Les grammaires de constructions à l'épreuve de l'empirie

Guillaume Desagulier

université PARIS DIDEROT PARIS 7

U.F.R. d'Études Anglophones

Synthèse présentée dans le cadre de
l'Habilitation à Diriger les Recherches

Les grammaires de constructions à l'épreuve de l'empirie

Guillaume Desagulier
Maître de conférences à l'Université Paris 8 - Vincennes–Saint Denis

Garant	Pr. Philip Miller Université Paris Diderot U.F.R. d'Études Anglophones
Premier rapporteur	Pr. Adele E. Goldberg Department of Psychology Princeton University
Second rapporteur	Pr. Martin Hilpert Département de Linguistique Université de Neuchâtel
Membre du jury	Pr. Agnès Celle U.F.R. d'Études Anglophones Université Paris Diderot
Membre du jury	Pr. Em. Jacques François Université de Caen–Basse Normandie
Membre du jury	Pr. Dylan Glynn Département d'Études des Pays Anglophones Université Paris 8 - Vincennes–Saint Denis

9 décembre 2016

Guillaume Desagulier
Les grammaires de constructions
à l'épreuve de l'empirie
Synthèse présentée dans le cadre de
l'Habilitation à Diriger les Recherches, 9 décembre 2016
Rapporteurs : Pr. Adele E. Goldberg and Pr. Martin Hilpert
Garant : Pr. Philip Miller

Université Paris Diderot
Bâtiment Olympe de Gouges
U.F.R. d'Études Anglophones
5 rue Thomas Mann
75205 and Paris cedex 13

Remerciements

La recherche n'est pas un chemin solitaire. Mon parcours est jalonné de rencontres, trop nombreuses pour être toutes mentionnées ici.

J'adresse mes plus sincères remerciements à Philip Miller, qui a bien voulu se porter garant de mon Habilitation à Diriger les Recherches. Ses précieux conseils et ses encouragements m'ont permis de mener à bien ce travail.

Je remercie les membres de mon jury d'avoir accepté de lire ou relire mes travaux et de participer à la soutenance de mon Habilitation à Diriger les Recherches.

Depuis 2007, j'ai la chance de travailler dans un laboratoire où règne une atmosphère aussi professionnelle que conviviale : l'UMR 7114 MoDyCo. Je remercie très chaleureusement les collègues qui ont fait de chaque journée passée au quatrième étage du bâtiment A de l'Université Paris Ouest Nanterre un moment très agréable. J'adresse les mêmes remerciements à mes collègues du Département d'Études des Pays Anglophones de l'Université Paris 8 pour leur compréhension pendant les ultimes journées d'écriture.

Je suis redevable à mes parents et à mes ami-e-s pour leurs encouragements tout au long de la rédaction de cette synthèse. William, ce sera bientôt à mon tour de t'encourager.

J'aimerais adresser une mention spéciale à deux collègues dont les qualités humaines sont à la mesure de leur talent professionnel : Brigitte Félix et Antoine Chambaz.

Toute ma reconnaissance va à Fatima, ma compagne, pour son soutien sans faille, et nos enfants, Idris et Hanaé, pour leur infinie patience. Toutes ces heures passées à écrire ont été des heures en moins en votre compagnie. Il me tarde de les rattraper avec vous. Je vous dédie ce travail.

Oui, Hanaé, Papa a bientôt fini son livre.

Table des matières

Table des figures

Liste des tableaux

1 Introduction : de la théorisation de l'usage à sa capture

> *Cognitive linguistics, as a discipline, would have much greater status within the cognitive sciences if they paid more attention to explicating the methods they use, and demonstrate that these provide for consistent, replicable research results.*
>
> **— Raymond W. Gibbs**
> (GIBBS, 2007)

1.1 Retour sur mon parcours depuis la thèse

Mon parcours depuis 2005 comprend deux moments :

— une réflexion théorique dans le prolongement de ma thèse dans le domaine de la socio-pragmatique des constructions ;

— l'adoption des méthodes empiriques de la linguistique de corpus quantitative et l'étude de phénomènes sémantiques dans le cadre des grammaires de constructions.

