



HAL
open science

Le statut de la fréquence dans les grammaires de constructions : simple comme bonjour ?

Guillaume Desagulier

► **To cite this version:**

Guillaume Desagulier. Le statut de la fréquence dans les grammaires de constructions : simple comme bonjour ?. 2014. halshs-01056861v3

HAL Id: halshs-01056861

<https://shs.hal.science/halshs-01056861v3>

Preprint submitted on 12 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Le statut de la fréquence dans les grammaires de constructions : *simple comme bonjour ?*

Guillaume Desagulier

MoDyCo – Université Paris 8, CNRS, Université Paris Ouest Nanterre La Défense

1. Introduction

Dans la théorie de la Grammaire Cognitive, la grammaire est la représentation psychologique du système linguistique (Langacker, 1987, p. 57). Cette représentation est à la fois hiérarchisée et dynamique. Elle est hiérarchisée car les unités symboliques qui la composent se combinent pour former un réseau de constructions (Langacker, 1986, p. 29). Elle est dynamique car la forme du réseau s'adapte sans cesse à l'usage.

On ne peut cerner en quoi la grammaire est fondée sur l'usage sans faire appel à l'articulation entre unité symbolique, répétition et ancrage cognitif (*entrenchment*) : « [w]ith repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit (...) » (1987, p. 59). Plus un composant sémantique et un composant phonologique sont associés dans l'usage, plus cet assemblage tendra à s'ancrer cognitivement dans la grammaire au point d'acquérir le statut d'unité symbolique. Dès lors, la grammaticalité n'est plus affaire de jugement binaire, consistant pour le linguiste à décider de manière introspective qu'une unité fait partie de la grammaire ou qu'elle en est rejetée, mais de degré d'ancrage, partant de convention. Ce qui définit le degré d'ancrage d'une unité (et donc son degré de convention), c'est la fréquence avec laquelle une unité apparaît dans l'usage : « (...) units are variably entrenched depending on the frequency of their occurrence (*driven*, for example, is more entrenched than *thriven*) » (*ibid.*).

Formalisée à l'origine par Langacker (1987) et Goldberg (1995) et affinée par la suite (Goldberg, 2003, 2006, 2009; Langacker, 2008, 2009), la Grammaire de Constructions Cognitive (GxCC) reprend les grands principes de la Grammaire Cognitive, dont l'idée d'un réseau de constructions hiérarchisé et dynamique en prise avec l'usage. Toute structure linguistique peut prétendre au statut de construction dès lors qu'une partie de sa forme ou de son sens n'est pas dérivable d'une autre construction (Goldberg, 2003, p. 219). En vertu de ce postulat, la GxCC est non-réductionniste puisque l'inventaire des constructions ne se limite pas à des schémas syntaxiques abstraits (ex. la construction passive, l'inversion sujet-auxiliaire). Il comprend aussi d'autres parties du discours, à savoir des lexèmes (ex. *État*, *voyou*, *État-voyou*), des morphèmes (ex. *déconstruire*) ou des séquences idiomatiques à divers niveaux de productivité (ADJ *de chez* ADJ, V_{INF} *plus pour* V_{INF} *plus*).¹

Par contraste, la Grammaire de Constructions d'inspiration Fillmoreenne (GxCF) est réductionniste (Fillmore, 1997; Fillmore et al., 1988). N'ont le statut de construction que les schémas généraux productifs tels que *let alone* (Fillmore et al., 1988), *what's X doing Y* (Kay & Fillmore, 1999), ou les constructions clivées en *all* (Kay, 2013). Tout ce qui n'est pas nécessaire et suffisant pour interpréter et générer d'autres expressions linguistiques est relégué en périphérie de la grammaire.² Selon Kay (2013), <ADJ *as* GN> fait partie de ces schémas non-productifs exclus de l'inventaire des constructions :

- (1) a. *stiff as a board*
raide comme un piquet (lit. « raide comme une planche »)
b. *cool as a cucumber*
d'un calme olympien (lit. « calme comme un concombre »)
c. *flat as a pancake*
plat comme une limande (lit. « plat comme une crêpe »)

¹ La Grammaire de Constructions prend ainsi très au sérieux le principe de non-séparation du lexique et de la grammaire postulé par la Grammaire Cognitive.

² L'idée selon laquelle la grammaire a un centre et une périphérie rend la GxCF proche du générativisme dont elle émane.

En (1b), connaître le sens de *cool* et celui de *cucumber* ne suffit pas à prédire l'association des deux lexèmes ni le sens de leur appariement. De plus, ce schéma ne peut pas être étendu à d'autres paires ADJ-GN similaires (^{??}*cool as a tomato*, ^{??}*hot as a zucchini*). Selon Kay, cela suffit à exclure de la grammaire le schéma <ADJ as GN> car il ne s'agit que d'un schéma dérivé, ou *pattern of coining*. Pourtant, ce schéma est fréquent et productif : « [t]here are many members of the A as NP pattern, and it is likely that new ones come into existence now and then as analogical creations ». Mais contrairement à ce que postule la Grammaire Cognitive, la productivité n'est pas indexée sur la fréquence dans la GxCF.

Kay ne cache pas la radicalité de son approche. Pourtant, en GxCC, rien ne s'oppose a priori à ce que le schéma <ADJ as GN> soit considéré comme une construction. Dès lors qu'une combinaison ADJ-GN est suffisamment ancrée, le sens de l'ensemble ne correspond plus exactement à la somme du sens des deux éléments. En (2) ci-dessous, le sens cible (« très heureux ») n'est pas strictement dérivable de la combinaison de *happy* (« heureux ») et *a clam* (« une palourde ») :

- (2) *I'm happy as a clam.*
Je suis heureux comme un poisson dans l'eau.

Dans la mesure où le statut de <ADJ as GN> diffère en fonction de la grille de lecture théorique qui en est faite (construction selon la GxCC, schéma dérivé selon la GxCF), nous nous proposons de tester la frontière que propose Kay entre ce qui relève d'une construction et ce qui n'en relève pas. Notre hypothèse de travail est la suivante : on ne peut rendre compte de la nature constructionnelle de <ADJ as GN> sans s'interroger sur la productivité respective des éléments qui la composent. Bien que d'apparence homogène, il se peut que <ADJ as GN> regroupe des combinaisons ADJ-GN productives et d'autres plus figées.

De manière à nous affranchir le plus possible d'un parti pris théorique, nous optons pour une démarche empirique visant à mesurer la productivité de <ADJ as GN> dans un corpus d'anglais américain. Notre approche se veut originale car une telle démarche empirique ne fait pas partie du projet initial des grammaires de constructions.

L'article est structuré comme suit. La deuxième section est épistémologique. Nous y abordons le statut de la fréquence dans le contexte du tournant quantitatif en linguistique cognitive et en particulier dans les grammaires de constructions. Si dès l'origine la fréquence y est décrite comme un phénomène central dans l'ancrage cognitif des constructions, il faut attendre l'analyse collocationnelle pour voir émerger une prise en compte empirique et quantitative du statut de la fréquence. La troisième section est consacrée à l'étude de <ADJ as GN> en corpus. À la lumière des méthodes quantitatives mises en avant dans le sillage de l'analyse collocationnelle, il est possible de mesurer la productivité d'une construction, indépendamment de postulats théoriques. Nous présentons et utilisons quatre outils : l'analyse collexémique covariante, la classification ascendante hiérarchique, l'analyse factorielle des correspondances et ΔP , une mesure d'association issue de la théorie de l'information permettant de repérer les collocations asymétriques. La quatrième section propose une discussion critique des résultats.

2. Le tournant quantitatif de la linguistique cognitive

2.1. La fréquence dans une approche centrée sur l'usage

Lorsque Ronald Langacker pose les bases de la Grammaire Spatiale (Langacker, 1982), les pratiques méthodologiques s'articulent autour de trois axiomes issus de *Syntactic Structures* (Chomsky, 1957) et *Aspects of the Theory of Syntax* (Chomsky, 1965).

- i. le langage est un système autonome et inné ;
- ii. la syntaxe est au cœur de la grammaire ; les composantes lexicale et sémantique sont reléguées en périphérie ;
- iii. la structure conceptuelle de la grammaire relève d'une logique formelle vériconditionnelle et réductionniste.

À cela, Langacker oppose que :

- i. le langage n'est pas une faculté cognitive autonome ;
- ii. la grammaire est à la fois conceptualisée et vecteur de conceptualisation ;
- iii. la connaissance linguistique est tributaire de l'usage.

Ces trois principes, propres à la Grammaire Cognitive que nous connaissons aujourd'hui (Langacker, 1987, 1991)³, forment l'ossature de la linguistique cognitive. Le troisième principe est la clé de voûte de la linguistique dite « fondée sur l'usage » (*usage-based linguistics*). Alors que la tradition mentaliste cherche à définir les règles minimales qui permettent de générer une infinité de phrases, procédant ainsi du haut (les règles) vers le bas (les phrases)⁴, les approches centrées sur l'usage procèdent en sens inverse : c'est en observant l'usage de formes linguistique en contexte qu'on parvient à déterminer les règles de leur emploi.⁵ D'un côté, nous avons affaire à un modèle de la compétence, par essence introspectif puisque centré sur un locuteur idéal, et de l'autre un modèle inductif, par définition exploratoire et empirique.

Le modèle fondé sur l'usage prend doublement le contrepied du modèle génératif. D'une part, il institue une gradation dans l'acceptabilité d'une expression. D'autre part, il indexe ce degré d'acceptabilité non pas sur le jugement expert du linguiste mais sur la fréquence, ce en vertu d'une règle simple : « [e]very use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact » (*ibid.*). Ce changement de point de vue au bénéfice de la fréquence pose deux types de problèmes.

Le premier problème relève de la nature ambiguë de la fréquence. Si la fréquence est le principe organisateur de la grammaire, aucune méthode n'est proposée pour l'appréhender empiriquement (Glynn, 2010b; Tummers et al., 2005). De fait, tout porte à croire que la Grammaire Cognitive et, au delà, la plupart des théories fondées sur l'usage⁶, s'appuient sur une définition théorisée et abstraite de la fréquence. Ce qui compte pour évaluer l'ancrage d'une unité linguistique ne serait pas tant la fréquence *mesurée* en corpus par le linguiste que la fréquence *perçue* par les locuteurs natifs. C'est sans doute pour cette raison que la frontière entre ce qui relève d'une unité ancrée et ce qui n'en relève pas est subjective, pour ne pas dire introspective : « [i]s there some particular level of entrenchment, with special behavioral significance, that can serve as a nonarbitrary cutoff point in defining units? There are no obvious linguistic grounds for believing so » (*ibid.*). Pourtant, Langacker admet plus tard que cette frontière peut être identifiée de manière quantitative en prenant en compte la fréquence observée :

In practical terms, we cannot always know whether a particular expression is established as a unit. While this may not be of any great importance (in the grand scheme of things), in principle the degree of entrenchment can be determined empirically. Observed frequency provides one basis for estimating it. Experimentally, one can look for measurable differences in the processing of novel vs. unit expressions (Harris 1998). We have more direct evidence of unit status when an expression consistently displays some idiosyncrasy that does not follow from any regular pattern. (Langacker, 2008, p. 238)

De subjective, la fréquence devient objective car mesurable, ce qui ouvre la voie aux techniques statistiques :

With large samples and appropriate statistical techniques, for example, speaker judgments could help determine whether *ring* 'circular piece of jewelry' and *ring* 'arena' represent alternate senses of a polysemous lexical item (as opposed to being

³ Les deux premiers se retrouvent également chez Lakoff (1987), selon qui la linguistique est une discipline nécessairement holistique, fondée sur deux postulats : le Postulat de Généralisation (*Generalization Commitment*) et le Postulat Cognitif (*Cognitive Commitment*) (Lakoff, 1990, p. 40). En effet, en vertu du Postulat de Généralisation, le linguiste n'a pas de raison a priori de séparer les niveaux d'analyse que sont la syntaxe, la sémantique, la pragmatique, la phonologie, etc. Il se doit au contraire de reconnaître que ces niveaux obéissent à des principes langagiers plus généraux. Le Postulat Cognitif relève de la même logique : le linguiste n'a aucune raison a priori de séparer la faculté de langage d'autres facultés cognitives. Il se doit donc de reconnaître que la structure linguistique est la trace d'opérations cognitives plus générales.

⁴ On parle alors de *top-down approach*.

⁵ On parle de *bottom-up approach*.

