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

Guillaume Desagulier

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 Fillmoréenne (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 génerativisme 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 (…) » (2013, p. 38). Mais contrairement à ce que postule la Grammaire Cognitive, la productivité n'est pas indexée sur la fréquence dans la GxCF.

L'approche de Kay est radicale sur le plan méthodologique. 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.

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 dans les approches centrées sur l'usage. La troisième section est consacrée à l'étude de <ADJ *as* GN> en corpus. Nous présentons et utilisons trois outils : l'analyse collexémique covariante, la classification ascendante hiérarchique 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 de l'usage

2.1. Fréquence perçue vs. fréquence mesurée

La tradition mentaliste héritée de *Syntactic Structures* (Chomsky, 1957) et *Aspects of the Theory of Syntax* (Chomsky, 1965) 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). Par contraste, les approches centrées sur l'usage dans le sillage de Langacker (1982, 1987, 1991) 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. 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 a priori 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 changement de point de vue 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, 2010a, 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'usage[3], 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 » (Langacker, 1987, p. 59). Pourtant, Langacker admet bien plus tard que cette frontière peut être identifiée de manière quantitative en prenant en compte la fréquence observée : « (…) in principle the degree of entrenchment can be determined empirically. Observed frequency provides one basis for estimating it » (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 (…), or whether *computer* is in fact more analyzable than *propeller*. (Langacker, 2008, p. 86)

Ce revirement intervient parallèlement à l'essor d'études quantitatives (sur corpus) et expérimentales en linguistique cognitive.[4] Toutefois, fidèle aux principes de l'approche centrée sur l'usage, Langacker ne sépare pas les méthodes quantitatives de l'intuition du locuteur (« speaker judgments»).

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 proéminence cognitive. Sur le plan linguistique, la fréquence et la saillance peuvent être à la fois en résonnance 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 proéminente. Cette corrélation directe entre fréquence et saillance relève de la résonnance. 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 rapide 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 systématise le couplage entre linguistique de corpus et méthodes quantitatives en linguistique cognitive.

---

[3] 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 & Kemmer (2000) et Langacker (2000).

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

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),[5] 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 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 schématisme et de convention (des lexèmes aux schémas les plus abstraits). Au même titre que d'autres méthodes quantitatives destinées à cerner les collocations, y compris dans leur dimension phraséologique, l'analyse collostructionnelle exploite une idée bien connue en linguistique quantitative : 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 collostructionnelle 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 les lexèmes apparaissant dans cette même construction (ex. les substantifs spécifiques du quantifieur *quelques*). L'analyse collexémique distinctive mesure la préférence de lexèmes 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*[6]). Enfin, l'analyse collexémique covariante mesure le degré d'attraction ou de répulsion de lexèmes dans différentes positions d'une même construction (ex. <V-inf *plus pour* V-inf *plus*>, <ADJ *comme* GN>).

L'analyse collostructionnelle s'appuie sur des mesures d'association aptes à mesurer l'attraction ou la répulsion entre deux lexèmes.[7] 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 collostructionnelle

La plupart des mesures d'association se fondent sur un tableau d'entrée tel que le Tableau 2 pour générer une *p*-valeur. 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	$\neg L_1$	total (lignes)
L_2	a	b	a+b
$\neg L_2$	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 L_1 et L_2 (¬ : « autre que »)

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

[6] Cf. Gréa (2008), Gréa & Haas (ce volume).

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

La case L_2/L_1 indique le nombre de fois où l'on trouve les deux lexèmes ensemble, $L_2/\neg L_1$ le nombre de fois où l'on trouve le lexème L_2 sans L_1, $\neg L_2/L_1$ le nombre de fois où l'on trouve le lexème L_1 sans L_2 et $\neg L_2/\neg L_1$ le nombre de fois où l'on ne trouve ni L_1 ni L_2. Le principe est le même en analyse collostructionnelle, 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 collostructionnelle » pour permettre un classement des éléments les plus attirés vers les moins attirés.[8] 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 C_1 et C_2

	$L_{créneau\_1}$	$\neg L_{créneau\_1}$	total (lignes)
$L_{créneau\_2}$	a	b	a+b
$\neg L_{créneau\_2}$	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.[9] 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 (la co-occurrence de mots) et colligation (la co-occurrence de formes lexicales et de phénomènes grammaticaux). 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

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

[8] 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.

[9] L'accueil de l'analyse collostructionnelle en France est mitigé pour une raison principalement historique. Au début des années 2000, l'école française de lexicométrie est déjà bien implantée (Lafon, 1980, 1981, 1984; Lebart & Salem, 1994; Muller, 1964, 1973, 1977). On reproche notamment à l'analyse collostructionnelle de réinventer le calcul des spécificités de Lafon, ou de faire doublon avec les études sur la colligation.

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 deux 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 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'analyse collostructionnelle ne rend pas compte de l'asymétrie dans l'attraction entre deux éléments (lexèmes ou constructions). Or, dans un appariement de lexèmes ou de constructions, l'attraction est rarement symétrique (ex. dans *ad hominem*, *hominem* attire plus *ad* que vice versa). 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'analyse collexémique covariante.

