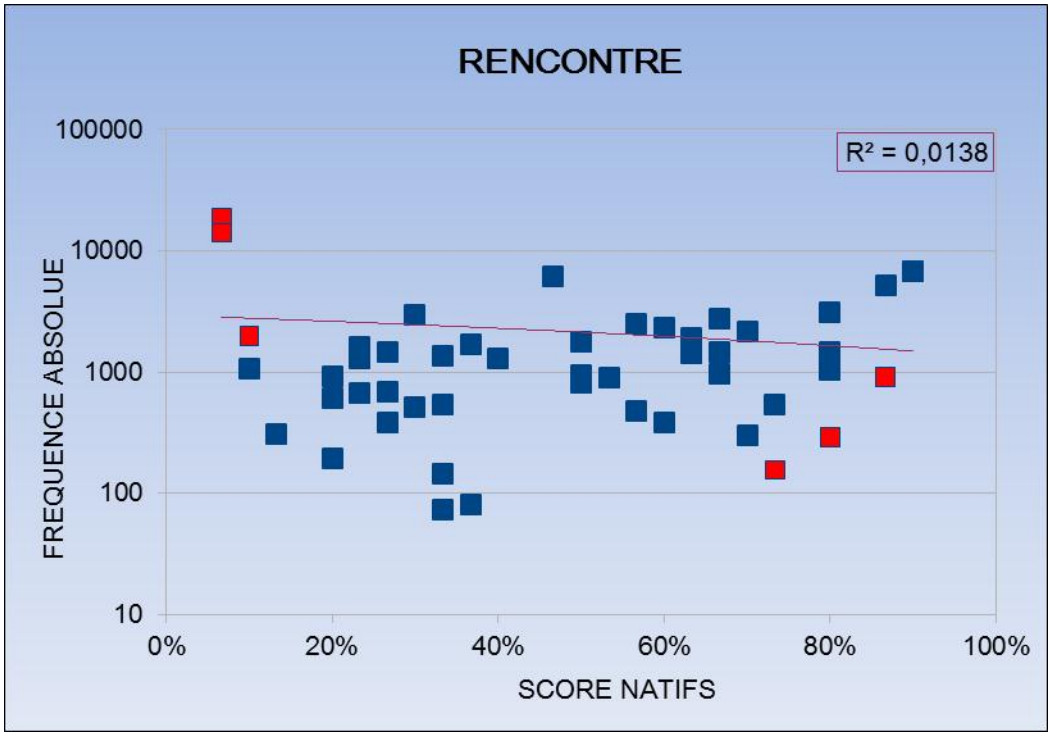


**FIGURE A.9 - Pivot *réunion*** : corrélation positive entre fréquence et score des natifs. Quelques points singuliers de fréquence élevée et score faible (*la prochaine réunion :* 4303, 23% ; *participer à la réunion* : 3806, 23%), et de fréquence basse et score élevé (*une réunion consultative* : 379, 87% ; *une réunion préparatoire* : 767, 93%).