⁶ Par exemple en linguistique exemplariste (Bybee, 1985, 2006, 2010; Bybee & Hopper, 2001), dans laquelle la fréquence joue un rôle tout aussi central. Voir également Barlow and Kemmer (2000) et Langacker (2000).

unrelated homonyms), or whether *computer* is in fact more analyzable than *propeller*. (Langacker, 2008, p. 86)

On notera que ce revirement intervient parallèlement à l'essor d'études quantitatives (sur corpus) et expérimentales en linguistique cognitive.⁷

Le second problème concerne le lien direct entre fréquences (hautes) et ancrage cognitif. Celui-ci ne va pas de soi. Des recherches menées en sémantique cognitive ont montré que les unités linguistiques n'étaient pas ancrées du seul fait de leur haute fréquence (Schmid, 2010). Un phénomène complémentaire entre en jeu : la saillance. Mis en avant par des travaux sur la prototypicalité (Geeraerts et al., 1994), le concept de saillance est encore difficile à cerner. On en trouve une typologie chez Schmid (2007) et Geeraerts (2000). Sans faire une épistémologie nécessairement risquée de ce terme, il ressort qu'une unité est saillante lorsqu'elle atteint un certain degré de prééminence cognitive. Sur le plan linguistique, la fréquence et la saillance peuvent être à la fois en résonance et en dissonance. L'usage nous dit que plus une forme est fréquente, plus son degré d'ancrage dans la grammaire est élevé. Si nous poursuivons ce raisonnement, nous sommes en droit de penser que plus une unité est ancrée (parce que fréquente), plus sa place dans la grammaire est prééminente. Cette corrélation directe entre fréquence et saillance relève de la résonance. Cependant, les travaux sur la grammaticalisation (notamment sur la « décoloration sémantique », ou *semantic bleaching*) montrent qu'une unité fréquente subit généralement un affaiblissement potentiel de son sens.⁸ Il en résulte une moindre saillance au niveau sémantique et une dissonance entre fréquence et saillance. A contrario, une unité peu fréquente mais rendue saillante dans une situation discursive particulière peut prétendre à un statut ancré. Dans ce cas de figure, fréquence et saillance sont en dissonance. Ce détour par la saillance nous permet d'émettre des réserves sur le lien entre fréquence et ancrage.

Pour résumer, si la linguistique cognitive de première génération est empirique dans ses principes, il est difficile d'y retrouver la trace d'une méthodologie empirique dans sa pratique, en dépit d'un possible revirement tardif.

2.2. L'apport de la linguistique de corpus

La linguistique de corpus repose sur l'hypothèse selon laquelle le contexte d'une variable lexicale ou phrastique révèle des aspects importants de sa syntaxe ou de sa sémantique (Biber, 1998; Sinclair, 1991). Il n'est donc pas surprenant de voir que la linguistique de corpus a acquis un rôle central dans l'étude des schémas d'usage en linguistique cognitive au sens large (Gries & Stefanowitsch, 2006).

L'utilisation des corpus en linguistique cognitive est attestée depuis le début des années 80 (cf. par exemple Dirven et al., 1982; Dirven & Taylor, 1988). C'est d'ailleurs une spécificité de la tradition européenne par rapport à la tradition américaine, cette dernière étant nettement plus introspective. Mais cet emploi des corpus n'implique pas de méthode quantitative. C'est véritablement avec l'émergence de l'analyse collostructionnelle au début des années 2000 que se systématisent le couplage entre linguistique de corpus et méthodes quantitatives en linguistique cognitive.

L'analyse collostructionnelle propose une extension de l'analyse des collocations au domaine des grammaires de constructions (en particulier de la GxCC). En vertu du principe selon lequel on connaît le sens d'un mot en étudiant son voisinage (Firth, 1957),⁹ le profil sémantique d'un mot cible dépend très largement de son contexte lexical. Dès Firth figure l'intuition selon laquelle les collocations sont déterminées quantitativement et statistiquement. En effet, le mot cible et son voisinage ne sont pas que de simples juxtapositions mais des unités reliées par des « attentes mutuelles » (Firth, 1957 : 181). Des recherches menées à la suite de Firth ont confirmé les intuitions de ce dernier (Sinclair, 1966, 1987; Sinclair & Carter, 2004) et ont étendu les collocations au delà des unités lexicales pour inclure le

⁷ Pour une épistémologie plus précise, voir Geeraerts (2010, pp. 263-264).

⁸ Cela est vérifié par la loi de Zipf (Zipf, 1935), selon laquelle la fréquence d'un mot est inversement proportionnelle à son rang dans un tableau de fréquences. Les premiers rangs, qui concernent les mots les plus fréquents dans un texte, sont généralement tenus par les mots grammaticaux (au sémantisme plus abstrait) tandis que les rangs les plus éloignés sont tenus par des mots lexicaux (à plus forte coloration sémantique).

⁹ « You shall know a word by the company it keeps » (Firth, 1957, p. 179).

phénomène de cooccurrence à l'intersection du lexique et de la grammaire (Sinclair, 1991, 1996; Stubbs, 2001). Les unités linguistiques concernées par le phénomène de collocation se situent désormais à différents niveaux de schématisation et de convention. Elles incluent des GN (*État providence*), des GV (*se passer de quelque chose*), des locutions adverbiales (*grosso modo*) et des constructions plus complexes (ADJ *comme* GN). Parmi les méthodes quantitatives censées cerner les collocations, y compris dans leur dimension phraséologique, l'analyse collocationnelle est l'avatar le plus récent et sans doute le plus avancé au sens où elle a su formaliser et diffuser l'idée suivante : pour qu'il y ait collocation, il ne suffit pas qu'au moins deux unités lexicales apparaissent côte-à-côte en corpus ; la cooccurrence de ces unités doit être plus fréquente que ce à quoi on peut s'attendre.

L'analyse collocationnelle regroupe trois méthodes : l'analyse collexémique (Stefanowitsch & Gries, 2003), l'analyse collexémique distinctive (Gries & Stefanowitsch, 2004b) et l'analyse collexémique covariante (Gries & Stefanowitsch, 2004a; Stefanowitsch & Gries, 2005). L'analyse collexémique mesure le degré de répulsion ou d'attraction entre une construction et une position instanciable dans cette même construction (ex. les substantifs spécifiques du quantifieur *quelques*). L'analyse collexémique distinctive mesure la préférence d'un lemme pour une construction par rapport à une autre construction équivalente (ex. les substantifs spécifiques du quantifieur *quelques* par rapport aux substantifs spécifiques du quantifieur *plusieurs*¹⁰). Enfin, l'analyse collexémique covariante mesure le degré d'attraction ou de répulsion de lemmes dans différentes positions d'une même construction (ex. <V-inf *plus pour* V-inf *plus*>, <ADJ *comme* GN>).

L'analyse collocationnelle s'appuie sur des mesures d'association développées au cours des cinq décennies précédentes, le but de ces méthodes étant de mesurer l'attraction ou la répulsion entre deux lexèmes.¹¹ Ces mesures sont résumées dans le Tableau 1.

méthodes	mesures d'association
analyse collexémique	test exact de Fisher, rapport de log-vraisemblance, information mutuelle, test du χ^2 , odds ratio
analyse collexémique distinctive	test exact de Fisher, rapport de log-vraisemblance
analyse collexémique distinctive multiple	test multinomial
analyse collexémique covariante	test exact de Fisher, rapport de log-vraisemblance, odds ratio

Tableau 1. Mesures d'association intervenant dans l'analyse collocationnelle

La plupart des mesures d'association se fondent sur un tableau d'entrée tel que le Tableau 2 pour générer une *p*-valeur.¹² Plus cette *p*-valeur tend vers 0 et plus on peut rejeter l'hypothèse nulle selon laquelle il n'y a pas d'attraction entre un lexème et un autre.

	L ₁	¬L ₁	total (lignes)
L ₂	a	b	a+b
¬L ₂	c	d	c+d
total (colonnes)	a+c	b+d	a+b+c+d

Tableau 2. Tableau d'entrée pour mesurer l'association entre deux lexèmes L₁ et L₂ (¬ : « autre que »)

Le principe est le même en analyse collocationnelle, si ce n'est que les fréquences de lexèmes sont rapportées à leurs distributions constructionnelles et que les *p*-valeurs sont converties par log-transformations en « force collocationnelle » pour permettre un classement des éléments les plus

¹⁰ Voir Gréa (2008) qui compare les spécificités de *quelques* et *plusieurs* grâce à la loi hypergéométrique, technique très proche de l'analyse collexémique distinctive.

¹¹ On compte à ce jour plusieurs dizaines de méthodes, dont on trouvera un inventaire critique chez Evert (2005) et Pecina (2010).

¹² L'information mutuelle procède différemment puisqu'elle attribue un score d'association.

attirés vers les moins attirés.¹³ Cette similitude de fonctionnement se voit à travers les tableaux d'entrée employés dans l'analyse collocationnelle traditionnelle (Tableau 2) et ceux utilisés dans l'analyse collexémique (Tableau 3), l'analyse collexémique distinctive (Tableau 4) et l'analyse collexémique covariante (Tableau 5).

	L	¬L	total (lignes)
C	a	b	a+b
¬C	c	d	c+d
total (colonnes)	a+c	b+d	a+b+c+d

Tableau 3. Tableau d'entrée pour mesurer l'association entre un lexème L et une construction C (C : construction)

	L	¬L	total (lignes)
C ₁	a	b	a+b
C ₂	c	d	c+d
total (colonnes)	a+c	b+d	a+b+c+d

Tableau 4. Tableau d'entrée pour mesurer l'association entre un lexème L et deux constructions C₁ et C₂

	L _{créneau 1}	¬L _{créneau 1}	total (lignes)
L _{créneau 2}	a	b	a+b
¬L _{créneau 2}	c	d	c+d
total (colonnes)	a+c	b+d	a+b+c+d

Tableau 5. Tableau d'entrée pour mesurer l'association entre un lexème L dans le premier créneau d'une construction et un autre lexème dans le deuxième créneau d'une construction

L'originalité de l'analyse collostructionnelle ne vient donc ni de son principe de fonctionnement, ni des mesures d'association qu'elle fait intervenir.¹⁴ Son apport se résume plutôt en deux aspects : (a) son domaine d'application, en l'occurrence les constructions lexico-syntaxiques dans une optique fondée sur l'usage; (b) sa remise en cause de la distinction artificielle entre collocation et colligation. Ce second aspect est d'autant plus pertinent qu'il semble improbable que les locuteurs soient sensibles à la fréquence d'un lexème donné en dehors du contexte grammatical dans lequel celui-ci est employé.

2.3. Au-delà des collocations, l'approche multivariée

Le succès de l'analyse collostructionnelle ne doit pas en cacher les limites, la plupart desquelles s'appliquent à l'ensemble des mesures d'association. La première de ces limites est formulée par Firth lui-même : il est vain de croire que les collocations suffisent à circonscrire le sens d'un mot. Elles sont, tout au plus, un moyen aisé d'accéder approximativement au sens au niveau strictement lexical (Firth, 1957 : 181).

Concernant plus spécifiquement l'analyse collostructionnelle, les limites sont de trois ordres. Premièrement, les tableaux de sortie peuvent contenir jusqu'à des milliers de lignes correspondant à autant de lexèmes spécifiques à une ou plusieurs constructions, ce qui en rend la synthèse difficile à

¹³ Vu que les *p*-valeurs sont infinitésimales, les classer n'a aucun sens. La log-transformation a le double avantage d'amplifier les différences (par exemple une très forte attraction se traduit par une force collostructionnelle qui tend vers l'infini) et de permettre un classement plus lisible.

¹⁴ L'accueil de l'analyse collostructionnelle en France fut mitigé pour deux raisons principales, l'une historique, l'autre théorique. Historiquement, l'école française de lexicométrie est déjà bien implantée (Lafon, 1980, 1981, 1984; Lebart & Salem, 1994; Muller, 1964, 1973, 1977). Théoriquement, on reproche par exemple à l'analyse collostructionnelle de réinventer le calcul des spécificités de Lafon, ou de faire doublon avec les études sur la colligation.

l'œil nu. Par conséquent, la description d'une grammaire de l'usage fondée uniquement sur ces tableaux crée un décalage avec l'expérience des locuteurs car ces derniers parviennent sans mal à percevoir des régularités dans la masse des données propre à leur pratique langagière. Si l'analyse collostructionnelle se veut fondée sur l'usage, il est paradoxal qu'elle fournisse en sortie des tableaux dont la nature ne reflète pas l'usage.