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 millions de mots répartis en près de 190 000 textes d'anglais américain compilés entre 1990 et 2012. Le COCA est un corpus dit « équilibré » au sens ou 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. Sa taille et son échantillonnage permettent 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 la classification ascendante hiérarchique 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 montrer 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 dans l'ensemble des constructions <ADJ *as* GN>.

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

Tableau 6. Les dix types les plus fréquents de <ADJ *as* GN>.[10]

<ADJ *as* GN> 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. Selon Leroy (2004), ce type de structure opère une comparaison dite « à parangon », le GN jouant le rôle de parangon[11]. La majorité des adjectifs du tableau 6 sont prototypiquement gradables.[12] 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). *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 d'un drap a valeur de paragon par convention.

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)
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 dilution sémantique, le sens lexical littéral faisant place à un sens lexical grammatical :

(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

[10] Les fréquences de l'adjectif et du GN sont rapportées à la construction A as GN.

[11] Sur le rôle de *comme*, très proche de celui de *as*, voir également Fournier & Fuchs (2007).

[12] La gradabilité étant même une propriété définitoire des adjectifs selon Goes (1999).

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.

Indépendamment de l'allitération, et pour n'importe quelle expression, le contexte peut conduire à remotiver certains éléments comme en (9). Dans cet exemple de défigement, l'or de *good as gold* (« infaillible/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 métonymique du type *white as a sheet* (la blancheur est une qualité saillante du drap blanc), voire *clear as a bell* (la clarté est une qualité saillante 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, dans tous les cas, le lien sémantique entre les unités qui composent cette construction est conventionnel et non intrinsèque ou objectif (un drap n'est pas intrinsèquement blanc, pas plus que le son d'une cloche intrinsèquement clair). Ce lien est établi au-delà du niveau compositionnel par la construction. Par conséquent, nous sommes bien dans une problématique constructionnelle au sens où l'entendent les grammaires 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 que 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). Nous obtenons le Tableau 7.[13] Celui-ci indique, pour chaque paire ADJ-GN, la fréquence totale de l'adjectif dans la construction, la fréquence totale du GN dans la construction, 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).

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

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

À première vue, la différence entre les fréquences brutes (Tableau 6) et l'ACC (Tableau 7) est infime. On retrouve les mêmes paires ADJ-GN parmi les dix les plus associées. Seul l'ordre du classement change. La 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 (trait plein). 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*) et par la forme de la courbe Lowess (en pointillés).[14] Ceci est un effet du calcul de la force collostructionnelle, qui fait intervenir une transformation logarithmique.

[14] Par rapport à une droite de regression 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 (Cornillon & Matzner-Løber, 2010).

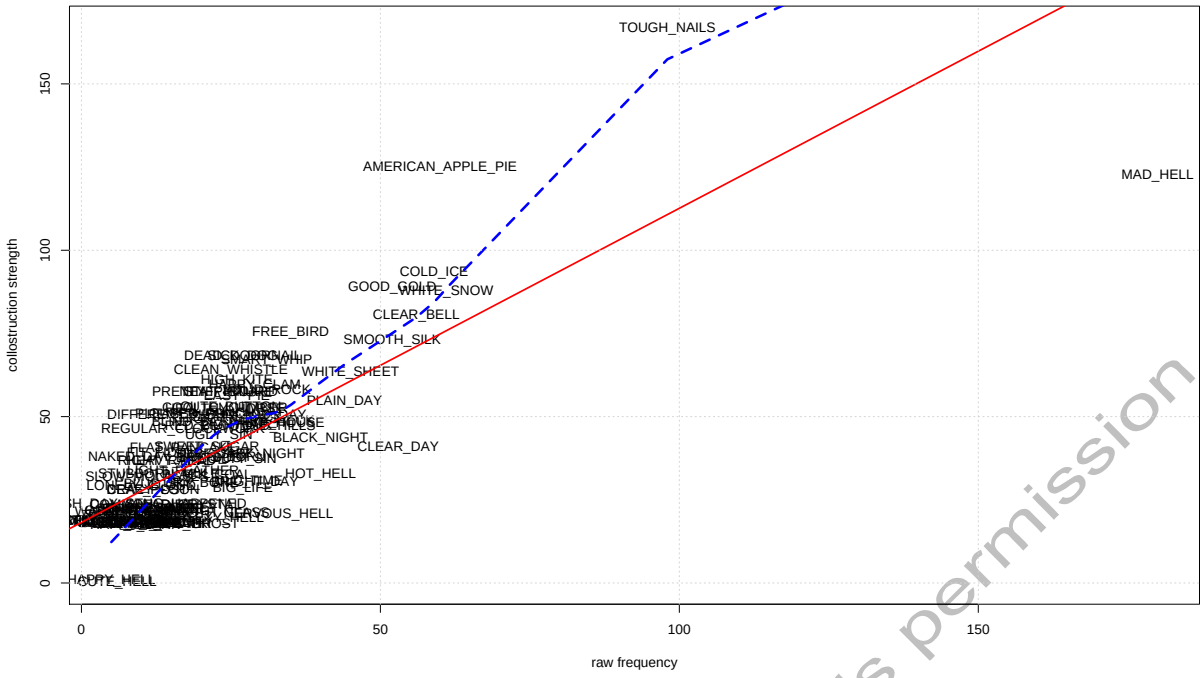


Figure 1. Croisement entre fréquence brute et force collostructionnelle (en trait plein : droite de régression ; en pointillés : courbe Lowess)