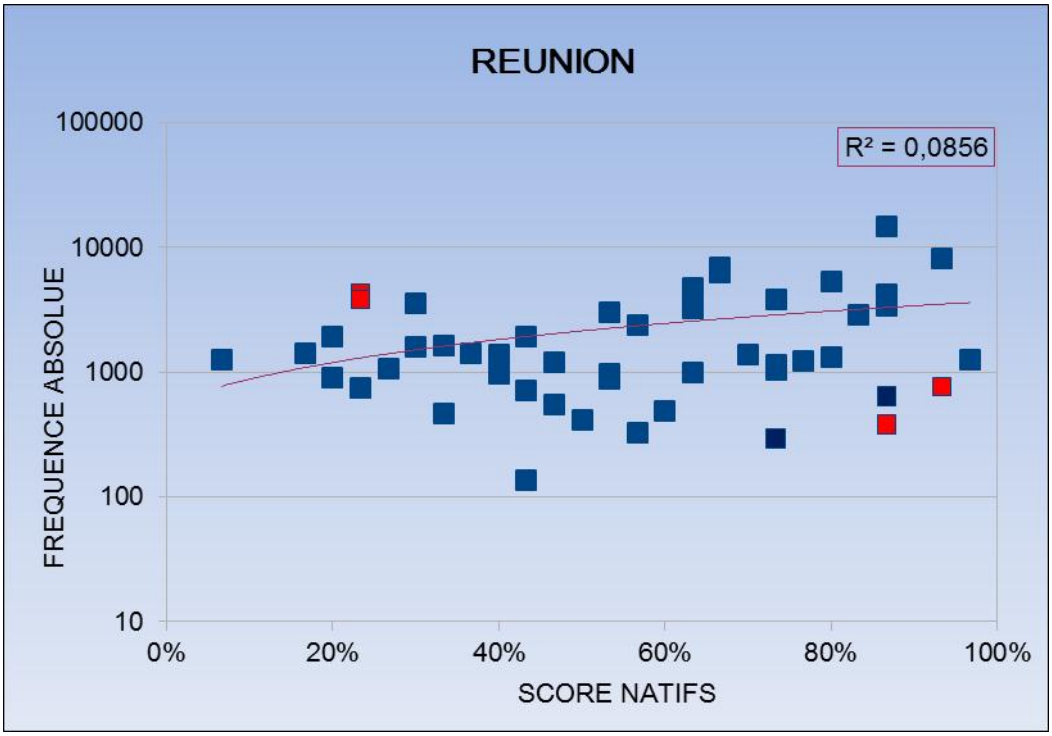
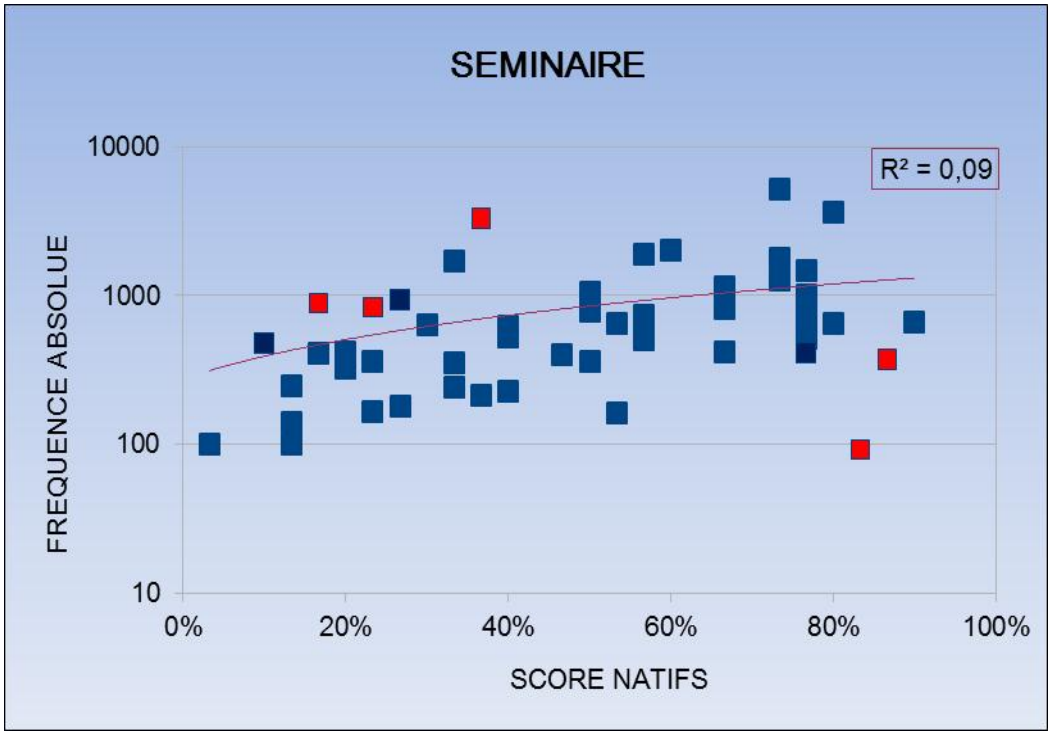


**FIGURE A.10 - Pivot *séminaire*** : corrélation positive entre fréquence et score des natifs. Quelques points singuliers de fréquence élevée et score faible (*avoir un séminaire* : 3297, 37% ; *les actes du séminaire* : 888, 17% ; *le séminaire se propose de …* : 836, 23%), et de fréquence basse et score élevé (*aller à un séminaire* : 92, 83% ; *le séminaire se déroule …* : 373, 87%).