Deuxièmement, à l'instar des mesures d'association sur lesquelles elle s'appuie, l'ACC ne rend pas compte de l'asymétrie dans l'attraction entre deux lexèmes. Or, dans un appariement de lexèmes, l'attraction est rarement symétrique. C'est particulièrement évident dans *ad hominem*, où *hominem* attire plus *ad* que vice versa.

Troisièmement, l'analyse collostructionnelle est univariée et monodimensionnelle car elle ne prend en compte qu'un seul type de variable, à savoir la fréquence d'occurrence d'un ou plusieurs lexèmes dans une ou plusieurs constructions. C'est un inconvénient en linguistique cognitive. En effet, cette dernière considère que la structure du sens est en prise directe avec l'expérience humaine dans son ensemble (et pas seulement langagière). Dans la mesure où l'expérience est confrontée à une multiplicité de variables concomitantes, il est naturel de penser que la grammaire est également soumise à l'influence simultanée de variables linguistiques, paralinguistiques et extralinguistiques. Une approche fondée sur l'usage ne peut donc pas se contenter de l'étude des collocations, y compris dans ses développements phraséologiques, mais doit s'étendre aux cooccurrences entre phénomènes linguistiques et phénomènes contextuels. Ceci explique pourquoi les derniers développements méthodologiques en linguistique cognitive se concentrent sur les méthodes statistiques multifactorielles (Desagulier, sous presse; Divjak & Fieller, to appear; Glynn, 2010a, 2010b; Glynn & Robinson, à paraître; Gries, 2003). Ces dernières mettent en regard une ou plusieurs variables dépendantes à la lumière de plusieurs variables indépendantes. Elles sont suffisamment puissantes pour tout à la fois explorer de gigantesques tableaux de contingence et générer des synthèses graphiques pertinentes d'un point de vue centré sur l'usage.

Dans l'étude de cas qui suit, nous présentons des méthodes qui permettent de dépasser ces limites tout en se fondant sur les résultats de l'ACC.

3. Etude de cas : ADJ *as* GN

3.1. Corpus et méthode

Le corpus retenu est le Corpus of Contemporary American English (Davies, 2008-). Il contient 464 020 256 mots répartis en 189 431 textes d'anglais américain compilés entre 1990 et 2012. Le COCA est un corpus dit « équilibré » au sens où il est divisé en cinq genres de tailles en mots équivalentes : transcriptions d'anglais parlé (~ 90 millions de mots), fiction (~ 90 millions de mots), magazines populaires (~ 95 millions de mots), journaux (~ 92 millions de mots) et écrits universitaires (~ 91 millions de mots). L'équilibre est relatif au sens où l'anglais parlé représente moins de 20% du corpus. Ce défaut est largement compensé par le fait que le COCA est le plus grand corpus d'anglais annoté disponible publiquement et gratuitement via une plateforme de requêtes en ligne. De par sa taille, il permet d'obtenir un nombre représentatif d'occurrences, y compris pour des exemples rares.

À partir de l'extraction d'occurrences du corpus, nous allons procéder à une analyse collexémique covariante de manière à déterminer les paires ADJ-GN pour lesquelles il y a une attraction statistiquement significative dans la construction <ADJ *as* GN>. Nous allons ensuite explorer ces paires à l'aide de deux méthodes multifactorielles exploratoires – la classification ascendante hiérarchique et l'analyse factorielle des correspondances – de manière à repérer des classes sémantiques d'adjectifs et de GN intervenant dans la construction. Nous allons enfin soumettre les paires à une mesure directionnelle, ΔP , afin de révéler les attractions asymétriques entre adjectifs et GN et séparer les paires productives des paires figées. Le but de ces méthodes est de proposer un panorama représentatif des traitements de la fréquence dans un cadre constructionnel et de montrer plus spécifiquement que la construction <ADJ *as* GN> est plus productive que ne le dit Kay (2013).

3.2. Extraction et première analyse

L'extraction nous permet d'obtenir 3 653 occurrences de la construction <ADJ as GN> réparties en 270 types. Le tableau 6 contient les dix types les plus fréquents de la construction (classés par fréquence d'occurrence) ainsi que la fréquence respective des adjectifs et des GN qui les composent.

construction	freq. de la C	freq. de l'ADJ	freq. du GN
<i>mad as hell</i>	181	210	582
<i>tough as nails</i>	98	110	107
<i>white as snow</i>	61	165	61
<i>American as apple pie</i>	60	65	60
<i>cold as ice</i>	59	89	71
<i>clear as a bell</i>	56	162	56
<i>clear as day</i>	53	162	126
<i>good as gold</i>	52	79	58
<i>smooth as silk</i>	52	96	71
<i>white as a sheet</i>	45	165	45

Tableau 6. Les dix types les plus fréquents de <ADJ as GN>.

La construction <ADJ as GN> est très proche de son équivalent en français <ADJ comme GN>. C'est une structure comparative qui opère la mise en relation d'un adjectif (le comparé) et d'un GN (le comparant) de manière à intensifier la valeur de l'adjectif du côté du haut degré. On s'attend donc à ce que les adjectifs qui interviennent dans la construction soient gradables.¹⁵ Conformément à cette attente, la majorité des adjectifs du tableau 6 sont prototypiquement gradables. On notera que même si les adjectifs *American* et *white* ne sont pas naturellement gradables, ils tendent néanmoins vers le haut degré une fois associés au GN. *American as apple pie* aura ainsi le sens de « très américain », renvoyant au cœur de la notion d'« américanité », avec toutes les connotations que cette notion peut avoir (en l'occurrence on s'attend à ce que la tarte aux pommes soit copieusement sucrée et saupoudrée de cannelle, conformément au canon de la pâtisserie américaine).¹⁶ *White* étant un adjectif de couleur, son emploi gradable est contraint (Kleiber, 2007; Whittaker, 2002). On sait pourtant qu'il apparaît volontiers dans des structures d'intensification (Van de Velde, 1995, pp. 147, 157). La disposition des adjectifs à sortir de leur comportement prototypiques est peut-être l'effet de leur combinaison avec une entité référentielle, en l'occurrence le GN. Pour ce qui est de *white as snow* « blanc comme neige » ou *white as a sheet* « pâle comme un linge » (littéralement « blanc comme un drap »), le blanc immaculé de la neige fraîche ou du drap contraste avec la blancheur relative du riz ou des dents par exemple.

En fonction du contexte, un jeu sémantique entre l'adjectif et le GN est possible :

- (3) *If American music is truly as American as apple pie, then it is a pie spiced with multiple ingredients and flavors.* (2008 - ACAD - MusicEduc)
Si la musique américaine est aussi américaine que la tarte aux pommes, c'est une tarte relevée de nombreux ingrédients et de nombreuses saveurs.
- (4) *Makes my skin smooth as silk and irresistible to the touch.* (2011 - FIC - Bk:LiliesInMoonlight)

¹⁵ La gradabilité étant même une propriété définitoire des adjectifs selon Goes (1999).

¹⁶ Voir les remarques socio-historiques de Wierzbicka (1988, pp. 488-491) au sujet de la distinction entre les « adjectifs de nationalité » (ex. *Irish, English, German*) et les « noms de nationalité » (ex. *American, Australian*). Les premiers renverraient à l'« ethnicité » tandis que les seconds seraient plus larges et engloberaient, pour des raisons liées à l'histoire du peuplement, des considérations culturelles :
Words such as *American, Australian* or *Brazilian* tend to be used as adjectives only in combination with nouns which impose a certain qualitative interpretation on the adjective. For example, one can speak of the 'American way of life', 'American culture', 'American cities' or 'American cigarettes', on the assumption that the referents differ in intuitively clear (...) ways from other ways of life, other cultures, other cities, and other cigarettes. But it would be odd to speak of 'American forks', 'American chairs' or 'American leaves', since it is hard to imagine what unique quality could be meant, in such contexts (1988 : 490).

Ainsi ma peau est soyeuse (lit. douce comme la soie) et irrésistible au toucher.

Ce jeu n'est toutefois pas systématique :

- (5) *He's as American as apple pie* (...). (1996 - SPOK - NPR_Weekend)
Plus américain que lui, tu meurs.
- (6) *That implant came out smooth as silk* (...). (2001 - FIC - Bk:DeckHalls)
L'implant est sorti tout seul.

Pour certaines constructions, il semble même que le GN ait atteint un stade avancé de transparence sémantique :

- (7) *Let me ask you a question, are you mad as hell?* (2009 - SPOK - Fox_Hannity)
Permettez-moi de vous poser la question suivante : êtes-vous furieux ?

Pour d'autres constructions, le lien entre l'adjectif et le GN semble motivé par un jeu sur les sons avant même toute considération sémantique. Dans l'exemple (8), ce jeu prend la forme d'une allitération puisque l'adjectif et le GN sont tous deux des monosyllabes commençant respectivement par [g] et [d] :

- (8) a. *Trained nurses are good as gold*. (2004 - MAG - Skiing)
Les infirmières qualifiées valent de l'or.
b. *Cause sometimes I think I'm dumb as dirt*. (2008 - FIC - Bk:FranklyMyDearIm)
Parce que des fois je pense que je suis bête comme mes pieds.

Toutefois, même dans ce cas de figure, le jeu sémantique n'est jamais véritablement absent. Il est même mis en avant si le contexte l'exige, comme en (9) où l'or de *good as gold* « infallible/irréprochable » fait écho à l'or olympique de la gymnaste en question :

- (9) (...) *she's as good as gold; olympic gymnast Dominique Moceanu struts her stuff*. (2000 - SPOK - NBC_Today)
Elle vaut de l'or ; la gymnaste olympique Dominique Moceanu fait le spectacle.

Dans certains cas, le lien sémantique entre l'adjectif et le GN est étroit. C'est le cas des combinaisons métonymiques du type *white as a sheet* (la blancheur est une qualité significative du drap), voire *clear as a bell* (la clarté est une qualité significative du son d'une cloche). Dans d'autres cas, le lien sémantique est beaucoup plus lâche, comme c'est le cas pour *mad as hell* (il est difficile de considérer la colère comme un trait significatif de l'enfer). Ceci dit, quel que soit l'étroitesse du lien sémantique entre les unités qui composent cette construction, c'est le sens contextuel de l'ensemble qui est déterminant. Par conséquent, nous sommes bien dans une problématique constructionnelle au sens où l'entendent les grammairiens de constructions : en contexte, le sens d'une construction n'est pas toujours réductible au sens des éléments qui la composent. Il est donc difficile de généraliser au niveau de l'adjectif ou du GN sur la seule base du tableau 6.

3.3. L'analyse collexémique covariante

Dans la mesure où il existe une interdépendance entre l'adjectif et le groupe nominal dans la construction <ADJ as GN>, la première étape consiste à s'affranchir des fréquences brutes pour ne retenir que les paires statistiquement significatives. C'est précisément ce qui permet de faire l'analyse collexémique covariante (ci-après ACC). À la différence d'une analyse fondée sur un comptage absolu, l'ACC identifie les combinaisons qui apparaissent de manière statistiquement significative, c'est à dire dont la fréquence attestée est supérieure à la fréquence attendue. Pour obtenir la fréquence attendue, l'ACC met en regard la fréquence absolue des combinaisons ADJ-GN attestées dans la construction <ADJ as GN> avec la fréquence respective des adjectifs et des GN. Pour chaque paire ADJ-GN, l'ACC soumet une tabulation sous la forme du Tableau 5 à l'une des trois mesures d'association figurant dans le Tableau 1 (nous avons choisi le test exact de Fisher)¹⁷. Nous obtenons le

¹⁷ Le choix d'une mesure d'association plutôt qu'une autre est affaire de subjectivité. Ces mesures ne se recouvrent pas exactement, mais aucune ne peut être considérée comme nettement supérieure à une autre.

Tableau 7.¹⁸ Celui-ci indique, pour chaque paire ADJ-GN, la fréquence de l'adjectif dans la construction, la fréquence du GN, la fréquence observée de la combinaison ADJ-GN dans la construction, la fréquence attendue de cette combinaison, le type de relation entre l'adjectif et le GN dans la construction (attraction ou répulsion), et la force collostructionnelle (qui mesure le degré d'attraction ou de répulsion en fonction de la mesure d'association choisie).