On note que *mad as hell*, qui figurait en première position au Tableau 6, ne figure qu'en troisième position au Tableau 7. Ceci est dû au fait que les fréquences respectives de *mad* et de *hell* sont élevées en dehors de la construction. La fréquence attendue de leur collocation dans la construction est donc très proche de la fréquence observée. Au vu de leurs fréquences absolues respectives, il est moins surprenant de voir appariés *mad* et *hell* dans la construction <ADJ *as* GN> que *tough* et *nails* ou *American* et *apple pie*. Ceci dit, comparer des différences de classement aussi infimes n'a pas vraiment de sens linguistiquement parlant. On pourrait donc penser que l'apport de l'ACC vis-à-vis des fréquences brutes est minime, mais ce serait oublier que pour un autre jeu de données la différence entre les deux pourrait être plus importante. L'ACC a le mérite de rappeler que les fréquences brutes ne montrent pas en quoi une association lexicale est surprenante.

On relève deux types de liens sémantiques entre l'adjectif et le GN : un lien assez strict de nature métonymique et un lien plus lâche fondé sur la connotation du GN. Le premier type concerne les paires suivantes : *tough as nails*, *cold as ice*, *white as snow*, *smooth as silk*, *white as a sheet*, *high as a kite*, *solid as a rock*, et *stiff as board*. À chaque fois, l'adjectif dénote une propriété distinctive du GN (la dureté des ongles, la solidité du rocher, la raideur de la planche, etc.) Le second type concerne les paires suivantes : *American as apple pie*, *mad as hell*, *good as gold*, *clear as a bell*, *free as a bird*, *dead as a doornail*, *sick as a dog*, *smart as a whip*, *clean as a whistle*, *happy as a clam*, *neat as a pin* et *pretty as a picture*. Comme nous l'avons vu plus haut, le lien sémantique est plus imagé (ex. *free as a bird* lit. « libre comme un oiseau ») et parfois fantasque (ex. *happy as a clam* lit. « heureux comme une palourde »)[15].

Si les résultats de l'ACC fournissent des associations plus sûres, car prenant en compte des fréquences relatives, le Tableau 7 ne permet cependant pas de déterminer facilement les profils sémantiques de la construction <ADJ *as* GN>. Si le problème se pose avec un tableau de 20 lignes, il se pose avec encore plus d'acuité pour le tableau complet qui en comporte 265.

Nous proposons d'explorer les données de sortie de l'ACC avec la classification hiérarchique ascendante. C'est une méthode multifactorielle exploratoire. Elle permet d'observer des tendances à travers des tableaux de données de grande taille. L'observateur ne formule aucune hypothèse quant aux tendances sous-jacentes au tableau de données, même si en linguistique la constitution d'un tableau de données suppose que l'on a une raison de croiser les données que l'on fait intervenir.

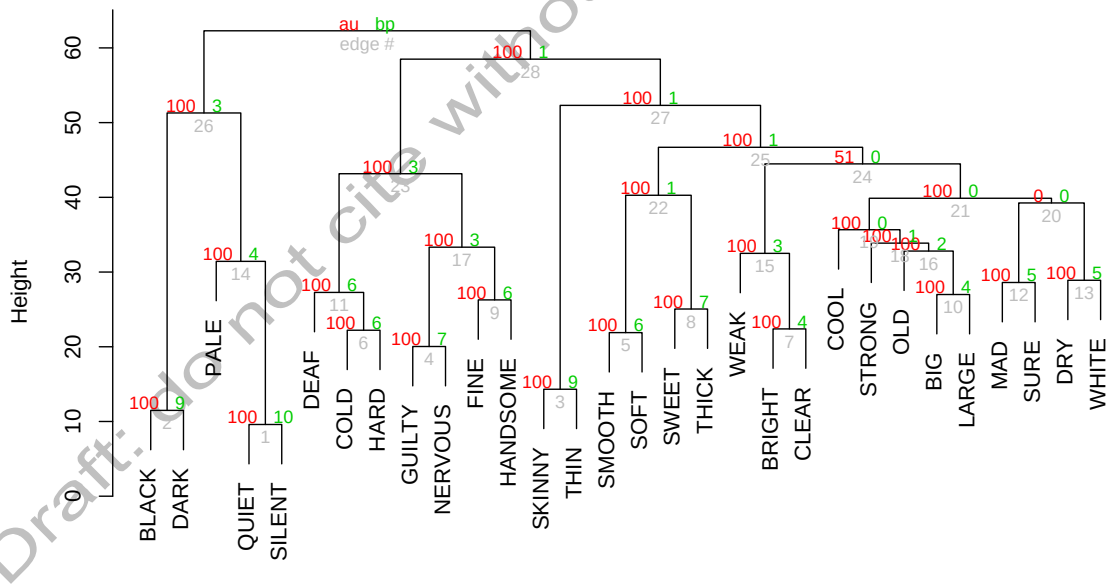
[15] Au-delà du Tableau 7, on trouve également *naked as a jaybird* « nu comme un geai ».