# ANNEXES B

# Tableaux des associations fondamentales

**TABLEAU B.1 – Pivot *colloque*** : liste des associations fondamentales, surlignées en gris. Les unités les plus fréquentes sont indiquées par **F**, celles moins fréquentes « disponibles » sont indiquées par **D**. En rouge, les points singuliers. Seuil de significativité de la fréquence : 399.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | les **actes** du * | 7056 | 47% |
| | **organiser** un * | 5772 | 43% |
| | un * **international** | 5056 | 27% |
| | **être** à un * | 4565 | 23% |
| | **avoir** un * | 3220 | 73% |
| | les **communications** du * | 1605 | 17% |
| | l'**organisation** du * | 1451 | 37% |
| | le **thème** du * | 1215 | 17% |
| | **tenir** un * | 1188 | 80% |
| | un * **national** | 1007 | 83% |
| | un * **scientifique** | 1003 | 3% |
| | une **journée** de * | 886 | 80% |
| | le **programme** du * | 886 | 23% |
| | **à l'occasion du** * | 790 | 33% |
| | **participer** au * | 699 | 73% |
| | **dans le cadre du** * | 686 | 63% |
| | le * **a lieu …** | 673 | 63% |
| | un * **consacré à ...** | 664 | 77% |
| | un * **d'étude** | 601 | 63% |
| | **présenter** un * | 587 | 43% |
| | **faire** un * | 562 | 50% |
| | un * **annuel** | 559 | 60% |
| | la **participation** au * | 535 | 43% |
| | l'**ouverture** du * | 530 | 40% |
| | un * **intitulé X** | 515 | 20% |
| | les **participants** au * | 513 | 37% |
| | l'**intervention** au * | 495 | 80% |
| | **réunir** un * | 491 | 47% |
| | la **présentation** du * | 399 | 50% |
| D | le **prochain** * | 374 | 73% |
| | la **contribution** au * | 288 | 10% |
| | le * **se déroule**... | 282 | 7% |
| | le **compte-rendu** du * | 255 | 90% |
| | la **clôture** du * | 234 | 87% |
| | l'**organisateur** du * | 221 | 73% |
| | un * **interdisciplinaire** | 183 | 73% |
| | un * **singulier** | 133 | 47% |
| | un * **pluridisciplinaire** | 116 | 73% |
| | un * **co-organisé** | 78 | 50% |

**TABLEAU B.2 – Pivot *conférence*** : liste des associations fondamentales, surlignées en gris. Les unités les plus fréquentes sont indiquées par **F**, celles moins fréquentes « disponibles » sont indiquées par **D**. En rouge, les points singuliers. Seuil de significativité de la fréquence : 1043.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | une * **de presse** | 14207 | 97% |
| | un **maître de** *s | 14106 | 87% |
| | **être à** une * | 10418 | 57% |
| | **avoir** une * | 8069 | 20% |
| | **organiser** une * | 5589 | 70% |
| | une * **internationale** | 3729 | 73% |
| | **donner** une * | 3487 | 83% |
| | un **cycle** de *s | 3111 | 40% |
| | la **salle** de * | 3045 | 93% |
| | le **thème** de la * | 2478 | 77% |
| | **au cours de la** * | 2439 | 63% |
| | **tenir** une * | 2045 | 57% |
| | le **programme** de la * | 1749 | 50% |
| | **faire** une * | 1674 | 37% |
| | **animer** une* | 1331 | 73% |
| | **participer à** une * | 1233 | 73% |
| | le **président** de la * | 1203 | 17% |
| | **à l'occasion de la** * | 1166 | 53% |
| | **assister à** une * | 1130 | 93% |
| | la * **de rentrée** | 1053 | 7% |
| | une * **annuelle** | 1043 | 43% |
| D | l'**animation** de la * | 809 | 27% |
| | une * **de consensus** | 790 | 7% |
| | une * **épiscopale** | 609 | 17% |
| | **prononcer** une* | 608 | 13% |
| | la * **s'intitule ...** | 595 | 23% |
| | une * **mensuelle** | 569 | 27% |
| | une * **ministérielle** | 524 | 40% |
| | une * **intergouvernementale** | 492 | 40% |
| | une * **téléphonique** | 459 | 83% |
| | le **compte-rendu** de la * | 436 | 70% |
| | la **tenue** de la * | 382 | 10% |
| | une * **inaugurale** | 248 | 23% |
| | une * **tripartite** | 147 | 17% |
| | une * **introductive** | 102 | 17% |