ADJ	GN	freq. ADJ	freq GN	obs ADJ_GN dans C	att ADJ_GN dans C	relation	coll.strength
<i>tough</i>	<i>nails</i>	110	107	98	3.21	attraction	166.72
<i>American</i>	<i>apple pie</i>	65	60	60	1.06	attraction	124.84
<i>mad</i>	<i>hell</i>	210	582	181	33.29	attraction	122.49
<i>cold</i>	<i>ice</i>	89	71	59	1.72	attraction	93.26
<i>good</i>	<i>gold</i>	79	58	52	1.25	attraction	88.71
<i>white</i>	<i>snow</i>	165	61	61	2.74	attraction	87.51
<i>clear</i>	<i>bell</i>	162	56	56	2.47	attraction	80.41
<i>free</i>	<i>bird</i>	45	35	35	0.43	attraction	75.18
<i>smooth</i>	<i>silk</i>	96	71	52	1.86	attraction	72.8
<i>dead</i>	<i>doornail</i>	27	27	27	0.2	attraction	68.17
<i>sick</i>	<i>dog</i>	27	27	27	0.2	attraction	68.17
<i>smart</i>	<i>whip</i>	42	31	31	0.35	attraction	66.91
<i>clean</i>	<i>whistle</i>	25	25	25	0.17	attraction	63.89
<i>white</i>	<i>sheet</i>	165	45	45	2.02	attraction	63.39
<i>high</i>	<i>kite</i>	31	26	26	0.22	attraction	60.81
<i>happy</i>	<i>clam</i>	48	29	29	0.38	attraction	59.32
<i>solid</i>	<i>rock</i>	36	49	31	0.48	attraction	57.91
<i>stiff</i>	<i>board</i>	25	32	25	0.22	attraction	57.37
<i>neat</i>	<i>pin</i>	22	22	22	0.13	attraction	57.35
<i>pretty</i>	<i>picture</i>	22	22	22	0.13	attraction	57.35

Tableau 7. Les 20 paires de collexèmes covariants les plus spécifiques de la construction <ADJ as GN>.

Attardons-nous un instant sur le véritable apport de l'ACC (Tableau 7) sur les fréquences brutes (Tableau 6). À première vue, la différence entre les deux classements est infime. On retrouve les mêmes paires ADJ-GN parmi les dix les plus associées. Seul l'ordre du classement change. Le graphique en Figure 1 croise la fréquence brute et la force collostructionnelle pour chacune des constructions attestées dans le corpus. Si la corrélation était strictement linéaire, toutes les constructions seraient alignées sur la droite de régression (droite pleine). Ce n'est pas vraiment le cas, à en juger par l'éloignement progressif vis-à-vis de la droite de régression des constructions prenant des valeurs élevées sur les deux axes (en particulier *American as apple pie*, *tough as nails* et *mad as hell*). Ceci est un effet du calcul de la force collostructionnelle, en l'occurrence l'inverse de \log_{10} de la *p*-valeur obtenue par un test exact de Fisher unilatéral. De fait, la courbe Lowess (en pointillés)¹⁹ porte la trace de la courbe logarithmique décrite ci-dessus au niveau des paires ADJ-GN les plus fortement associées.

¹⁸ L'ACC a été réalisée sous R (R Core Team, 2013) avec le script Coll.analysis 3.2 (Gries, 2007).

¹⁹ La courbe Lowess (abréviation de *locally weighted scatterplot smoothing*) est un lisseur non paramétrique obtenu par régression par polynômes locaux itérée (Cornillon & Matzner-Løber, 2010). Par rapport à une droite de régression traditionnelle, la courbe Lowess s'ajuste de manière plus souple à l'hétérogénéité des points sur un diagramme tel que celui de la Figure 1.

3.4. La classification ascendante hiérarchique à partir d'une ACC

Nous cherchons à savoir s'il existe des régularités dans le choix des adjectifs en fonction du GN et vice-versa. En nous inspirant de la méthode de Gries and Stefanowitsch (2010), nous sélectionnons les types de paires ADJ-GN dont la force collocationnelle et la fréquence de cooccurrence sont les plus grandes. Nous obtenons un tableau de contingence croisant les fréquences de cooccurrence de 30 adjectifs et de 34 GN. Nous soumettons ce tableau à la classification ascendante hiérarchique (Everitt, Landau, Leese, & Stahl, 2011, section 4.2). Cette méthode requiert une matrice de distances. On obtient cette matrice en convertissant le tableau de fréquences en distances à l'aide d'une mesure définie par l'utilisateur.²¹ Les éléments de la matrice de distances sont ensuite agrégés en clusters à l'aide d'un indice d'agrégation.²² Les résultats sont représentés sous la forme d'un dendrogramme. A l'aide du même tableau de contingence, nous avons généré deux dendrogrammes.²³

Le premier dendrogramme (Figure 2) représente la classification des 30 adjectifs en fonction des 34 GN avec lesquels ils se combinent. Il se lit du bas vers le haut.²⁴ Les individus les plus proches s'amalgament en clusters en premier. Idéalement, les clusters les mieux structurés regroupent les mots par unités de sens, ce qui permet une lecture sémantologique du graphique. Parmi les premiers clusters, on trouve les synonymes et quasi-synonymes suivants : *quiet* et *silent* (« silencieux », cluster 1), *black* et *dark* (« noir » / « sombre », cluster 2), *skinny* et *thin* (« maigre », cluster 3), *smooth* et *soft* (« doux », cluster 5), *bright* et *clear* (cluster 7), *fine* et *handsome* (« fin, raffiné » / « beau » cluster 9) et *big* et *large* (« grand », cluster 10). Lors que les individus ne sont pas des synonymes, on devine qu'ils sont reliés par les multiples propriétés du GN avec lequel ils se combinent (ex. cluster 6 : *cold/hard as stone*).

Le second dendrogramme (Figure 3) représente la classification des 34 GN en fonction des 30 adjectifs qu'ils servent à intensifier dans la construction. Le lien entre les GN regroupés sous les mêmes clusters est moins étroit qu'avec les adjectifs de la Figure 2. Ils appartiennent toutefois au mêmes domaines conceptuels : *rail* et *stick* sont des objets allongés et fins (cluster 3), *horse* et *mule* sont des équidés (cluster 4), *sky* fait partie de *world* et sont tous deux de grande dimension (cluster 5), *tomb* est un aspect de *death* (cluster 11), *glass* est un contenant de *water* (cluster 12), *hell* et *sin* relèvent initialement du domaine religieux (cluster 13), *sun* est la cause de *day* et *daylight* (clusters 14 et 16). D'autres GN sont reliés par la polysémie de l'adjectif. On note que *honey* « miel » s'amalgame avec *cotton* « coton », les deux étant reliés par le GN *thick* « épais ».

²¹ Il existe plusieurs mesures de distance : euclidienne, Canberra, city-block (Manhattan), Tchebychev, etc. Nous avons retenu Canberra car c'est elle qui gère le mieux les matrices creuses, i.e. les matrices contenant de nombreux effectifs faibles voire nuls.

²² Il existe plusieurs indices : saut minimum, saut maximum, moyenne pondérée des groupes associés, etc. Nous avons retenu la méthode de Ward (1963), qui évalue la distance entre plusieurs clusters à l'aide de l'analyse de variance. Cette méthode a l'avantage de générer des clusters de taille moyenne.

²³ Pour cela, nous avons utilisé le package *pvclust* pour R (<http://www.is.titech.ac.jp/~shimo/prog/pvclust/>).

²⁴ Chaque cluster est accompagné de trois nombres. Celui du bas indique le rang du cluster (du premier cluster généré par la classification au 28^e). Les deux nombres au-dessus sont des degrés de confiance. Le nombre de droite est obtenu par la méthode BP (*bootstrap probability*) tandis que celui de gauche est obtenu par la méthode AU (*approximately unbiased*), qui est plus rigoureuse que la première. Plus la valeur de BP ou AU est proche de 100 et plus le cluster est « solide ».

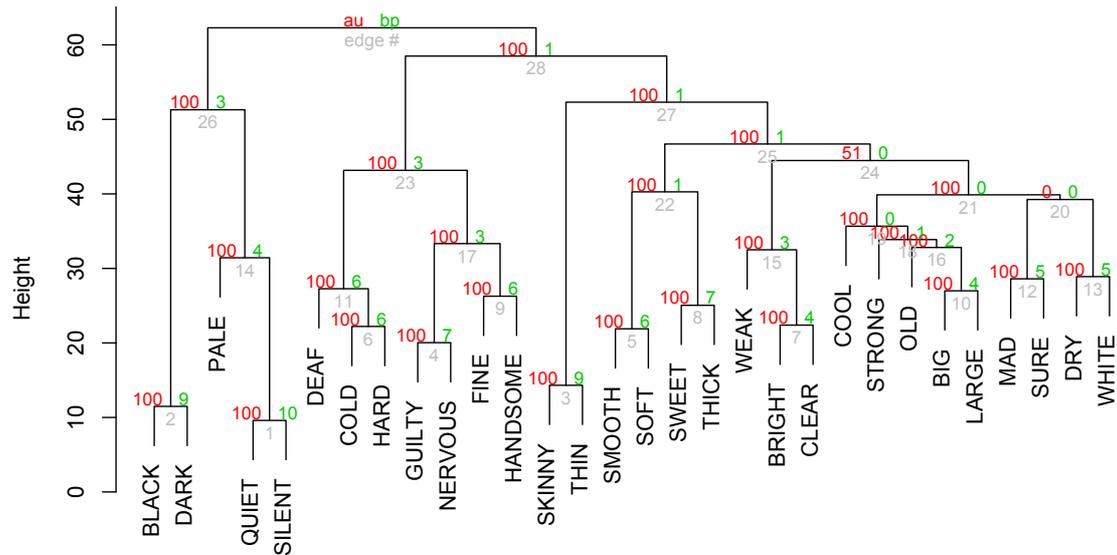


Figure 2. Classification ascendante hiérarchique de 30 adjectifs classés en fonction des 34 GN avec lesquels ils se combinent dans la construction.

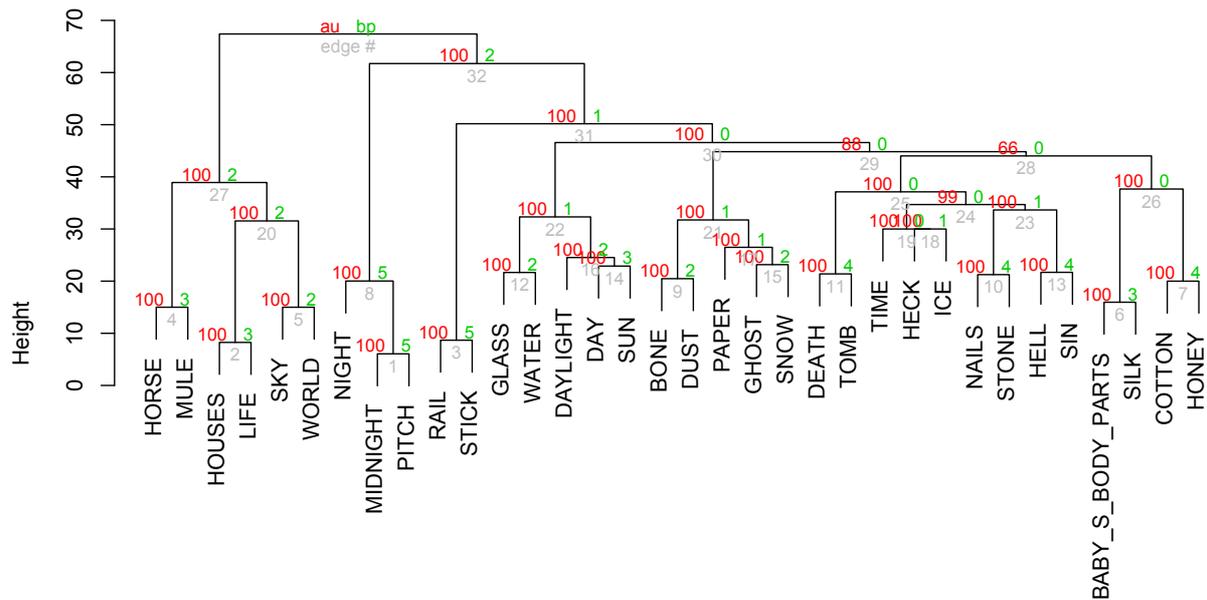


Figure 3. Classification ascendante hiérarchique de 34 GN classés en fonction des 30 adjectifs avec lesquels ils se combinent dans la construction.

Nos classifications ont trois défauts: (a) il est d'usage de classer des individus au regard de plusieurs types de variables, alors que nous ne l'avons fait qu'au regard d'une seule, (b) nous n'avons pris en compte que 30 des 121 adjectifs et 34 des 171 GN et (c) en séparant adjectifs (Figure 2) et GN (Figure 3), nous ne pouvons pas corrélérer leurs profils sémantiques respectifs (nous proposons une solution dans la sous-section suivante). En dépit de ces faiblesses, des régularités émergent tant au niveau des adjectifs qu'au niveau des GN. Parce que la construction <ADJ as GN> intensifie des familles d'adjectifs sur la base de familles de GN, son comportement est plus général et régulier que ne le laisse entendre Kay (2013).