Ces deux moments sont à l'image des deux générations qui ont marqué la linguistique cognitive. La première propose une théorie de l'usage sur des fondations essentiellement théoriques. La seconde s'investit dans les méthodes empiriques, notamment le recours aux corpus et à l'expérimentation.

1.1.1 Grammaires de constructions et espaces mentaux

J'ai eu la chance de pouvoir rédiger une thèse sur la linguistique cognitive en lien direct avec ses principaux acteurs de l'époque, notamment George Lakoff et Eve Sweetser (dont j'ai pu suivre les cours lors de mon année en qualité de Graduate Student Instructor à l'Université de Californie à Berkeley en 2001-2002), Charles Fillmore, Gilles Fauconnier, Mark Turner, Suzanne Kemmer et Ronald Langacker.

Ma thèse fait le lien entre la Théorie de l'Intégration Conceptuelle (FAUCONNIER, 1985 ; FAUCONNIER, 1997 ; FAUCONNIER et TURNER, 2002) et les grammaires de constructions dans la modélisation des constructions émergentes en anglais contemporain. Ce travail donne lieu à deux publications : DESAGULIER (2003) et DESAGULIER (2005).

1.1.2 Une linguistique de la subjectivité

Dans les mois qui suivent ma soutenance de thèse, fin 2006, Pierre Encrevé et Michel de Fornel m'invitent à les rejoindre au Centre de Linguistique Théorique (CELITH) de l'École des Hautes Études en Sciences Sociales [1]. Les axes forts de ce laboratoire sont la sociolinguistique variationniste, la linguistique interactionniste et l'anthropologie linguistique. À l'image de l'EHESS, qui se veut pluridisciplinaire, le CELITH est un centre expérimental ouvert à de nombreuses disciplines – dont l'ethnométhodologie et la sociologie bourdieusienne – et à plusieurs courants linguistiques théoriques – dont la linguistique cognitive [2].

Le CELITH invite chaque année une personnalité scientifique d'envergure internationale. C'est l'occasion d'écouter mais aussi de connaître personnellement chacune d'entre elles. J'ai le plaisir de rencontrer Bill Hanks (Université de Californie, Berkeley) lors de son cycle de conférences sur la deixis, Charles Goodwin et Marjorie Harness Goodwin (Université de Californie, Los Angeles) lors de leurs cycles de conférence sur l'analyse de conversation en micro-interaction (C. GOODWIN, 2003 ; M. H. GOODWIN, 2006) et Marina Sbisà (Université de Trieste, Italie), invitée pour présenter ses travaux sur les actes de langage. Le père de Charles Goodwin étant aphasique suite à un accident vasculaire cérébral, Charles Goodwin en a fait un sujet d'étude. Les études de cas présentées sur la négociation non- ou quasi-verbale du sens sont certainement les plus marquantes car je travaille à l'époque sur les actes de langage directifs indirects et ses composantes gestuelles et situationnelles.

Michel de Fornel m'associe rapidement à la tenue de journées d'étude (dont une sur l'interaction) et de séminaires. Grâce à lui, je participe activement au Programme de Recherches Interdisciplinaires (PRI) « Anthropologie et Linguistique » entre 2006 et 2008 [3]. L'expérience du PRI est la plus passionnante de mon passage à l'EHESS. En décembre 2006, nous y organisons plusieurs séances sur l'idéologie linguistique. Nous y discutons d'un débat entre d'un côté Ray Jackendoff et Steven Pinker et de l'autre Marc Hauser, Noam Chomsky et Tecumseh Fitch. HAUSER et al. (2002) défendent l'argument déjà soutenu par CHOMSKY (1995) selon lequel la faculté de langage s'articule autour d'un cœur syntaxique abstrait et récursif excluant le lexique. JACKENDOFF et PINKER (2005a) remettent en cause cet argument en soulignant la fragilité empirique de cette théorie. Suite à une clarification des premiers ouvrant la voie à des fondements empiriques (FITCH et al., 2005), JACKENDOFF et PINKER (2005b) s'appuient sur les grammaires de constructions (GOLDBERG, 1995 ; GOLDBERG, 2006), Head-Driven Phrase Structure Grammar (POLLARD et SAG, 1994 ; GINZBURG et SAG, 2000), la Grammaire Cognitive (LANGACKER, 1998), Lexical Functional Grammar (BRESNAN, 1982) et Simpler Syntax (Peter William CULICOVER et JACKENDOFF, 2005) pour appeler à une remise en question de la distinction entre règles et lexique, c'est-à-dire un point central du programme minimaliste. Jackendoff et Pinker ne développent pas explicitement les détails de cette remise en question. Je le fais par l'entremise d'un article publié dans la base de données du PRI (DESAGULIER, 2007c).