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 collostructionnelle 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, et al., 2011, section 4.2 ; Divjak & Fieller, 2014 ; Desagulier, 2014)[16]. Les résultats sont représentés sous la forme d'un dendrogramme qui se lit du bas vers le haut.[17] Les mots d'autant plus proches qu'ils partagent une distribution similaire s'amalgament en clusters en premier. Idéalement, les clusters les mieux structurés regroupent les mots par unités de sens.

A l'aide du même tableau de contingence, nous avons généré deux dendrogrammes.[18] Le premier dendrogramme (Figure 2) représente la classification des 30 adjectifs en fonction des 34 GN avec lesquels ils se combinent. 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). Lorsque 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 ».



[16] Voir également Beliaõ, Lacheret et Kahane (ici-même).

[17] Chaque cluster est accompagné de trois nombres. Celui du bas indique le rang du cluster (du premier cluster généré par la classification au 28e). 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 ».

[18] Nous avons utilisé pour cela le package *pvclust* pour R.

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

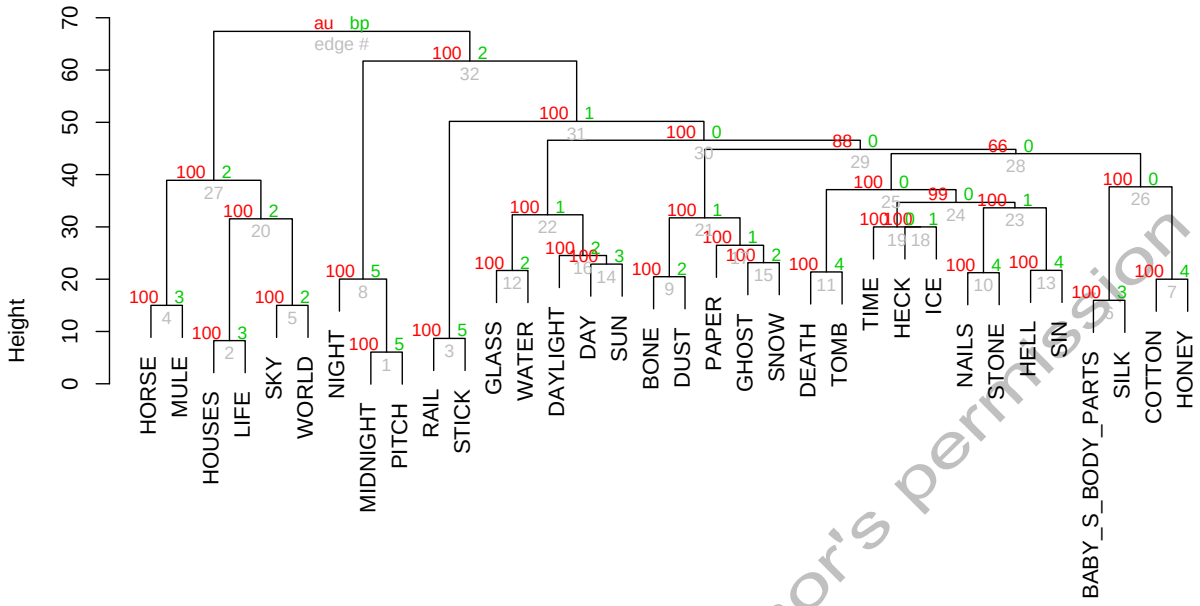


Figure 3. Classification ascendante hiérarchique de 34 GN classés en fonction des 30 adjectifs avec lesquels ils se combinent dans la construction.

Nos classifications ont deux défauts. Premièrement, nous n'avons pris en compte que 30 des 121 adjectifs et 34 des 171 GN. Deuxièmement, en séparant adjectifs (Figure 2) et GN (Figure 3), nous ne pouvons pas corréler leurs profils sémantiques respectifs. 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. 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*(ADJ|GN), à savoir la probabilité de tel adjectif sachant que l'on a tel GN, et *p*(GN|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 8) :

	O	¬O
C	a	b
¬C	c	d

Tableau 8. 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, etc.). ΔP se calcule de la manière suivante :

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

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|¬C). Lorsque ces probabilités sont identiques (…) il n'y a pas de covariation entre les deux événements et ΔP = 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 collostructions. Dans le cadre de la construction <ADJ *as* GN>, nous devons calculer deux valeurs de ΔP :

(ii) $\Delta P_{GN|ADJ} = p(GN|ADJ) - p(GN|\neg ADJ)$
$= a/(a + b) - c/(c + d)$

(iii) $\Delta P_{ADJ|GN} = p(ADJ|GN) - p(ADJ|\neg GN)$
$= a/(a + c) - c/(b + d)$

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 »)[19] et appliquons les formules (ii) et (iii) au Tableau 9 en (iv) et (v) respectivement. Il apparaît que *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.

	hatter : présent	*hatter* : absent	totaux
mad : présent	14	196	210
mad : absent	0	464 020 046	464 020 046
totaux	14	464 020 242	464 020 256

Tableau 9. Tableau de contingence pour *mad as a hatter* dans le COCA.