3.5. L'analyse factorielle des correspondances à partir d'une ACC

Nous souhaitons ici compenser une lacune de la CAH en comparant d'un même élan les profils des adjectifs et les profils des GN. L'extraction initiale nous a permis d'obtenir tous les types de la construction dans le COCA et leurs fréquences respectives dans les sous-parties du corpus (anglais parlé, fiction, magazines, presse, anglais universitaire). Nous pouvons croiser cette extraction avec les données de l'ACC pour ne plus retenir que les adjectifs et les GN pour lesquels il y a une attraction. Nous obtenons le Tableau 8. Celui-ci comporte :

- 5 colonnes correspondant aux variables contextuelles ;
- 291 lignes correspondant aux 170 GN et 121 adjectifs des constructions pour lesquelles il y a une attraction significative selon l'ACC (une ligne par individu).

	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
GN_AIR	1	15	8	2	5
GN_APPLE_PIE	23	4	18	11	4
GN_BABY	0	11	0	0	0
GN_BABY_S_BODY_PARTS	2	16	8	6	0
GN_BARN	0	4	2	0	0
GN_BASEBALL	0	1	1	3	0
GN_BAT	4	12	2	0	0
GN_BEET	1	4	0	0	0
GN_BELL	11	24	11	7	3
GN_BIBLE	1	0	2	3	0
-----	-----	-----	-----	-----	-----
ADJ_AMERICAN	23	5	19	14	4
ADJ_ANCIENT	1	2	1	0	1
ADJ_ANGRY	1	5	0	1	0
ADJ_ANNoyING	2	8	1	1	0
ADJ_BALD	0	3	1	1	0
ADJ_BIG	9	105	32	10	2
ADJ_BLACK	4	94	14	2	3
ADJ_BLIND	4	12	2	0	0
ADJ_BLUE	1	8	3	0	1
ADJ_BOLD	1	6	3	4	0
-----	-----	-----	-----	-----	-----
...

Tableau 8. Fréquences des composants adjectivaux et nominaux de <ADJ as GN> en fonction de 5 variables contextuelles (aperçu).

Pour synthétiser ce tableau mathématiquement et graphiquement, nous le soumettons à l'analyse factorielle des correspondances (AFC) (Benzécri, 1984; Greenacre, 2007; Husson, Le, & Pages, 2011). Tout comme la CAH, l'AFC est une méthode multifactorielle exploratoire. Elle résume des tableaux de données de manière à croiser les profils d'individus (lignes) avec les profils de variables (colonnes). Le profil d'une ligne s'obtient en divisant la fréquence de chacune de ses cellules par la somme des fréquences de la ligne. De la même manière, le profil d'une colonne s'obtient en divisant la fréquence de chacune de ses cellules par la somme des fréquences de la colonne.²⁵ Ces profils ont des propriétés géométriques que l'AFC convertit en coordonnées de manière à les visualiser sur un espace euclidien à deux dimensions. La Figure 4 est le résultat d'une telle conversion. Nous n'y faisons apparaître que les individus contribuant le plus à la construction de l'axe horizontal (dimension 1) et de l'axe vertical (dimensions 2)²⁶ de manière à ne pas rendre le graphe illisible (les points moins dont la contribution est moindre sont sans étiquette). La lecture d'un graphique d'AFC est simple : plus deux points sont proches dans l'espace, plus il y a une corrélation entre les deux.²⁷ Par ailleurs, l'intersection des deux axes signale le profil moyen des lignes et des colonnes. Ce point de repère est utile car il plus un point en est éloigné, plus son profil est atypique.

²⁵ Parce que ce n'est pas l'objet de cet article, nous n'entrerons pas plus avant dans les détails de l'AFC. Nous renvoyons le lecteur à Desagulier (sous presse) et Desagulier (soumis).

²⁶ Pour une définition mathématique de la contribution d'un point à la construction des axes, voir Husson et al. (2011, pp. 81-82). Seuls les 40 points contribuant le plus à la construction des deux dimensions ont été projetés.

²⁷ A ceci près que le graphique ne fait apparaître que deux dimensions sur cinq (voir note suivante). Certains points en apparence proches peuvent être éloignés si on les représente dans un espace à trois dimensions.

Observons à présent les individus (adjectifs et GN). Certains d'entre eux affichent une préférence contextuelle. Ainsi, l'intensification des adjectifs dénotant la noirceur (*black*), l'obscurité (*dark*) ou au contraire la blancheur (*white*) et la pâleur (*pale*) sont une spécificité de la variable FICTION. L'intensification de l'adjectif *old* est quant à elle une spécificité quasiment exclusive des variables ACADEMIC et MAGAZINE (en combinaison avec *time*, *warfare*, *humanity*, *civilization*, *history* et *man*).

Par ailleurs, la proximité spatiale de certains individus sur le graphique fait apparaître des associations dont le degré de figement est élevé :

- FICTION (moitié gauche) : *dark as night*, *black as the night* ;
- ACADEMIC/MAGAZINE (en haut à droite) : *different as night and day*, *lonely as a cloud*, *old as time*, *queer as folk*, *good as gold* ;
- SPOKEN/NEWSPAPER (en bas à droite) : *American as apple pie*, *solid as a rock*, *tough as nails*, *mad as heck*.

L'intérêt de ce type de graphique en deux dimensions est précisément de fournir une manière d'indexer le degré de figement sur la distance entre les points. En mesurant l'espace entre adjectifs et GN, on constate par exemple que l'expression *American as apple pie* est plus figée que *good as gold* et bien plus encore que *good as sex*.

Notre AFC n'offre qu'un aperçu limité de la distribution des adjectifs et des GN en fonction du contexte (sachant que nos variables contextuelles sont tout aussi limitées). Toutefois, cet aperçu est suffisant pour faire émerger des niveaux de régularité dans les appariements ADJ-GN au sein de la construction <ADJ as GN>.

3.6. Une mesure directionnelle appliquée à l'ACC : ΔP

L'ACC et l'ensemble des mesures d'association, partent du principe implicite que l'attraction entre deux unités linguistiques est symétrique. Rien n'est moins sûr. À titre d'exemple, si l'on extrait au hasard un mot d'un corpus et que ce mot est *ipso*, on a de très fortes chances pour que *facto* apparaisse juste après. On estime qu'*ipso* est un bon prédicteur de *facto*. Réciproquement, *facto* n'est pas un aussi bon prédicteur d'*ipso* car *facto* peut aussi être précédé de *de* dans la locution *de facto*. L'attraction entre *ipso* et *facto* est dite « asymétrique » ou « directionnelle ». Plus un lexème est un bon prédicteur d'un autre lexème, moins il s'emploie dans d'autres contextes, par conséquent moins il est productif. L'intérêt des associations asymétriques est donc de fournir une mesure de la productivité d'une paire de lexèmes.

Concernant la construction <ADJ as GN>. L'ACC que nous avons réalisée confond deux types de probabilités : $p(\text{ADJ}|\text{GN})$, à savoir la probabilité de tel adjectif sachant que l'on a tel GN, et $p(\text{GN}|\text{ADJ})$, à savoir la probabilité de tel GN sachant que l'on a tel adjectif. Parmi les mesures permettant de différencier ces probabilités, la plus compatible avec une approche psycholinguistiquement réaliste de la langue est ΔP , une mesure directionnelle issue de la théorie de l'information (Allan, 1980). D'abord décrite par Ellis (2006) dans le domaine de l'acquisition, ΔP prend comme point de départ un tableau de contingence classique (Tableau 9) :

	O	¬O
C	a	b
¬C	c	d

Tableau 9. Tableau d'entrée générique pour le calcul de ΔP .

Dans ce tableau, O représente le résultat (*outcome*), C l'indice (*cue*), et *a*, *b*, *c*, et *d* sont des fréquences (par exemple *a* est la fréquence de la conjonction du résultat et de l'indice, *c* est la fréquence du résultat sans la présence de l'indice). ΔP se calcule de la manière suivante :

$$(i) \quad \Delta P = \frac{p(O|C) - p(O|\neg C)}{a/(a+b) - c/(c+d)} = \frac{(ad - bc)/[(a+b)(c+d)]}{a/(a+b) - c/(c+d)}$$

Ellis (2006, p. 11) résume cette formule ainsi :

ΔP est la probabilité d'obtenir le résultat sachant que l'indice est présent $P(O|C)$ moins la probabilité du résultat en l'absence de l'indice $P(O|\neg C)$. Lorsque ces probabilités sont identiques (...) il n'y a pas de covariation entre les deux événements et $\Delta P = 0$. On considère que plus ΔP est proche de 1.0, plus la présence de l'indice augmente la probabilité du résultat, alors que plus ΔP est proche de -1.0, plus la présence de l'indice diminue la probabilité du résultat (...).

On doit à Gries (2013) d'avoir transposé cette mesure aux collocations. Nous souhaitons ici la transposer aux collocations. Dans le cadre de la construction <ADJ as GN>, nous devons calculer deux valeurs de ΔP :

$$\begin{aligned} \text{(ii)} \quad \Delta P_{GN|ADJ} &= p(GN|ADJ) - p(GN|\neg ADJ) \\ &= a/(a + b) - c/(c + d) \\ \text{(iii)} \quad \Delta P_{ADJ|GN} &= p(ADJ|GN) - p(ADJ|\neg GN) \\ &= a/(a + c) - b/(b + d) \end{aligned}$$

Dans la première, le GN est le résultat et l'adjectif l'indice. Si $0 < \Delta P_{GN|ADJ} \leq 1$, alors l'adjectif est un prédicteur du GN. Dans la seconde, l'adjectif est le résultat et le GN l'indice. Si $0 < \Delta P_{ADJ|GN} \leq 1$, alors le GN est un prédicteur de l'adjectif. Prenons l'exemple de *mad as a hatter* (lit « fou comme un chapelier »)²⁹ et appliquons les formules (ii) et (iii) au Tableau 10 en (iv) et (v) respectivement. Le résultat est frappant : *mad* n'est pas un bon indice de *hatter* tandis que *hatter* est un excellent indice de *mad*. L'exemple est extrême puisque *hatter* n'intervient dans la construction <ADJ as GN> qu'avec *mad* tandis que *mad* intervient en conjonction avec d'autres GN (*mad as a hornet*, *mad as hell*, *mad as heck*). On devine donc que dans la construction, et dans le corpus, si le GN est *hatter*, l'adjectif sera nécessairement *mad*. En dépit de l'évidence, cette asymétrie est absente de l'ACC, qui ne révèle qu'une attraction bidirectionnelle entre les deux lexèmes dans la construction.

	<i>hatter</i> : présent	<i>hatter</i> : absent	totaux
<i>mad</i> : présent	14	196	210
<i>mad</i> : absent	0	464 020 046	464 020 046
totaux	14	464 020 242	464 020 256

Tableau 10. Tableau de contingence pour *mad as a hatter* dans le COCA.

$$\begin{aligned} \text{(iv)} \quad \Delta P_{hatter|mad} &= p(hatter|mad) - p(hatter|adjectifs autres que mad) \\ &= 14/(14 + 196) - 0/(0 + 463\,020\,046) \\ &\approx 0,067 \\ \text{(v)} \quad \Delta P_{mad|hatter} &= p(mad|hatter) - p(mad|GN autres que hatter) \\ &= 14/(14 + 0) - 196/(196 + 463\,020\,046) \\ &\approx 1 \end{aligned}$$

Plus une valeur de ΔP est faible, plus le lexème en position de résultat rend la construction productive (car plus il apparaît en conjonction avec d'autres lexèmes en position d'indice). Inversement, plus une valeur de ΔP est élevée, plus c'est le lexème en position d'indice qui rend la construction productive. L'apport de cette mesure directionnelle est considérable dans le cadre d'une étude de cas comme la nôtre puisque la productivité de la construction peut être fonction soit de l'adjectif, soit du GN. Le profil de la construction au regard de la productivité est donc plus complexe que Kay (2013) ne le pense.