1. Depuis 2010, le CELITH est intégré à l'Institut Marcel Mauss (IMM – UMR 8178 CNRS/EHESS) sous le nom de LIAS (Linguistique Anthropologique et Sociolinguistique) – `http://lias.ehess.fr/`.
2. Michel de Fornel est le traducteur en français de *Metaphors We Live By* de George Lakoff et Gilles Fauconnier a travaillé avec Michel de Fornel dans les années 80, au moment où il a élaboré sa théorie des espaces mentaux.
3. Les PRI sont des programmes de recherche exclusivement pluridisciplinaires. Ces programmes sont fortement encouragés par la présidence de l'École pour favoriser les échanges entre les chercheurs.

En 2008, au cours d'une séance consacrée à la théorie des faces et au marquage linguistique de la politesse, je présente le dernier ouvrage de R. J. WATTS (2003) et le relie à des travaux de psychologie sociale en et sur le japonais (DOI, 1973 ; DOI, 1986). Cette intervention fait l'objet d'une publication trois ans plus tard (DESAGULIER, 2011b). Lors du séminaire, je prends conscience du fossé épistémologique qui existe entre les priorités propres à l'anthropologie linguistique et celles de la linguistique. Mon approche centrée sur les structures linguistiques et les manifestations grammaticales de phénomènes culturels parle aux linguistes mais peu aux spécialistes de l'anthropologie linguistique. Cette impression se confirme lorsque, quelques semaines plus tard, le PRI se réoriente sur des problématiques spécifiques à l'anthropologie linguistique, à l'initiative du Directeur d'Études indianiste Francis Zimmerman. Le nouveau programme de recherche se concentre sur l'hypothèse Sapir-Whorf et le parler-chanter. Après une expérience certes très riche en échanges interdisciplinaires, je suis toutefois convaincu (peut-être à tort) que l'anthropologie linguistique à laquelle j'ai été initié privilégie l'étude des structures anthropologiques au détriment des structures linguistiques. Je me distancie du nouveau programme du PRI afin de recentrer mes recherches sur un terrain qui m'est beaucoup plus familier : la subjectivité à l'œuvre dans les grammaires de constructions et ses manifestations (DESAGULIER, 2007b ; DESAGULIER, 2008a ; DESAGULIER, 2011a).

Mon passage au CELITH m'apprend à faire de chaque publication le fruit d'une longue maturation. C'est aussi le produit d'un long questionnement au cours duquel chaque concept fait l'objet d'une épistémologie encyclopédique. Il est successivement défini, appliqué, remis en question, retravaillé et appliqué de nouveau avant d'être une fois de plus soumis à la discussion. Cette expérience est fondatrice au sens ou elle est marque pour moi une ouverture à d'autres champs disciplinaires en sciences humaines.