(iv) $\Delta P_{hatter|mad} = p(hatter|mad) - p(hatter|\text{adjectifs autres que } mad)$
$= 14/(14 + 196) - 0/(0 + 463\,020\,046)$
$\approx 0{,}067$

(v) $\Delta P_{mad|hatter} = p(mad|hatter) - p(mad|\text{GN autres que } hatter)$
$= 14/(14 + 0) - 0/(196 + 463\,020\,046)$
$= 1$

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

[19] En référence au chapelier fou dans *Alice au pays des merveilles* de Lewis Caroll. 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).

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 collostructionnelle 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 collostructions sont asymétriques. En effet, sur les 265 types que compte la construction <ADJ *as* GN>, 132 affichent une différence importante en termes de valeurs de ΔP (≥ 0,5 ou ≤ –0,5). Seuls 24 affichent une différence 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%).

La Figure 4 donne une image détaillée des asymétries. Chaque cercle représente une paire ADJ-GN.

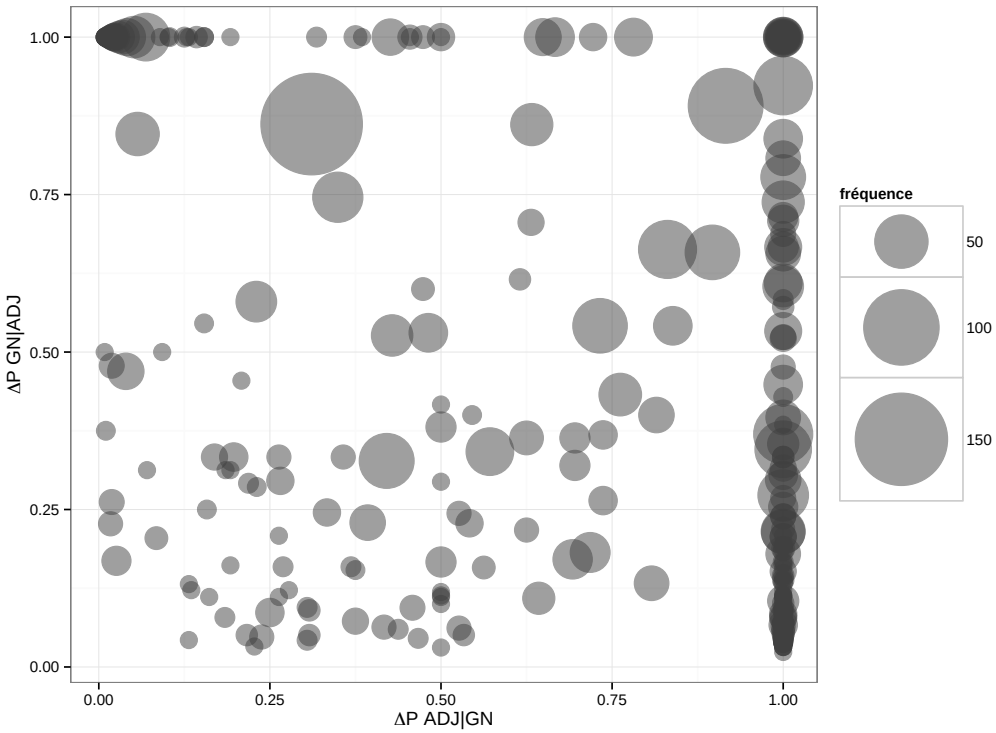


Figure 4. Croisement des valeurs de $\Delta P_{GN|ADJ}$ et $\Delta P_{ADJ|GN}$

Pour un grand nombre de paires ADJ-GN, l'asymétrie est importante, à en juger par la concentration importante de cercles 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[20] et des différences par paires des valeurs de ΔP. Cette distribution est visible en Figure 5.

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 productivité. 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 :

[20] Le rapport de log-vraisemblance (Dunning, 1993), qui est une mesure très proche du test exact de Fisher.

old *old as war/warfare, old as the Bible, old as history, old as politics, old as the century, old as man/mankind/humanity, old as civilization* ;
black *black as ebony, black as soot, black as tar, black as ink, black as obsidian* ;
white *white as ghost, white as snow, white as sheet, white as chalk, white as milk* ;
big *big as life, big as saucers, big as mountains/a mountain, big as quarters, big as softballs, big as a barn, big as the Earth, big as the universe* ;
strong *strong as a bull, strong as steel, strong as iron, strong as pillars* ;
sharp *sharp as razors/a razor, sharp as a knife, sharp as a tack.*

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*)[21], 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*.