Afin de montrer les asymétries dans l'attraction collocationnelle à l'œuvre dans la construction <ADJ as GN>, nous avons calculé $\Delta P_{GN|ADJ}$ et $\Delta P_{ADJ|GN}$ pour l'ensemble des paires ADJ-GN affichant une attraction au prisme de l'ACC. ΔP indique que bon nombre de collocations sont asymétriques (Figure 5). En effet, sur les 265 types que compte la construction <ADJ as GN>, 132 affichent une différence importante en termes de valeurs de ΔP ($\geq 0,5$ ou $\leq -0,5$). Seuls 24 affichent une différence

²⁹ En référence au chapelier fou dans *Alice au pays des merveilles* de Lewis Carroll. On trouve un équivalent en français : *travailler du chapeau*. La folie des chapeliers serait due aux produits toxiques utilisés dans la confection des chapeaux de feutre, notamment le mercure (Waldron, 1983).

nulle (soit une absence de covariation selon Ellis). On observe un plus grand nombre de cas pour lesquels le GN est un meilleur prédicteur de l'adjectif que vice versa (67% contre 33%).

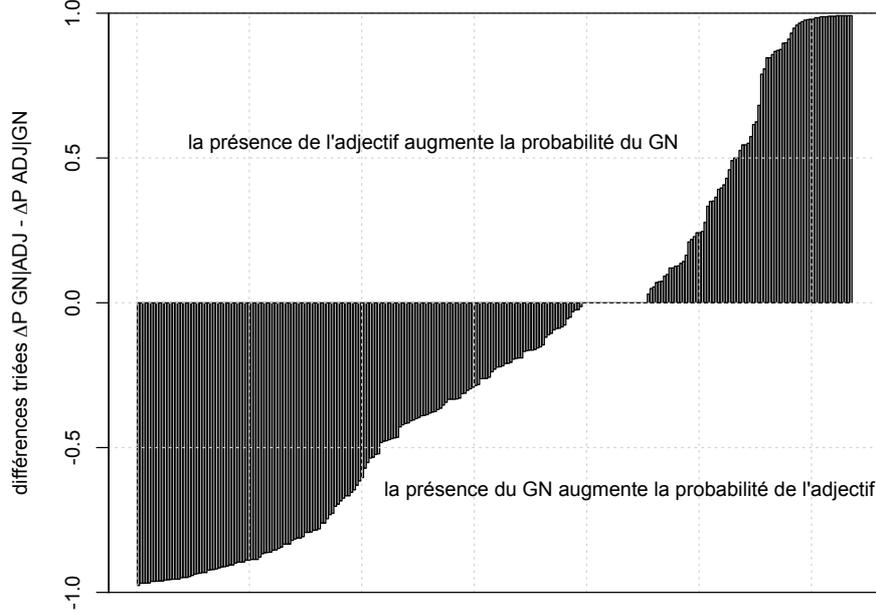


Figure 5. Tri des différences par paires des valeurs de ΔP

La Figure 6 donne une image plus détaillée des asymétries. Les valeurs de $\Delta P_{ADJ|GN}$ figurent sur l'axe des abscisses, et celles de $\Delta P_{GN|ADJ}$ sur l'axe des ordonnées. Chaque cercle représente une paire ADJ-GN, la taille du cercle étant proportionnelle à sa fréquence. Les cercles qui se recouvrent sont caractérisés par des nuances de gris.

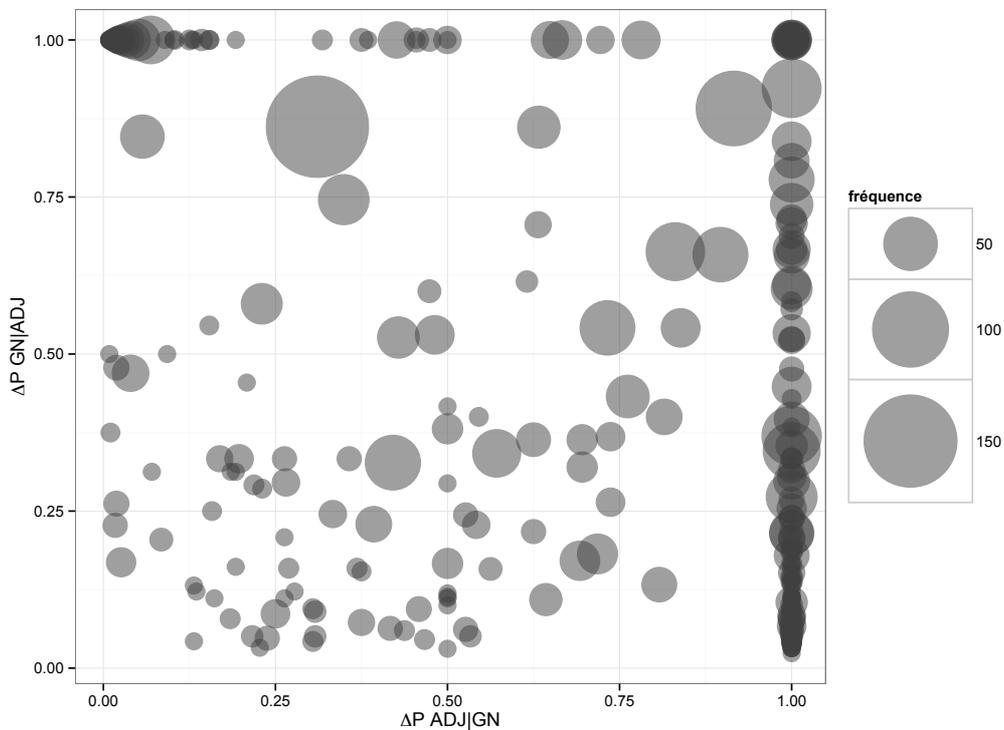


Figure 6. Croisement des valeurs de $\Delta P_{GN|ADJ}$ et $\Delta P_{ADJ|GN}$

Pour un grand nombre de paires ADJ-GN, l'asymétrie est flagrante, à en juger par la concentration importante de cercles de diamètre plus ou moins grands sur des valeurs de ΔP très proches de ou égales à 1. Le cercle le plus gros (quart supérieur gauche du diagramme) correspond à la paire *mad-hell* dans *mad as hell*. L'ACC a montré qu'en plus d'une haute fréquence de co-occurrence entre *mad*

et *hell* dans la construction (181 occurrences), il y avait également une attraction significative entre ces deux lexèmes (coll.strength = 122.49, Tableau 7). Nous voyons à présent que ni la fréquence brute, ni la force collostructionnelle ne rend compte du fait que *mad* attire bien plus *hell* que *hell* n'attire *mad*.

Pour apprécier dans le détail la nature des différences, nous choisissons d'observer la distribution des paires ADJ-GN à l'aune de la mesure d'association propre à l'ACC (logarithme décimal du rapport de log-vraisemblance)³⁰ et des différences par paires des valeurs de ΔP . Cette distribution est visible en Figure 7 et Figure 8 (la seconde est un zoom de la première afin de permettre une lecture plus aisée des valeurs extrêmes). À chaque fois, l'adjectif est au dessus du GN.

Dans chacune des deux figures, le diagramme de gauche représente les paires pour lesquelles le GN est un meilleur prédicteur de l'adjectif que vice-versa. On y retrouve des adjectifs à forte (Figure 7) voire très forte productivité (Figure 8). Parmi ces adjectifs très productifs dans la construction, certains dénotent des qualités « primaires », ayant trait au domaine physique, telles que l'âge illustre, la couleur, la taille, la force :

- old* *old as war/warfare* « vieux comme la guerre », *old as the Bible* « vieux comme la bible », *old as history* « vieux comme l'histoire », *old as politics* « vieux comme la politique », *old as the century* « vieux comme le siècle », *old as man/mankind/humanity* « vieux comme l'humanité », *old as civilization* ;
- black* *black as ebony* « noir comme l'ébène », *black as soot* « noir comme la suie », *black as tar* « noir comme le goudron », *black as ink* « noir comme l'encre », *black as obsidian* « noir comme l'obsidienne » ;
- white* *white as ghost* « blanc comme un spectre », *white as snow* « blanc comme neige », *white as sheet* « pâle comme un linge », *white as chalk* « blanc comme la craie », *white as milk* « blanc comme le lait » ;
- big* *big as life* « grandeur nature », *big as saucers* « grand comme des soucoupes », *big as mountains/a mountain* « grand comme une montagne », *big as quarters* « gros comme des pièces de 25 cents », *big as softballs* « gros comme des balles de softball », *big as a barn* « gros/grand comme une grange », *big as the Earth* « gros comme la Terre », *big as the universe* « grand comme l'univers » ;
- strong* *strong as a bull* « fort comme un taureau », *strong as steel* « solide comme l'acier », *strong as iron* « solide comme le fer », *strong as pillars* « solide comme des piliers » ;
- sharp* *sharp as razors/a razor* « aiguisé comme un rasoir », *sharp as a knife* « aiguisé comme un couteau », *sharp as a tack* « pointu comme une punaise ».

Certains GN sont recrutés par la construction pour une de leurs qualités intrinsèques : la hauteur d'un arbre (*tall as trees*), la noirceur du charbon (*black as coal*), la solidité du cuir (*tough as leather*), la légèreté de la plume (*light as feathers/a feather*), la grandeur de l'univers (*big as the universe*), la blancheur de la neige ou de la craie (*white as snow/chalk*), le goût du sucre (*sweet as sugar*)³¹, ou l'« américanité » du baseball (*American as baseball*). Deux GN sont recrutés pour leur valeur intensive, avant leur valeur référentielle, en particulier lorsqu'ils intensifient l'adjectif épistémique *sure* : *sure as shootin'*, *sure as shit*.

³⁰ Le rapport de log-vraisemblance (Dunning, 1993) est une mesure très proche du test exact de Fisher. Nous pouvons donc considérer le logarithme décimal du rapport de log-vraisemblance comme une mesure de la force collostructionnelle quasiment identique à celle de notre ACC initiale.

³¹ Avec une polysémie possible de *sweet* (« sucré », « doux », « tendre », « attendrissant »).

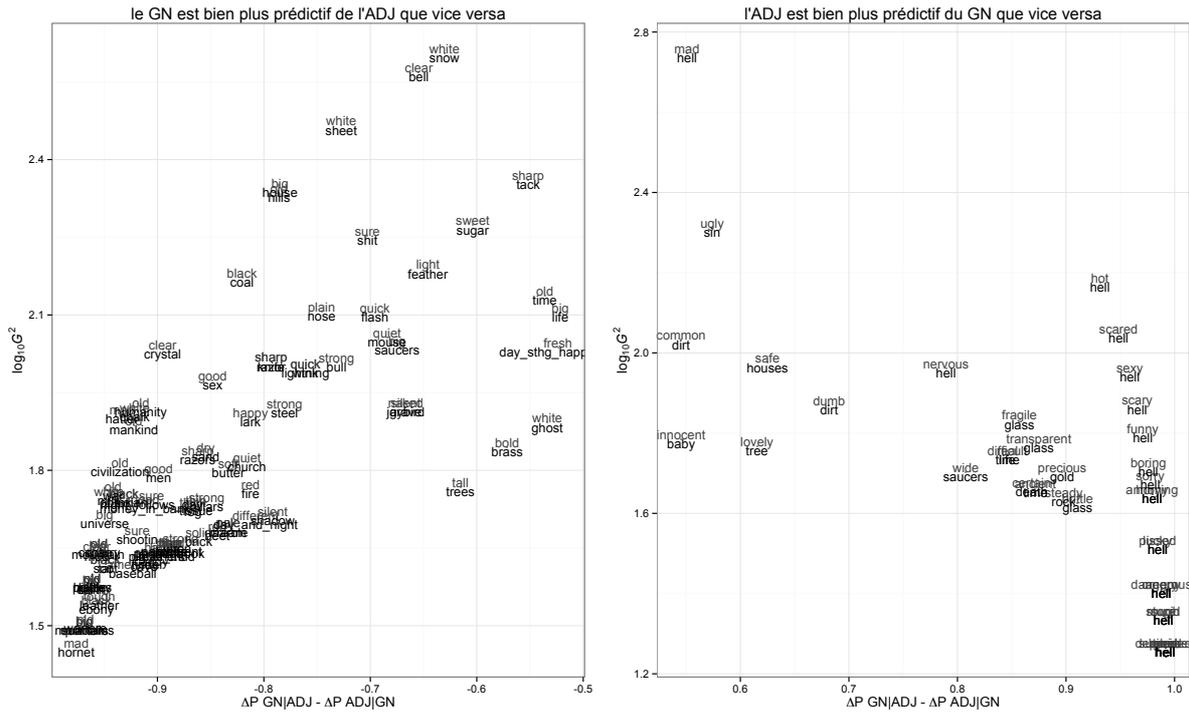


Figure 7. Distribution des paires ADJ-GN selon le rapport de log-vraisemblance (en ordonnée) et les moitiés supérieures des différences $\Delta P_{GN|ADJ} - \Delta P_{ADJ|GN}$ (en abscisse) ($\leq -0,5$ à gauche et $\geq 0,5$ à droite).