En 2007, le CELITH est intégré temporairement à l'UMR 7114-MoDyCo (Paris Ouest Nanterre La Défense/CNRS). N'étant pas enseignant chercheur statutaire à l'EHESS et souhaitant poursuivre ma recherche en tant que chercheur titulaire, mon transfert est, quant à lui, permanent. Ce changement est facilité par le fait que le laboratoire est l'organisateur du 3^{e} colloque international de l'Association Française de Linguistique Cognitive en 2009 (`http://www.modyco.fr/aflico3`). Je dirige le comité d'organisation du colloque conjointement avec Philippe Gréa, maître de conférences au département de sciences du langage à Paris Ouest. Nous choisissons d'articuler le colloque autour du thème des grammaires de constructions et de leurs avatars (Fig. 1.1). Nous invitons un panel international de chercheuses et de chercheurs en lien avec les grammaires de constructions de manière à proposer un état de l'art critique sur la question : Hans C. Boas (Université du Texas, Austin, USA), Gilles Fauconnier (Université de Californie, San Diego), Jacques François (Université de Caen), Adele Goldberg (Université de Princeton, USA), Stéphane Robert (CNRS, Llacan), Bernard Victorri (CNRS, Lattice, ENS) et Richard Watts (Université de Berne, Suisse).

Le colloque attire près de 230 participants et donne lieu à un numéro spécial de la revue *CogniTextes* (`https://cognitextes.revues.org/314`).

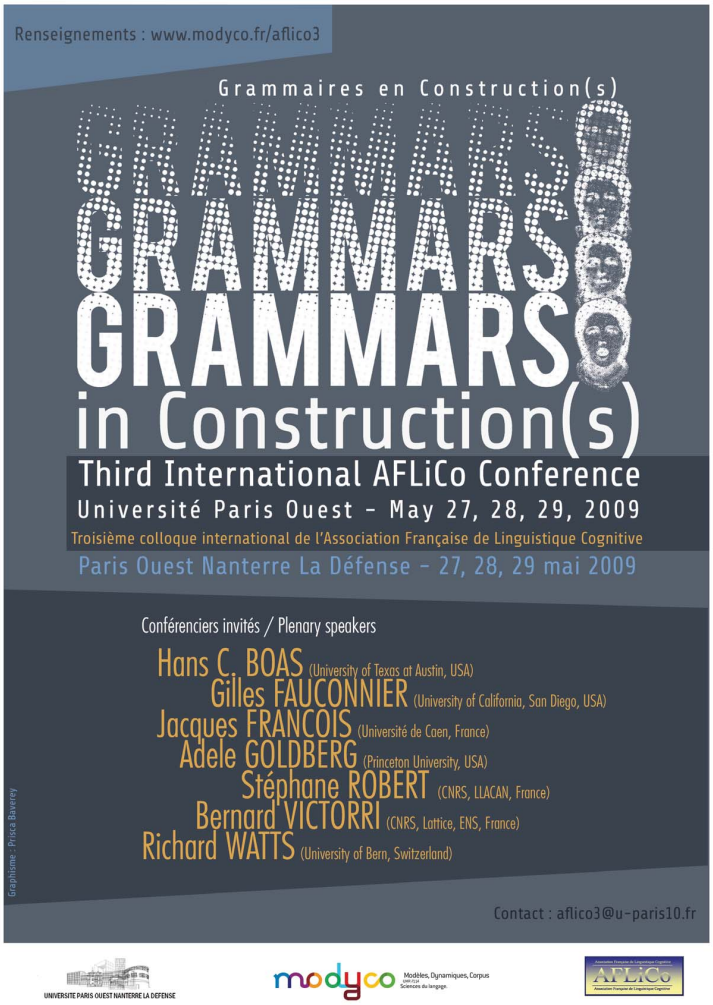


Figure 1.1: L'affiche du 3e colloque de l'AFLiCo

1.1.3 La linguistique de corpus

L'acronyme du laboratoire se décompose comme suit : « Modèles, Dynamiques, Corpus ». Les recherches qui y sont menées sur la constitution et le traitement des corpus m'ont influencé au point de vouloir me former rapidement aux techniques qui y sont liées.