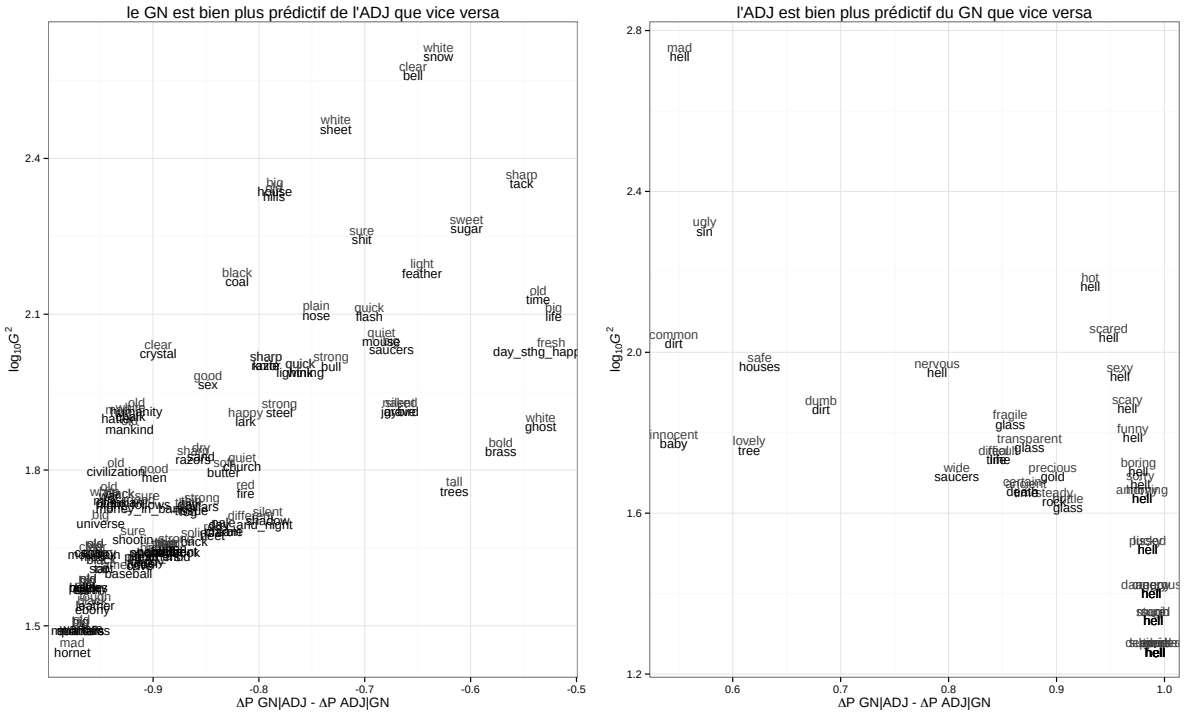


Figure 5. Distribution des paires ADJ-GN selon le rapport de log-vraisemblance (en ordonnée) et les moitiés supérieures des différences $\Delta P_{GN|ADJ} - \Delta P_{ADJ|GN}$ (en abscisse) (≤ –0,5 à gauche et ≥ 0,5 à droite).

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* s'est grammaticalisé et n'est employé que pour sa fonction intensive. C'est de loin le GN le plus productif de la construction. Il s'emploie dans la construction <ADJ *as* GN> pour intensifier des adjectif dénotant des états psychologiques et

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

physiologiques (*glad*, *surprised*, *angry*, *pissed*, *horny*, *sorry*, *scared*, *nervous*, *bored*, *depressed*, *tired*, *stupid*, *lucky*), divers stimuli (*annoying*, *boring*, *creepy*, *dangerous*, *eerie*, *funny*, *hot*[22], *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	*bored*, *boring*
colère	*angry*, *pissed*
étrangeté	*eerie*, *creepy*
séduction, excitation	*sexy*, *hot*, *horny*

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 6).