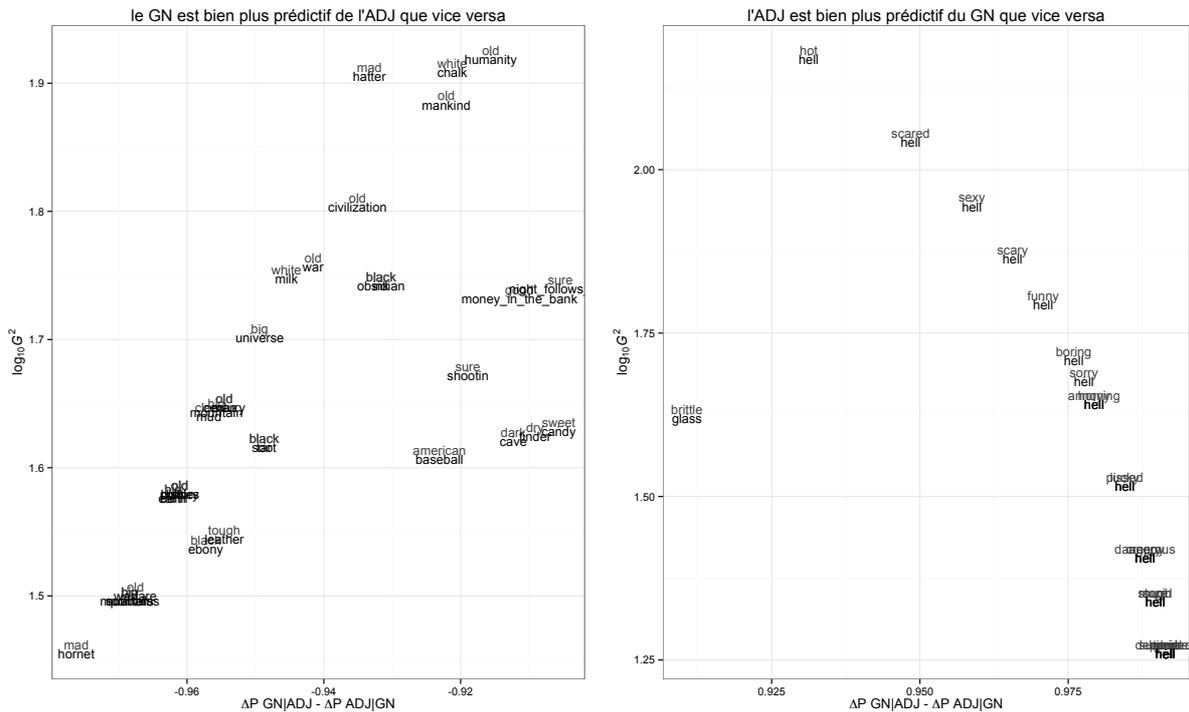


Figure 8. Distribution des paires ADJ-GN selon le rapport de log-vraisemblance (en ordonnée) et les deux extrêmes des différences $\Delta P_{GN|ADJ} - \Delta P_{ADJ|GN}$ (en abscisse, $\leq -0,9$ à gauche et $\geq 0,9$ à droite).

Les diagrammes de droite représentent les paires pour lesquelles l'adjectif est un meilleur prédicteur du GN que vice-versa. On retrouve certains GN des diagrammes de gauche pour intensifier cette fois des adjectifs sémantiquement plus spécifiés, par exemple *saucers* (*wide as saucers*, vs. *big as saucers*) et *gold* (*precious as gold*, vs. *good as gold*). *Tree*, qui lorsqu'il est prédicteur de l'adjectif renvoie à sa grandeur, est ici retenu pour son aspect plaisant (*lovely*). Les GN qui figurent sur le diagramme de droite ne sont pas tous productifs de la même manière. Certains conservent leur sens lexical. C'est le cas de *glass*, qui intensifie la nature cassante (*brittle*), transparente (*transparent*) ou fragile (*fragile*) du GN modifié par l'adjectif. Par contraste, *hell* « enfer » s'est grammaticalisé et n'est employé que pour sa fonction intensive. C'est de loin le GN le plus productif de la construction (Figure 8). Il s'emploie dans la construction <ADJ as GN> pour intensifier des adjectifs dénotant des états psychologiques et physiologiques (*glad*, *surprised*, *angry*, *pissed*, *horny*, *sorry*, *scared*, *nervous*, *bored*, *depressed*, *tired*, *stupid*, *lucky*), divers stimuli (*annoying*, *boring*, *creepy*, *dangerous*, *eerie*, *funny*, *hot*³², *scary*, *sexy*), ainsi que des perceptions physiques et psychologiques (*rough*, *sore*). Pour certains d'entre eux, il est possible de procéder à des regroupements par domaines conceptuels :

ennui	<i>bored, boring</i>
colère	<i>angry, pissed</i>
étrangeté	<i>eerie, creepy</i>
séduction, excitation	<i>sexy, hot, horny</i>

Cet emploi quasi-grammatical de *hell* contraste avec la coloration sémantique des autres constructions figurant sur le même diagramme : *safe as houses* « très sûr » (jeu de mot avec *safe house* « planque, cachette »), *common as dirt* « très banal », *innocent as a baby* « l'innocence même ». Pour deux constructions en apparence très proches – ex. *mad as a hornet* et *mad as hell* « fou/folle de rage » – on a deux comportements bien distincts en termes d'attraction : *mad* est un bien meilleur prédicteur de *hell* que de *hornet* « frelon ».

Enfin, il est utile d'examiner les paires ADJ-GN pour lesquelles la différence $\Delta P_{GN|ADJ} - \Delta P_{ADJ|GN}$ est faible (Figure 9).

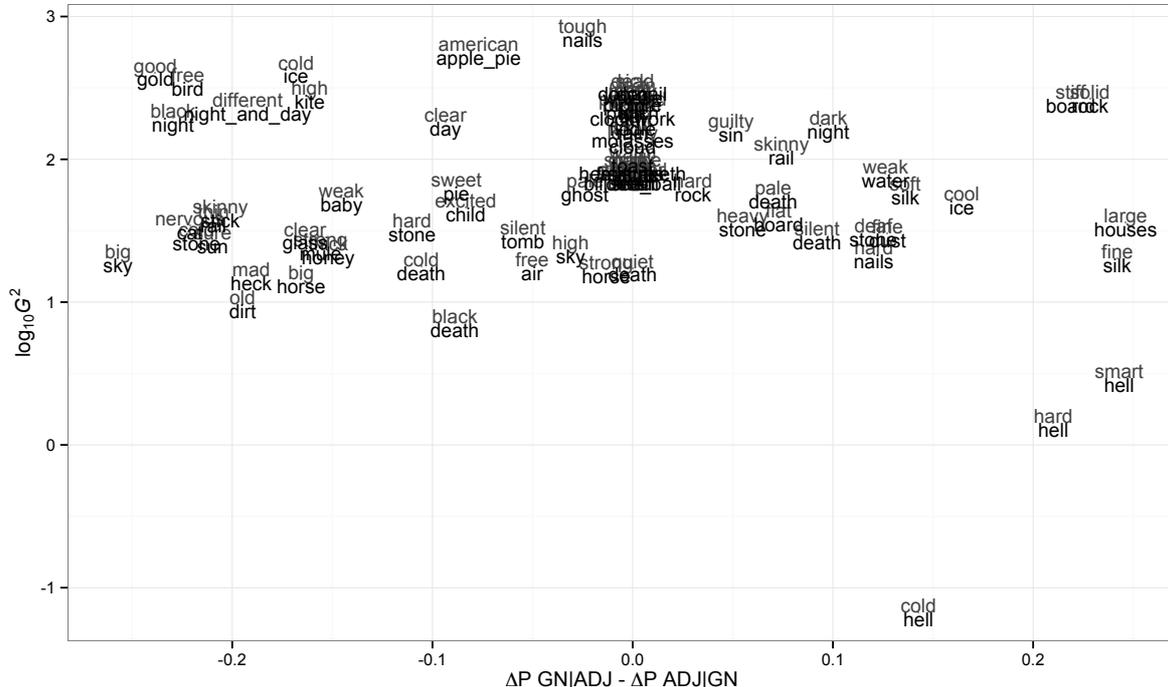


Figure 9. Distribution des paires ADJ-GN selon le rapport de log-vraisemblance (en abscisse) et les deux extrêmes des différences $\Delta P_{GN|ADJ} - \Delta P_{ADJ|GN}$ (en ordonnée, $\leq -0,25$ à gauche et $\geq 0,25$ à droite).

³² Au sens de « sexy ».

Pour ces paires, il y a une relative absence de covariation car ΔP est proche de 0. On trouve plusieurs paires caractérisées par une forte attraction (cf. Tableau 7) : *tough as nails* (coll.strength = 166,72), *American as apple pie* (coll.strength = 124,84), *cold as ice* (coll.strength = 93,26), *good as gold* (coll.strength = 88,71). Il semble donc que la conjonction d'une force collostructionnelle élevée (ici mesurée par le logarithme décimal du rapport de log-vraisemblance) et d'une valeur faible de ΔP soit un indice du figement de <ADJ as GN>. A contrario, les constructions du type <ADJ as hell>, dont nous avons observé la productivité et donc le faible degré de figement plus haut, présentent soit des valeurs extrêmes de ΔP et des valeurs de force collostructionnelle faibles (Figure 8, diagramme de droite), soit des valeurs de ΔP proches de 0 et des valeurs de force collostructionnelle élevées (Figure 9, *smart as hell*, *hard as hell* et *cold as hell*).

Le croisement de ces deux mesures permet de voir que le figement ou la productivité d'une construction est une affaire de degré. Empiriquement, il n'y a pas lieu de formuler un jugement binaire consistant à dire qu'une construction est productive ou ne l'est pas.

4. Discussion

Si l'on ne se fie qu'aux deux critères introspectifs et réductionnistes de Kay, la construction <ADJ as GN> est effectivement non-générale et non-productive. En effet, il ne suffit pas de relier un adjectif et un GN par la préposition *as* pour obtenir une expression sanctionnée par l'usage du type *easy as pie* ou *easy as duck soup* ayant une fonction d'intensification de l'adjectif. De plus, nos résultats confirment une autre observation de Kay (2013, p. 39) : certaines expressions sont motivées par le sens du GN (*tall as a tree*, *cold as ice*, *smooth as silk*), d'autres moins car reposant sur un jeu de mots (*safe as houses*, *smart as a whip*) ou une association sémantique figurée (*plain as the nose on someone's face*, *skinny as a rail*), et d'autres pas du tout (*sure as shootin'*, *funny as hell*). Par ailleurs, il semble bien que <ADJ as GN> soit idiosyncrasique à plusieurs égards, comme le prouve la spécialisation contextuelle d'expressions telles que *cold as ice*, qui qualifiera sans problème un sportif en manque de réussite mais pas des conditions météorologiques (on préférera *cold as hell*). Inversement, *hot as hell* qualifiera naturellement une journée caniculaire, mais plus difficilement un sportif en réussite. C'est un défaut de notre approche empirique de ne pas avoir pris en compte systématiquement la variation contextuelle (nous y revenons plus bas). Enfin, certaines constructions se prêtent à une variante comparative (*flat as a pancake* > *flatter than a pancake*) tandis que d'autres non (*happy as a lark* > **happier than a lark*).

Cependant, si l'on se fie à l'empirie, la productivité de <ADJ as GN> est plus complexe que Kay ne le laisse entendre. D'un côté, nous avons des expressions motivées par l'adjectif (Figures 7 et 8, diagrammes de gauche). Sur la base de ces adjectifs (*old*, *black*, *white*, etc.), il est possible de former des constructions intensives non attestées dans le corpus sur le modèle de <ADJ as GN> : *old as a dinosaur* « vieux comme un dinosaure », *black as charcoal* « noir comme le charbon », *white as ivory* « blanc comme l'ivoire », etc. De l'autre, nous avons des expressions motivées par le GN (Figures 7 et 8, diagrammes de droite). Sur la base de ces GN (*glass*, *hell*, etc.), il est possible de créer des constructions intensives telles que *clear/smooth as glass* ou *weird/happy as hell* avec un haut degré de confiance quant à leur acceptabilité, même si ces constructions n'apparaissent pas dans le corpus. Entre ces deux pôles, nous trouvons des constructions moins productives, donc plus figées, au contenu sémantique majoritairement figuré.