**TABLEAU B.3 - Pivot *congrès*** : liste des associations fondamentales, surlignées en gris. Les unités les plus fréquentes sont indiquées par **F**, celles moins fréquentes « disponibles » sont indiquées par **D**. En rouge, les points singuliers. Seuil de significativité de la fréquence : 437.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | **être à** un * | 4824 | 63% |
| | **avoir** un * | 3670 | 17% |
| | le * **international** | 3175 | 70% |
| | le * **national** | 2796 | 43% |
| | le * **mondial** | 1730 | 30% |
| | le **prochain** * | 1174 | 87% |
| | **tenir** un * | 1114 | 37% |
| | **réunir** un * | 1110 | 23% |
| | **organiser** un * | 1059 | 83% |
| | la **préparation** d'un * | 1026 | 63% |
| | le * de l'**association** X | 991 | 37% |
| | un * **annuel** | 914 | 90% |
| | le **dernier** * | 904 | 47% |
| | **à l'occasion du** * | 863 | 73% |
| | l'**organisation** d'un * | 823 | 83% |
| | le * **européen** | 693 | 63% |
| | le * **adopte** (une réforme etc.) | 676 | 47% |
| | **participer à** un * | 645 | 87% |
| | **faire** un * | 644 | 20% |
| | le * **a lieu** | 640 | 77% |
| | un * **extraordinaire** | 579 | 40% |
| | le * **socialiste** | 553 | 37% |
| | un * **scientifique** | 505 | 70% |
| | la **synthèse** du * | 492 | 30% |
| | le **thème** du * | 438 | 70% |
| D | le * **communiste** | 382 | 23% |
| | le * **vote** (une réforme etc.) | 371 | 77% |
| | le **vote** du * | 371 | 53% |
| | le **fondateur** du * | 365 | 17% |
| | un * **départemental** | 358 | 23% |
| | le * **fédéral** | 306 | 13% |
| | les **participants au** * | 276 | 83% |
| | la **tenue** du * | 232 | 40% |
| | la **motion** du * | 231 | 20% |
| | **convoquer** un * | 200 | 20% |
| | le **délégué** du * | 191 | 33% |
| | la **clôture** du * | 182 | 40% |
| | l'**organisateur** du * | 163 | 67% |
| | un * **constitutif** | 146 | 10% |
| | le **compte-rendu** du * | 107 | 50% |

**TABLEAU B.4 - Pivot *conversation*** : liste des associations fondamentales, surlignées en gris. Les unités les plus fréquentes sont indiquées par **F**, celles moins fréquentes « disponibles » sont indiquées par **D**. En rouge, les points singuliers. Seuil de significativité de la fréquence : 98.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | **avoir** une * | 2220 | 80% |
| | une * **téléphonique** | 758 | 97% |
| | un **sujet** de * | 737 | 80% |
| | **suivre** la * | 591 | 60% |
| | **au cours d'une** * | 556 | 90% |
| | **engager** la * | 543 | 87% |
| | une **longue** * | 269 | 70% |
| | **entendre** la * | 249 | 33% |
| | une * **avec le jury** | 248 | 23% |
| | **tenir** une * | 218 | 43% |
| | **écouter** la * | 211 | 67% |
| | **entamer** la * | 176 | 83% |
| | **à la fin de la** * | 172 | 80% |
| | **enregistrer** une * | 171 | 63% |
| | la **suite** de la * | 161 | 50% |
| | **alimenter** la * | 148 | 77% |
| | **animer** la * | 143 | 53% |
| | **participer à la** * | 143 | 73% |
| | une **minute** de * | 140 | 27% |
| | **poursuivre** la * | 134 | 67% |
| | une * **en langue X** | 115 | 57% |
| | une * **écrite** | 112 | 7% |
| | **surprendre** une * | 112 | 70% |
| | **commencer** la * | 110 | 57% |
| | la * **courante** | 108 | 37% |
| | une **nouvelle** * | 107 | 33% |
| | **reprendre** la * | 103 | 77% |
| | l'**enregistrement** de la * | 102 | 43% |
| | une * **au téléphone** | 101 | 70% |
| | la * **orale** | 98 | 7% |
| D | **interrompre** une * | 93 | 93% |
| | une **simple** * | 93 | 43% |
| | **au détour d'une** * | 92 | 43% |
| | **continuer** la * | 91 | 50% |
| | **au fil de la** * | 88 | 87% |
| | **en pleine** * | 87 | 90% |
| | **tourner** la * | 87 | 3% |
| | **mener** une * | 75 | 40% |
| | la * **quotidienne** | 73 | 13% |
| | une * **intéressante** | 71 | 97% |
| | **mettre fin à la** * | 70 | 73% |
| | **lancer** une * | 69 | 47% |
| | une * **privée** | 66 | 83% |
| | une **courte** * | 57 | 37% |