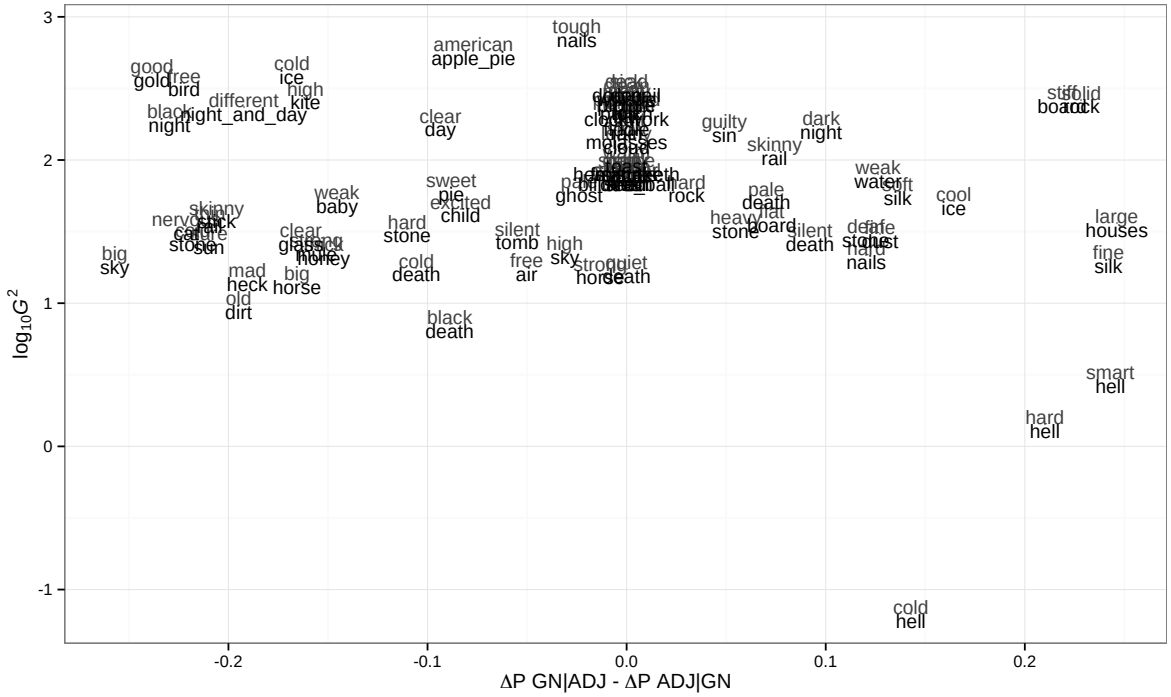


Figure 6. Distribution des paires ADJ-GN selon le rapport de log-vraisemblance (en abscisse) et les deux extrêmes des différences $\Delta P_{GN|ADJ} - \Delta P_{ADJ|GN}$ (en ordonnée, ≤ –0,25 à gauche et ≥ 0,25 à droite).

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 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 5, diagramme de droite), soit des valeurs de ΔP proches de 0 et des valeurs de force collostructionnelle élevées (Figure 6, *smart as hell*, *hard as hell* et *cold as hell*).

[22] Au sens de « sexy ».

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 préfèrera *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 instanciées au niveau de l'adjectif et productives au niveau du GN (Figures 5, diagramme 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 instanciées aux niveau du GN et productives au niveau de l'adjectif (Figures 5, diagramme 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.

On pourrait à juste titre objecter que les mesures d'association, qu'elles soient directionnelles ou asymétriques, ne rendent compte que partiellement et indirectement de la productivité constructionnelle. Premièrement, il y a fort à parier que la productivité d'une construction dépend d'une multiplicité de facteurs, à la fois linguistiques et paralinguistiques. On peut notamment supposer que même la construction la plus schématique ne générera pas le même nombre d'exemples en fonction du contexte ou du genre textuel ou du mode (écrit ou oral). À titre d'illustration, *ugly as sin* « hideux » ou *dark as night* « très sombre » ne sont productifs qu'à l'écrit. *Sure as shit*, équivalent vulgaire de « très probablement », n'est productif que dans un seul genre de l'écrit : la fiction.

Deuxièmement, nous n'avons pas ici, par manque de place, comparé la performance de nos mesures à des méthodes spécifiquement destinées à quantifier la productivité. Issues des travaux de Baayen en morphologie, ces dernières sont résumées dans Baayen (2009). Les plus notoires s'appuient sur l'hypothèse émise par Baayen et Renouf (1996) selon laquelle le nombre d'hapax legomena (i.e. d'occurrences de fréquence 1) pour une catégorie morphologique donnée (ex. un affixe) donne une indication fiable quant à la capacité de cette catégorie à former des néologismes. Exploitant ce postulat, Baayen a mis au point plusieurs mesures pour quantifier la productivité. $\mathscr{P}^*$ est le taux effectif de création de néologismes dans un corpus. On l'obtient en divisant V(1, C, N), à savoir le nombre de types de fréquence 1 de la catégorie C dans un corpus de N mots, par V(1, N), à savoir le nombre de types de fréquence 1 dans un corpus de N mots. $\mathscr{P}$ mesure la productivité non pas effective mais potentielle, c'est-à-dire la probabilité pour un processus morphologique de produire des néologismes. On l'obtient en divisant le nombre d'hapax legomena de la catégorie C par le nombre N d'occurrences de la catégorie dans le corpus. En principe, les valeurs de $\mathscr{P}^*$ et $\mathscr{P}$ sont comprises entre

0 et 1. 0 signifie qu'il n'y a pas d'hapax et 1 que tous les hapax du corpus proviennent de C (pour $\mathscr{P}$*) ou que toutes les C sont des hapax (pour $\mathscr{P}$). *P** mesure la productivité globale d'un processus. On l'obtient par le rapport du nombre de types sur $\mathscr{P}$. Enfin, en lien avec les mesures présentées ci-dessus, Baayen a développé une méthode reposant sur des courbes d'accroissement du vocabulaire (ou VGC pour *vocabulary growth curves*). Ces courbes sont établies à partir de modèles statistiques zipfiens. Elles permettent de modéliser le lien entre la productivité attestée (par interpolation) et la productivité estimée (par extrapolation) d'un processus donné. Les mesures décrites ci-dessus ont été optimalisées pour la morphologie. Zeldes (2012) les a adaptées à des problématiques syntaxiques.