Du 8 au 13 août 2010, j'effectue stage intensif d'une semaine en linguistique de corpus à l'Université de North Texas à Denton. Ce stage, que mon laboratoire accepte de financer, est animé par Stefan Gries, la figure de proue de l'analyse quantitative en linguistique cognitive. J'y apprends le langage informatique R (R Core Team, 2016) utilisé jusqu'alors principalement par les statisticiens. J'apprends les techniques clés de l'analyse de corpus, à savoir :

— les bases de la programmation (la syntaxe, les objets, les boucles, les expressions conditionnelles, etc.) ;

— le traitement des chaînes de caractères (l'encodage, les expressions régulières, la rédaction de requêtes, etc.) ;

— les différents types de corpus électroniques ;

— la manipulation de tableaux de données.

Le stage tenant toutes ses promesses, j'écris ma première étude sur la base d'un corpus constitué le mois d'après. Elle est publiée deux ans plus tard dans un volume collectif sur les constructions du français (Desagulier, 2012a). Mes publications ultérieures se font dans le prolongement de cet apprentissage initial.

1.1.4 Les statistiques appliquées à la linguistique

Une fois formé à la linguistique de corpus, je découvre rapidement les statistiques appliquées à la linguistique de corpus grâce à la lecture de Baayen (2008) et Gries (2013c). 2012 marque le début d'une collaboration intellectuellement stimulante avec Antoine Chambaz du laboratoire Modal'X (Université Paris Ouest Nanterre). Bernard Laks et Christophe Parisse, du laboratoire MoDyCo, Antoine Chambaz et moi créons et animons un séminaire commun. Chaque séance est l'occasion d'écouter deux chercheurs : l'un en mathématiques, l'autre en linguistique.

À la même période, nous soumettons un projet Projet Exploratoire Premier Soutien (PEPS) au CNRS, répondant ainsi à l'appel à projets pluridisciplinaires HuMaIn (Humanités–Mathématiques–Sciences de l'Information). Le projet est retenu et mené de 2013 à 2014. Dans ce cadre, nous organisons une journée d'étude au printemps 2013. Cette journée s'intitule « Contextual Statistics for Construction Grammar ». Nous y rassemblons des interventions des membres du séminaire mentionné au paragraphe précédent et invitons Adele Goldberg.

Ces projets me permettent d'avancer rapidement dans l'apprentissage des méthodes statistiques. Je m'oriente principalement vers l'analyse des collocations et la synthèse des tableaux de fréquences à l'aide de méthodes statistiques exploratoires. Parallèlement, je travaille avec Antoine Chambaz à l'application en linguistique de méthodes semi-paramétriques de pointe en biostatistiques.

En croisant ce que j'ai appris en linguistique de corpus et en statistiques, les quatre dernières années ont été productives. En plus de publier dans des volumes collectifs et des revues à comité de lecture (Desagulier, 2012a ; Desagulier, 2014 ; Desagulier, 2015c ; Desagulier, 2015a ; Desagulier, 2015b ; Chambaz et Desagulier, 2016), j'entreprends de consigner le fruit de mon apprentissage dans un ouvrage chez Springer (Desagulier, à paraître). À la fois un manuel (c'est une requête de l'éditeur) et un ouvrage théorique, ce livre est un objet hybride. Situé dans le prolongement des travaux de Stefan Gries, il reprend, développe et complète Gries (2009) et Gries (2013a) en incorporant de nouvelles méthodes. La version que je joins à cette synthèse est le manuscrit initial, qui doit être envoyé à l'éditeur en novembre 2016. Très proche de la version finale, le manuscrit ne comporte pas encore d'index, dans l'attente des commentaires des relecteurs. La publication est prévue pour juillet 2017.

Après ce bref résumé de mes recherches, je reviens un peu plus en détail sur ce qui m'a conduit, tout comme le courant de la linguistique cognitive, à reconsidérer la place de l'intuition et de l'introspection à l'aune d'une approche beaucoup plus empirique.

1.2 La remise en cause de l'introspection

En rejetant le réductionnisme propre au structuralisme et à la grammaire générative, la linguistique cognitive a mis en avant l'usage et a ouvert la voie à l'empirie. Ce faisant, elle a placé les linguistes de l'usage face aux conséquences de leur orientation, au prix d'une nécessaire remise en question.