**TABLEAU B.5 - Pivot *débat*** : liste des associations fondamentales, surlignées en gris. Les unités les plus fréquentes sont indiquées par **F**, celles moins fréquentes « disponibles » sont indiquées par **D**. En rouge, les points singuliers. Seuil de significativité de la fréquence : 1290.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | le * **public** | 11393 | 60% |
| | le * **politique** | 5787 | 83% |
| | un **grand** * | 4594 | 53% |
| | **organiser** le * | 4391 | 33% |
| | le * **national** | 4042 | 43% |
| | la **question** du * | 3681 | 27% |
| | le * **aura lieu ...** | 3628 | 63% |
| | **participer** à un * | 3567 | 80% |
| | **ouvrir** le * | 3418 | 67% |
| | **suivre** le * | 2842 | 67% |
| | **animer** le * | 2543 | 83% |
| | un * **démocratique** | 2471 | 53% |
| | le **cœur** du * | 2255 | 67% |
| | le * **parlementaire** | 2216 | 67% |
| | un * **d'idées** | 2192 | 73% |
| | **lancer** le * | 2043 | 83% |
| | l'**objet** de * | 2000 | 33% |
| | le **sujet** de * | 1997 | 37% |
| | le **thème** du * | 1997 | 63% |
| | le * **en cours** | 1937 | 53% |
| | **engager** un * | 1922 | 53% |
| | un **vrai** * | 1879 | 60% |
| | le * **porte sur ...** | 1856 | 73% |
| | un * **participatif** | 1706 | 23% |
| | **susciter** un * | 1664 | 70% |
| | un * **actuel** | 1607 | 53% |
| | un * **contradictoire** | 1480 | 37% |
| | un * **de fond** | 1439 | 77% |
| | le * **sur les enjeux …** | 1425 | 43% |
| | un * **de réflexion** | 1420 | 13% |
| | un * **d'orientation** | 1361 | 10% |
| | le * **budgétaire** | 1355 | 33% |
| | le * **citoyen** | 1346 | 37% |
| | la **contribution** au * | 1256 | 30% |
| | **relancer** le * | 1169 | 87% |
| | un * **télévisé** | 1101 | 93% |
| | une **séance** de * | 961 | 7% |
| | un * **en séance** (publique etc.) | 961 | 40% |

| | | | |
|---|---|---|---|
| D | le * **présidentiel** | 915 | 37% |
| | **alimenter** le * | 899 | 80% |
| | un * **vif** | 817 | 20% |
| | **argumenter** le * | 545 | 13% |
| | **clore** le * | 536 | 73% |
| | un * **constructif** | 417 | 77% |
| | un * **houleux** | 386 | 77% |
| | un * **stérile** | 244 | 60% |
| | **rouvrir le *** | 211 | 33% |
| | **recentrer** le * | 160 | 60% |
| | un * **escamoté** | 150 | 3% |
| | **dépassionner** le * | 145 | 27% |
| | **retransmettre** le * | 136 | 10% |
| | le * **historiographique** | 135 | 0% |
| | un * **pluraliste** | 105 | 0% |
| | le * **référendaire** | 104 | 0% |
| | un * **âpre** | 83 | 13% |

**TABLEAU B.6 - Pivot *fête*** : liste des associations fondamentales, surlignées en gris. Les unités les plus fréquentes sont indiquées par **F**, celles moins fréquentes « disponibles » sont indiquées par **D**. En rouge, les points singuliers. Seuil de significativité de la fréquence : 582.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | **être en** * | 14385 | 17% |
| | la **salle** des *s | 5661 | 97% |
| | **avoir** une * | 5291 | 7% |
| | une * **de fin d'année** | 5263 | 93% |
| | **faire** la * | 5250 | 93% |
| | un **jour** de * | 4000 | 93% |
| | de **bonnes** *s | 3787 | 67% |
| | **à l'occasion de** la * | 2647 | 47% |
| | une * **sainte** | 2281 | 7% |
| | **organiser** une * | 2145 | 70% |
| | une * **nationale** | 1816 | 97% |
| | une * **d'anniversaire** | 1515 | 93% |
| | le **comité** des *s | 1215 | 73% |
| | une * **religieuse** | 1209 | 73% |
| | l'**édition** d'une * | 1159 | 3% |
| | **célébrer** la * | 1153 | 23% |
| | une * **foraine** | 1091 | 97% |
| | **en période de** * | 1069 | 77% |
| | **souhaiter** de (bonnes etc.) *s | 1069 | 100% |
| | une **belle** * | 1027 | 77% |
| | de **joyeuses** * | 923 | 100% |
| | **participer** à la * | 903 | 60% |
| | une * **de famille** | 888 | 83% |
| | une * **de mariage** | 859 | 37% |
| | la * du **village** | 826 | 97% |
| | une * **populaire** | 823 | 73% |
| | une * **traditionnelle** | 819 | 57% |
| | le **repas** de * | 739 | 73% |
| | une * **de quartier** | 684 | 80% |
| | l'**animation** de la * | 670 | 20% |
| | **préparer** une * | 625 | 60% |
| | le **cadeau** de la * | 603 | 7% |
| | une **soirée** de * | 549 | 33% |
| | une * **annuelle** | 516 | 47% |
| | l'**ambiance** de la* | 462 | 77% |
| | une * **chrétienne** | 435 | 47% |
| | la * **se déroule** ... | 421 | 50% |
| | le **calendrier** de la * | 420 | 3% |