Notre méthodologie n'est pas exempte de défauts, le principal étant d'avoir utilisé deux méthodes multifactorielles de manière monofactorielle. Notre AFC n'examine la distribution des adjectifs et des GN qu'au regard d'un seul type de variable : le contexte. Dans la CAH, nous n'avons étudié le profil des adjectifs et des GN qu'au regard de leurs collocations respectives. Dans un travail futur, il serait intéressant d'inclure des variables autres que contextuelles, comme par exemple le GN sur lequel porte l'adjectif intensifié.³³

En dépit de ces limites, nous pensons avoir démontré par notre approche que l'on pouvait s'affranchir de postulats théoriques pour reconnaître que la distinction entre constructions (générales et

³³ Par exemple, l'expression *silent as a tomb* qualifie-t-elle de préférence les entités animées ou inanimées ?

productives) et schémas dérivés n'était pas hermétique et qu'elle ne pouvait pas se démontrer uniquement par des jugements de types introspectifs.

5. Conclusion

Nous avons souhaité accomplir deux objectifs : faire le point sur le statut de la fréquence dans les grammaires de constructions et proposer une étude de cas destinée à montrer que le traitement intuitif des phénomènes de fréquences au détriment de l'empirie menait à une vision faussée de l'usage.

Nous avons présenté plusieurs manières de traiter la fréquence dans une optique quantitative. L'ACC, la CAH et l'AFC (qu'on nous pardonne cet étalage d'acronymes) sont représentatives de ce sur quoi la linguistique cognitive de deuxième génération s'appuie à présent. ΔP est plus avant-gardiste. Sa remise en cause de la symétrie dans les attractions entre lexèmes est représentative du souci de proposer des mesures psycholinguistiquement plausibles.

Le danger serait de croire que les méthodes quantitatives appliquées à la linguistique de corpus sont le seul moyen valable d'accéder à l'usage. Fonder la linguistique sur l'usage suppose que l'on soit sensible à l'intuition des locuteurs natifs.

C'est un défaut de la linguistique cognitive de première génération d'avoir mis la fréquence au cœur de la grammaire sans vouloir la mesurer d'aucune manière, ignorant ainsi le fait que les locuteurs s'appuient sur une intuition statistique (Goldberg, 2011). Ce serait un défaut de la linguistique cognitive de deuxième génération, plus quantitative, d'ignorer le fait que la plupart des intuitions fondatrices de la théorie sont justes.

Pour que la linguistique soit véritablement fondée sur l'usage, il faut procéder à un aller-retour fécond entre analyse intuitive et méthodes quantitatives en ne perdant pas de vue que l'intuition qui nous intéresse est en définitive celle du locuteur.

Bibliographie

- Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, 15(3), 147-149.
- Barlow, M., & Kemmer, S. (2000). *Usage-based models of language*. Stanford, Calif.: CSLI Publications.
- Benzécri, J.-P. (1984). *Analyse des correspondances, exposé élémentaire* (Vol. 1). Paris: Dunod.
- Biber, D. (1998). *Corpus linguistics : investigating language structure and use*. Cambridge: Cambridge University Press.
- Bybee, J. L. (1985). *Morphology : a study of the relation between meaning and form*. Amsterdam ; Philadelphia: J. Benjamins.
- Bybee, J. L. (2006). From Usage to Grammar: The Mind's Response to Repetition. *Language*, 82(4), 711-733.
- Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge ; New York: Cambridge University Press.
- Bybee, J. L., & Hopper, P. (2001). *Frequency and the emergence of linguistic structure*. Amsterdam ; Philadelphia, PA: John Benjamins Pub. Co.
- Chomsky, N. (1957). *Syntactic structures*. 's-Gravenhage: Mouton en Co.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, Mass: M.I.T. Press.
- Cornillon, P.-A., & Matzner-Løber, É. (2010). *Régression avec R*. Paris Berlin Heidelberg [etc.]: Springer.
- Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 450 million words, 1990-present. Available online at <http://corpus.byu.edu/coca/>
- Desagulier, G. (sous presse). Visualizing distances in a set of near synonyms: rather, quite, fairly, and pretty. In D. Glynn & J. Robinson (Eds.), *Polysemy and Synonymy : Corpus Methods and Applications in Cognitive Linguistics*. Amsterdam: John Benjamins.
- Desagulier, G. (soumis). *Quite* new methods for a *rather* old issue: exploring and visualizing collocation data from the BNC with correspondence analysis.
- Dirven, R., Goosens, L., Putseys, Y., & Vorlat, E. (1982). *The scene of linguistic action and its perspectivization by speak, talk, say and tell*. Amsterdam: John Benjamins.

- Dirven, R., & Taylor, J. R. (1988). The conceptualisation of vertical space in English: the case of *tall*. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics*. Amsterdam ; Philadelphia: John Benjamins.
- Divjak, D., & Fieller, N. (to appear). Finding structure in linguistic data. In D. Glynn & J. Robinson (Eds.), *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics*. Amsterdam: John Benjamins.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.
- Ellis, N. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, 27(1), 1-24.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (Vol. 5). Oxford: Wiley-Blackwell.
- Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*. Unpublished doctoral dissertation, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. Retrieved from <http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/pdf/Evert2005phd.pdf>
- Fillmore, C. (1997). *Construction Grammar lecture notes*.
- Fillmore, C., Kay, P., & O'Connor, C. (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of *let alone*. *Language*, 64(3), 501-538.
- Firth, J. (1957). A Synopsis of Linguistic Theory, 1930-1955. In F. Palmer (Ed.), *Selected Papers of J.R. Firth 1952-1959* (pp. 168-205). London: Longman.
- Geeraerts, D. (2000). Saliency phenomena in the lexicon. A typology. In L. Albertazzi (Ed.), *Meaning and Cognition* (pp. 125-136). Amsterdam, Philadelphia: John Benjamins.
- Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford ; New York: Oxford University Press.
- Geeraerts, D., Grondelaers, S., & Bakema, P. (1994). *The structure of lexical variation : meaning, naming, and context*. Berlin ; New York: M. de Gruyter.
- Glynn, D. (2010a). Corpus-Driven Cognitive Semantics. An introduction to the field. In D. Glynn & K. Fischer (Eds.), *Corpus-Driven Cognitive Semantics. Quantitative approaches*. (pp. 1-42). Berlin: Mouton de Gruyter.
- Glynn, D. (2010b). Testing the hypothesis: Objectivity and verification in usage-based cognitive semantics. In D. Glynn & K. Fischer (Eds.), *Corpus-Driven Cognitive Semantics. Quantitative Approaches* (pp. 239-270). Berlin: Mouton de Gruyter.
- Glynn, D., & Robinson, J. (Eds.). (à paraître). *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics*. Amsterdam: John Benjamins.
- Goes, J. (1999). *L'adjectif : entre nom et verbe*. Paris: Duculot.
- Goldberg, A. E. (1995). *Constructions : a construction grammar approach to argument structure*. Chicago: University of Chicago Press.
- Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in cognitive sciences*, 7(5), 219-224.
- Goldberg, A. E. (2006). *Constructions at work : the nature of generalization in language*. Oxford ; New York: Oxford University Press.
- Goldberg, A. E. (2009). The nature of generalization in language. *Cognitive Linguistics*, 20(1), 93-127.
- Goldberg, A. E. (2011). Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics*, 22(1), 131-153.
- Greenacre, M. J. (2007). *Correspondence analysis in practice* (Vol. 2). Boca Raton: Chapman & Hall/CRC.
- Gries, S. T. (2003). *Multifactorial analysis in corpus linguistics: a study of particle placement*. New York: Continuum.
- Gries, S. T. (2007). Coll.analysis 3.2. A program for R for Windows 2.x.
- Gries, S. T. (2013). 50-something years of work on collocations: what is or should be next... *International Journal of Corpus Linguistics*, 18(1).
- Gries, S. T., & Stefanowitsch, A. (2004a). Co-varying collexemes in the *into*-causative. In M. Achard & S. Kemmer (Eds.), *Language, Culture, and Mind* (pp. 225-236). Stanford, Calif.: CSLI.
- Gries, S. T., & Stefanowitsch, A. (2004b). Extending collocation analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9(1), 97-129.

- Gries, S. T., & Stefanowitsch, A. (2006). *Corpora in cognitive linguistics : corpus-based approaches to syntax and lexis*. Berlin: Mouton de Gruyter.
- Gries, S. T., & Stefanowitsch, A. (2010). Cluster analysis and the identification of collexeme classes. In J. Newman & S. Rice (Eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford, CA: CSLI.
- Husson, F., Le, S., & Pages, J. (2011). *Exploratory multivariate analysis by example using R*. Boca Raton: CRC Press.
- Kay, P. (2013). The Limits of (Construction) Grammar. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.
- Kay, P., & Fillmore, C. (1999). Grammatical constructions and linguistic generalizations: the What's X doing Y? construction. *Language*, 75, 1-33.
- Kleiber, G. (2007). Adjectifs de couleur et gradation : une énigme... « très » colorée. *Travaux de linguistique*, 55(2), 9-44.
- Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1, 127-165.
- Lafon, P. (1981). Analyse lexicométrique et recherche des cooccurrences. *Mots*, 95-148.
- Lafon, P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève, Paris: Slatkine, Champion.
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things*: University Of Chicago Press.
- Lakoff, G. (1990). The Invariance Hypothesis: is abstract reason based on image-schemas? *Cognitive Linguistics*, 1(1), 39-74.
- Langacker, R. W. (1982). Space Grammar, analysability, and the English passive. *Language*, 58(1), 22-80.
- Langacker, R. W. (1986). An Introduction to Cognitive Grammar. *Cognitive Science*, 10(1), 1-40.
- Langacker, R. W. (1987). *Foundations of cognitive grammar* (Vol. 1). Stanford, CA: Stanford University Press.
- Langacker, R. W. (1991). *Foundations of cognitive grammar* (Vol. 2): Stanford.
- Langacker, R. W. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmer (Eds.), *Usage-based Models of Language* (pp. 1-64). Stanford: CSLI Publications.
- Langacker, R. W. (2008). *Cognitive grammar : a basic introduction*. Oxford: Oxford University Press.
- Langacker, R. W. (2009). Cognitive (Construction) Grammar. *Cognitive Linguistics*, 20(1), 167-176.
- Lebart, L., & Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.
- Muller, C. (1964). *Essai de statistique lexicale. L'illusion Comique de Pierre Corneille*. Paris: Klincksieck.
- Muller, C. (1973). *Initiation aux méthodes de la statistique linguistique*. Paris: Champion.
- Muller, C. (1977). *Principes et méthodes de statistique lexicale*. Paris: Hachette.
- Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1), 137-158.
- R Core Team (2013). R: a language and environment for statistical computing (3.0.1). R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Schmid, H.-J. (2007). Entrenchment, salience, and basic levels. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford handbook of Cognitive Linguistics* (pp. 117-138). Oxford: Oxford University Press.
- Schmid, H.-J. (2010). Does frequency in text really instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: corpus-driven approaches* (pp. 101-133). Berlin, New York: Mouton de Gruyter.
- Sinclair, J. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In Memory of J.R. Firth* (pp. 410-431). Longman: Longman.
- Sinclair, J. (1987). Collocation: A progress report. In R. Steele & T. Threadgold (Eds.), *Language Topics: Essays in Honour of Michael Halliday* (Vol. 2, pp. 319-331). Amsterdam: John Benjamins.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J. (1996). The search for units of meaning. *Textus*, 9, 75-106.
- Sinclair, J., & Carter, R. (2004). *Trust the text : language, corpus and discourse*. London: Routledge.
- Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209-243.

- Stefanowitsch, A., & Gries, S. T. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, 1(1), 1-46.
- Stubbs, M. (2001). *Words and phrases : corpus studies of lexical semantics*. Oxford: Blackwell.
- Tummers, J., Heylen, K., & Geeraerts, D. (2005). Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory*, 1(2), 225-261.
- Van de Velde, D. (1995). *Le spectre nominal*. Leuven: Peeters.
- Waldron, H. A. (1983). Did the Mad Hatter have mercury poisoning? *British Medical Journal*, 287, 1961.
- Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301), 236.
- Whittaker, S. (2002). *La notion de gradation. Application aux adjectifs* (Vol. 237). Bern, Berlin, Bruxelles, Frankfurt/M., New York, Oxford, Wien: Peter Lang.
- Wierzbicka, A. (1988). *The Semantics of Grammar*. Amsterdam ; Philadelphia: John Benjamins.
- Zipf, G. (1935). *The Psychobiology of Language: An Introduction to Dynamic Philology*: Routledge.