L'influence sur les résultats du contexte et de la dimension générique des textes du corpus ainsi que la comparaison avec des mesures de la productivité en syntaxe sont deux points que nous abordons ailleurs (Desagulier, soumis). Nous pensons néanmoins avoir démontré ici que l'on pouvait s'affranchir de postulats théoriques pour reconnaître que la frontière entre « constructions » (générales et productives) et « schémas dérivés » était poreuse et qu'on ne pouvait pas la cerner uniquement par des jugements introspectifs.

5. Conclusion

Nous avons souhaité remplir 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 et la CAH (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 plausibles car plus en accord avec l'apprentissage mémoriel humain.

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. (eds) (2000), *Usage-based Models of Language*, Stanford: CSLI Publications.

BAAYEN, R. H. (2009), "Corpus linguistics in morphology: morphological productivity", *in* A. Luedeling & M. Kyto (eds.), *Corpus Linguistics. An international handbook*, Berlin: Mouton De Gruyter.

BAAYEN, R. H., & RENOUF A. (1996), "Chronicling the Times: Productive lexical innovations in an English newspaper", *Language* 72(1), 69-96.

BENZECRI, 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: John 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: Cambridge University Press.
BYBEE, J. L., & HOPPER, P. (2001), *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins.
CHOMSKY, N. (1957), *Syntactic Structures*, The Hague: Mouton.
CHOMSKY, N. (1965), *Aspects of the Theory of Syntax*, Cambridge: M.I.T. Press.
CORNILLON, P.-A., & MATZNER-LØBER, É. (2010), *Régression avec R*, Paris, Berlin, Heidelberg: Springer.
DAVIES, M. (2008-), The Corpus of Contemporary American English (COCA): 450 million words, 1990-present. Disponible en ligne à l'adresse suivante : http://corpus.byu.edu/coca/
DESAGULIER, G. (2014), "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), "A lesson from associative learning: what collostructional asymmetries reveal as to the productivity of constructions".
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: John Benjamins.
DIVJAK, D., & FIELLER, N. (2014), « Finding structure in linguistic data », *in* D. Glynn & J. Robinson (éds.), *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*, thèse de doctorat, Institut für maschinelle Sprachverarbeitung, Université de Stuttgart. Disponible à l'adresse suivante : http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/pdf/Evert2005phd.pdf
FILLMORE, C. (1997), *Construction Grammar lecture notes*. Disponible à l'adresse suivante : http://www1.icsi.berkeley.edu/~kay/bcg/lec02.html
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*, London: Longman, 168-205.
FOURNIER, N. & FUCHS C. (2007), « *Que* et *comme* marqueurs de comparaison », *Lexique* 18, 69-107.
GEERAERTS, D. (2000), "Salience phenomena in the lexicon. A typology", *in* L. Albertazzi (ed), *Meaning and Cognition*, Amsterdam: John Benjamins, 125-136.
GEERAERTS, D. (2010), *Theories of Lexical Semantics*. Oxford: Oxford University Press.
GEERAERTS, D., GRONDELAERS, S., & BAKEMA, P. (1994), *The Structure of Lexical Variation : Meaning, Naming, and Context*. Berlin: Mouton 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*, Berlin: Mouton de Gruyter, 1-42.
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*, Berlin: Mouton de Gruyter, 239-270.
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: 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*, Stanford: CSLI, 225-236.
GRIES, S. T., & STEFANOWITSCH, A. (2004b), "Extending collostructional 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: 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* 3, 95-148.
LAFON, P. (1984), *Dépouillements et statistiques en lexicométrie*, Genève, Paris: Slatkine, Champion.
LAKOFF, G. (1987), *Women, Fire, and Dangerous Things*, Chicago: 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: Stanford University Press.
LANGACKER, R. W. (1991), *Foundations of Cognitive Grammar*, vol. 2, Stanford: Stanford University Press.
LANGACKER, R. W. (2000), "A dynamic usage-based model", *in* M. Barlow & S. Kemmer (eds), *Usage-based Models of Language*, Stanford: CSLI Publications, 1-64.
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.
LEROY, S. (2004), « Sale comme un peigne et méchant comme une teigne. Quelques remarques sur les comparaisons à parangon », *Travaux Linguistiques du Cerlico* 17, 255-267.
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. 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*, Oxford: Oxford University Press, 117-138.

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*, Berlin: Mouton de Gruyter, 101-133.

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*, Longman: Longman, 410-431.

SINCLAIR, J. (1987), "Collocation: A progress report", *in* R. Steele & T. Threadgold (eds), *Language Topics: Essays in Honour of Michael Halliday*, vol. 2, Amsterdam: John Benjamins, 319-331.

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), 236-244.

WHITTAKER, S. (2002), *La notion de gradation. Application aux adjectifs*, Bern, Berlin, Bruxelles, Frankfurt/M., New York, Oxford, Wien: Peter Lang.

WIERZBICKA, A. (1988), *The Semantics of Grammar*, Amsterdam: John Benjamins.

ZELDES, A. (2012), *Productivity in Argument Selection: from Morphology to Syntax*, Berlin: Mouton de Gruyter.

ZIPF, G. (1935), *The Psychobiology of Language: An Introduction to Dynamic Philology*, Boston: Houghton Mifflin.