| | | | |
|---|---|---|---|
| D | une * **votive** | 417 | 10% |
| | la **cérémonie** de la* | 406 | 3% |
| | une * **familiale** | 387 | 20% |
| | la **célébration** de la * | 373 | 10% |
| | une * **patronale** | 364 | 17% |
| | la **veille** de la * | 352 | 33% |
| | le **lendemain** de la * | 320 | 50% |
| | une * **médiévale** | 304 | 43% |
| | une * **liturgique** | 223 | 10% |
| | **gâcher** la * | 196 | 83% |
| | une * **galante** | 139 | 0% |
| | **en habits de** * | 135 | 77% |
| | **commémorer** une * | 128 | 17% |
| | la **convivialité** de la * | 128 | 10% |
| | une * **centenaire** | 113 | 3% |
| | le **rite** de la * | 111 | 0% |
| | une * **costumée** | 91 | 50% |
| | une * **folklorique** | 89 | 27% |
| | le **banquet** de la * | 81 | 23% |
| | la **commémoration** d'une * | 81 | 10% |
| | une * **somptueuse** | 79 | 17% |
| | **convier à** la * | 78 | 70% |
| | une * **inoubliable** | 77 | 67% |
| | une * **solennelle** | 76 | 7% |
| | les **préparatifs** de la * | 76 | 53% |
| | **rythmer** la * | 74 | 13% |
| | une * **champêtre** | 69 | 27% |
| | une * **villageoise** | 66 | 10% |
| | une * **grandiose** | 55 | 27% |

**TABLEAU B.7 - Pivot *interview*** : liste des associations fondamentales, surlignées en gris. Les unités les plus fréquentes sont indiquées par F, celles moins fréquentes « disponibles » sont indiquées par D. En rouge, les points singuliers. Seuil de significativité de la fréquence : 226.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | **avoir** une * | 3093 | 47% |
| | **lire** une * | 1650 | 47% |
| | **accorder** une * | 1391 | 93% |
| | une * **vidéo** | 1348 | 13% |
| | **réaliser** une * | 1214 | 77% |
| | une * **exclusive** | 1067 | 80% |
| | **donner** une * | 992 | 77% |
| | **voir** une * | 837 | 20% |
| | **publier** une * | 664 | 53% |
| | **faire** une * | 639 | 57% |
| | l'* du **journal X** | 532 | 37% |
| | **retrouver** une * | 434 | 20% |
| | une * **radio** | 407 | 40% |
| | la **suite** de l'* | 382 | 63% |
| | l' * **est extraite de** ... | 346 | 63% |
| | **l'auteur** d'une * | 337 | 37% |
| | une * **à paraître** | 331 | 27% |
| | **écouter** une * | 308 | 50% |
| | une * **de presse** | 299 | 13% |
| | un **extrait** d'* | 277 | 70% |
| | **diffuser** une * | 275 | 43% |
| | l'* du **quotidien** X | 262 | 17% |
| | une **longue** * | 257 | 43% |
| | l'* du **magazine X** | 240 | 40% |
| | l'**intégralité** d'une * | 231 | 70% |
| | l'* du **journaliste X** | 229 | 67% |
| D | une * **récente** | 224 | 57% |
| | une * **intéressante** | 164 | 63% |
| | une * **télévisée** | 163 | 80% |
| | l'* d'une **personnalité** | 143 | 73% |
| | **filmer** une * | 127 | 43% |
| | une * **audio** | 119 | 20% |
| | une * **inédite** | 114 | 80% |