Mon expérience en tant que linguiste m'a appris qu'une remise en question est un signe de la bonne santé pour un paradigme théorique. Sur ce point, la linguistique cognitive fait preuve de vivacité. Les débats qui y sont menés sont féconds et souvent porteurs de changements.

1.2.1 Positionnement linguistique

La période mancunienne de William Croft le pousse à prendre du recul vis-à-vis de sa pratique au sein de la linguistique cognitive et est féconde en travaux de nature épistémologique (Croft, 2000 ; Croft, 2001). Ces travaux ont été pour moi des sources d'inspiration. J'aborde ci-dessous un débat qui a eu une influence sur la nature de mes recherches, m'invitant à repenser le statut de l'introspection et à m'ouvrir aux méthodes empiriques.

Croft (1998) s'interroge sur les preuves linguistiques (*linguistic evidence*) en mesure de fonder les grands modèles censés rendre compte du stockage mental des unités linguistiques sous forme de représentations. Croft distingue quatre modèles :

1. le modèle des entrées indépendantes (*the independent entries model*) ;
2. le modèle fondé sur la polysémie (*the polysemy model*) ;
3. le modèle dérivationnel (*the derivational model*) ;
4. le modèle pragmatique (*the pragmatic model*).

Soit la forme a dans une langue donnée et deux sens clairement distincts dans l'usage, U_1 et U_2. Prenons deux exemples délibérément simplifiés à outrance. En lexicologie, *arm* désigne originellement un membre – U_1, (1a) – et par extension l'accoudoir d'un fauteuil – U_2, (1b). En grammaire, *at* est une préposition spatiale statique – U_1, (2a) – appliquée par extension pour introduire un participant détrimentaire – U_2, (2b).

(1) a. I broke my arm. (U_1)
    b. I broke the arm of my favorite chair. (U_2)

(2) a. He is staying at a very nice hotel. (U_1)
    b. He threw a stone at me. (U_2)

Selon le modèle des entrées indépendantes, chaque appariement forme/sens est ancré de manière indépendante dans la grammaire mentale : $[a/U_1]$ et $[a/U_2]$[4]. Même si les deux unités sont liées par un lien étymologique, elles sont synchroniquement indépendante puisque sanctionnées par des usages distincts. Selon le modèle fondé sur la polysémie, chaque appariement forme/sens est également ancré de manière indépendante dans l'esprit, mais un lien sémantique relie $[a/U_1]$ et $[a/U_2]$ au sein d'un réseau. Ce lien est connu des locuteurs. Selon le modèle dérivationnel, seul $[a/U_1]$ est représenté dans la compétence du locuteur. L'appariement (a/U_2) n'est pas ancré ou stocké : il est dérivé par une règle spécifique à la langue[5]. Le modèle pragmatique est proche du modèle dérivationnel. Seule la forme schématique commune aux deux appariements $[a/U]$ est ancrée. Les réalisations des deux appariements (a/U_1) et (a/U_2) sont dérivées via des règles contextuelles et discursives.

Trois de ces modèles se retrouvent dans les approches formelles : le premier, le troisième et le quatrième. Ces modèles sont non-redondants : une unité ne peut pas être stockée plusieurs fois dans la grammaire. Le second modèle est considéré comme le modèle par défaut en linguistique cognitive (la réalité est en fait plus complexe puisque le modèle fondé sur la polysémie est amené à se combiner avec le modèle des entrées indépendantes lorsque l'on prend en compte la variation).