**TABLEAU B.8 - Pivot *rencontre*** : liste des associations fondamentales, surlignées en gris. Les unités les plus fréquentes sont indiquées par **F**, celles moins fréquentes « disponibles » sont indiquées par **D**. En rouge, les points singuliers. Seuil de significativité de la fréquence : 1058.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | **être** à une * | 18786 | 7% |
| | **avoir** une * | 14079 | 7% |
| | la * **aura lieu …** | 6839 | 90% |
| | un **lieu** de * | 6839 | 90% |
| | **organiser** la * | 6064 | 47% |
| | **aller** à la * | 5175 | 87% |
| | un **site** de * | 3069 | 80% |
| | une * **nationale** | 2979 | 30% |
| | une * **internationale** | 2777 | 67% |
| | une * **professionnelle** | 2515 | 57% |
| | **à l'occasion de la** * | 2331 | 60% |
| | **venir à** la * | 2136 | 70% |
| | **annoncer** une * | 1985 | 10% |
| | un **point** de * | 1930 | 63% |
| | **participer** à une * | 1770 | 50% |
| | une **journée** de * | 1676 | 37% |
| | une * **occasionnelle** | 1593 | 23% |
| | un **espace** de * | 1461 | 67% |
| | la **prochaine** * | 1448 | 80% |
| | le **thème** de la * | 1444 | 27% |
| | **favoriser** la * | 1423 | 63% |
| | la * des **acteurs** (sociales, etc.) | 1368 | 33% |
| | l'**organisation** de la * | 1299 | 40% |
| | une * **régionale** | 1293 | 23% |
| D | une * de **cultures** | 1056 | 10% |
| | une * **sportive** | 1033 | 80% |
| | l'**issue** de la * | 963 | 67% |
| | une * **annuelle** | 942 | 50% |
| | la * des **adhérents** | 912 | 20% |
| | une * **amicale** | 908 | 87% |
| | une * **culturelle** | 890 | 53% |
| | la * **se déroule …** | 811 | 50% |
| | le **compte-rendu** de la * | 687 | 27% |
| | une * **privilégiée** | 676 | 23% |
| | une * **régulière** | 612 | 20% |

| | | | |
|---|---|---|---|
| | **au hasard d'**une * | 537 | 73% |
| | **disputer** la * | 536 | 33% |
| | une * **conviviale** | 509 | 30% |
| | une * **informelle** | 476 | 57% |
| | une * **préparatoire** | 386 | 27% |
| | une * **fortuite** | 382 | 60% |
| | la **convivialité** de la * | 305 | 13% |
| | une * **improbable** | 300 | 70% |
| | une * **inattendue** | 291 | 80% |
| | une * **interrégionale** | 192 | 20% |
| | une * **enrichissante** | 156 | 73% |
| | une * **coquine** | 144 | 33% |
| | une * **inopinée** | 81 | 37% |
| | une * **intergénérationnelle** | 73 | 33% |

**TABLEAU B.9 – Pivot *réunion*** : liste des associations fondamentales, surlignées en gris. Les unités les plus fréquentes sont indiquées par F, celles moins fréquentes « disponibles » sont indiquées par D. En rouge, les points singuliers. Seuil de significativité de la fréquence : 1050.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | **être en** * | 14868 | 87% |
| | une * **publique** | 8170 | 93% |
| | **organiser** une * | 6826 | 67% |
| | une * **d'information** | 6218 | 67% |
| | la **salle de** * | 5358 | 80% |
| | une * **de travail** | 4699 | 63% |
| | la **prochaine** * | 4303 | 23% |
| | **se rendre à** la * | 4231 | 87% |
| | la * **se tiendra …** | 3860 | 73% |
| | **participer à** la * | 3806 | 23% |
| | la * **a lieu …** | 3598 | 30% |
| | la * du **comité** (central, etc.) | 3393 | 87% |
| | la * du **groupe** x | 3205 | 63% |
| | la * de la **commission** (mixte, etc.) | 3048 | 53% |
| | le **compte-rendu** de la * | 2899 | 83% |
| | la **dernière** * | 2396 | 57% |
| | **assister à** la * | 1944 | 43% |
| | **faire** une * | 1908 | 20% |
| | l'**organisation** de la * | 1628 | 33% |
| | la **date** de la * | 1596 | 30% |
| | une * **de bureau** | 1431 | 37% |
| | **animer** une * | 1406 | 17% |
| | une * **de concertation** | 1397 | 40% |
| | **prévoir** une * | 1368 | 70% |
| | l'**issue** de la * | 1310 | 80% |
| | une * **collective** | 1263 | 97% |
| | une * **plénière** | 1242 | 7% |
| | une * **de section** | 1221 | 77% |
| | la * de l'**assemblée** (générale, etc.) | 1196 | 47% |
| | une * **annuelle** | 1149 | 73% |
| | la * des **ministres** | 1073 | 27% |
| | une * **de quartier** | 1050 | 73% |
| | **préparation de** la * | 981 | 63% |
| | une * **municipale** | 974 | 53% |
| | une * **mensuelle** | 971 | 40% |
| | la **participation à** la * | 903 | 20% |

| | | | |
|---|---|---|---|
| **D** | le **procès-verbal** de la * | 873 | 53% |
| | une * **préparatoire** | 767 | 93% |
| | une * **thématique** | 741 | 23% |
| | la * **se déroulera** ... | 720 | 43% |
| | le **calendrier** de la * | 638 | 87% |
| | **présider** une * | 555 | 47% |
| | une * **hebdomadaire** | 488 | 60% |
| | une * **informelle** | 466 | 33% |
| | **convoquer** une * | 416 | 50% |
| | une * **consultative** | 379 | 87% |
| | une * **interministérielle** | 327 | 57% |
| | **convier** une * | 293 | 73% |
| | une * **tripartite** | 135 | 43% |

**TABLEAU B.10 - Pivot *séminaire*** : liste des associations fondamentales, surlignées en gris. Les unités les plus fréquentes sont indiquées par **F**, celles moins fréquentes « disponibles » sont indiquées par **D**. En rouge, les points singuliers. Seuil de significativité de la fréquence : 464.

| | ASSOCIATIONS | FREQUENCE | SCORE NATIFS |
|---|---|---|---|
| F | **être à un** * | 5201 | 73% |
| | **organiser** un * | 3643 | 80% |
| | **avoir** un * | 3297 | 37% |
| | un * **de recherche** | 2005 | 60% |
| | un * **de formation** | 1882 | 57% |
| | **dans le cadre du** * | 1787 | 73% |
| | les **séances** du * | 1700 | 33% |
| | **au cours du** * | 1482 | 77% |
| | l'**organisation** du * | 1280 | 73% |
| | un * **de travail** | 1143 | 67% |
| | la **salle** du * | 1042 | 50% |
| | le **programme** du * | 1017 | 77% |
| | un * **national** | 936 | 27% |
| | les **actes** du * | 888 | 17% |
| | le * **a lieu …** | 875 | 77% |
| | le * **se propose de...** | 836 | 23% |
| | une **journée de** * | 814 | 67% |
| | un * **d'étude** | 781 | 50% |
| | un * **international** | 736 | 57% |
| | le **thème** du * | 679 | 77% |
| | **animer** le * | 668 | 57% |
| | **participer** au * | 663 | 90% |
| | l'**objectif** du * | 648 | 80% |
| | la **présentation** du * | 642 | 53% |
| | **faire** un * | 630 | 30% |
| | **tenir** un * | 618 | 40% |
| | un * **consacré à …** | 606 | 77% |
| | **ouvrir** le * | 525 | 40% |
| | le **prochain** * | 516 | 77% |
| | un * **annuel** | 508 | 57% |
| | un * **étudiant** | 481 | 10% |
| | les **participants** au * | 428 | 77% |
| | la **participation** au * | 420 | 67% |
| | un * **gouvernemental** | 415 | 20% |
| | **à l'occasion du** * | 413 | 77% |
| | les **rencontres** du * | 412 | 17% |

| | | | |
|---|---|---|---|
| **D** | un * **thématique** | 399 | 47% |
| | le * **se déroule ...** | 373 | 87% |
| | le * **réunit …** (les chercheurs, etc.) | 357 | 50% |
| | le * **se réunit …** (chaque jeudi, etc.) | 357 | 23% |
| | un **cycle** de *s | 350 | 33% |
| | un * **doctoral** | 323 | 20% |
| | le * des **doctorants** | 246 | 13% |
| | un * **intitulé X** | 243 | 33% |
| | un * **mensuel** | 229 | 40% |
| | un * **interdisciplinaire** | 211 | 37% |
| | un * **transversal** | 182 | 27% |
| | la **méthodologie** du * | 166 | 23% |
| | le **compte-rendu** du * | 162 | 53% |
| | un * **méthodologique** | 140 | 13% |
| | un * **résidentiel** | 101 | 13% |
| | un * **de restitution** | 101 | 3% |
| | **aller à un** * | 92 | 83% |