Alors que les querelles entre approches formelles (principalement le modèle génératif) et les approches cognitives-fonctionnelles se déroulent en terrain théorique, Croft postule qu'aucun des quatre modèles ne peut être infirmé a priori sur la base de l'intuition. Par exemple, le rejet de la redondance par les théories formalistes au prétexte que celle-ci serait psychologiquement impossible a été contré par plusieurs travaux de nature exemplariste, notamment en morphologie (Bybee, 1985). Parallèlement, le rejet systématique du modèle des entrées indépendantes postulé par la linguistique cognitive est difficile à justifier pour une raison paradoxale : en vertu des fondamentaux de la linguistique de l'usage, rejeter ce modèle revient à dire que les locuteurs sont toujours en mesure d'établir un lien entre deux formes similaires et à jeter des ponts sémantiques entre eux. Si cela est fortement probable pour des cas très simples, il est difficile de le croire pour des exemples plus complexes (par exemple le lien entre un *legs* et le verbe *laisser*, qui explique la présence du <*s*> final dans le nom)[6]. En somme, si l'on peut dresser les modèles les uns contre les autres sur le plan théorique, il est impossible de privilégier l'un plutôt qu'un autre. Croft appelle non pas à se départir de l'intuition mais à ne pas surestimer les conclusions faites sur la seule base de l'intuition.

La linguistique cognitive objecte à la linguistique formelle de tomber dans la « rule/list fallacy » (Langacker, 1987, p. 29). Inversement, la linguistique formelle d'inspiration générativiste reproche à la linguistique cognitive de tomber dans la « generality fallacy ». L'honnêteté de Croft est de souligner qu'en l'absence d'empirie, la linguistique cognitive est dans un état de tension entre objectif de généralité et particularisme. C'est le propre de la linguistique de l'usage au sein de la linguistique cognitive d'accorder une importance

4. En Grammaire Cognitive, la forme est notée en minuscules et le sens en majuscules. Les crochets indiquent que l'appariement est ancré cognitivement. Le signe / dénote un lien symbolique entre la forme et le sens.

5. L'absence d'ancrage de l'appariement est dénoté par l'emploi de parenthèses au lieu des crochets.

6. A contrario, les locuteurs sont parfois amenés à jeter des ponts sémantiques entre des unités dont les sens originaux sont originellement séparés. Ce phénomène est connu sous le nom d'étymologie populaire (par exemple l'inexacte mise en relation sémantique entre *un legs* et le verbe *léguer*).

particulière aux idiosyncrasies, ce que la linguistique générativiste exclut des règles et consigne au rang de liste.

La démarche proposée par Croft retient mon attention pour trois raisons. La première est qu'elle n'exclut aucun modèle théorique de la représentation linguistique a priori. Plutôt que de placer ces quatre modèles dans un schéma d'exclusion mutuelle, Croft invite a les placer dans un continuum. Le recours a un modèle n'est pas une affaire de credo mais d'adaptation de l'outillage aux questions de recherche. Que l'on ne se trompe pas quant aux préférences théoriques de l'auteur. Il devine qu'au prisme de l'empirie, ce sont bien les approches cognitives-fonctionnelles qui tireront leur épingle du jeu.

La seconde raison est l'attrait pour une démarche qui rappelle le test d'hypothèses en statistique. Cette démarche consiste à postuler deux hypothèses, une hypothèse nulle (H_0, qui est l'hypothèse d'absence d'effet significatif) et une hypothèse alternative (H_1, à savoir l'hypothèse d'un effet significatif). Lorsque l'on teste une hypothèse dans une démarche statistique, on est amené non pas à prouver que l'une des deux hypothèses est vraie, mais à prouver que l'on a de bonnes raisons de rejeter l'une d'elles. De la même manière, partant du principe qu'aucune intuition ne peut considérer un seul modèle parmi les quatre comme étant le meilleur, Croft postule que la démarche empirique ne sert pas à établir la validité d'un modèle mais à rejeter les modèles les moins probables. En somme, c'est seulement sur la base de ce que nous apporte l'empirie que l'on est en mesure de restreindre le champ des possibles quant à la représentation des unités dans la grammaire mentale. Cette méthodologie étant la base de l'approche inférentielle, antérieure à la constitution de la linguistique comme discipline, elle mérite qu'on la retienne.

La troisième raison est liée au choix des preuves empiriques et à l'exposé de leur utilisation. Croft en propose deux types : l'expérimentation psycholinguistique et les données de corpus (1998, p. 168–170). Elles interviennent à trois niveaux. Je les résume dans le tableau 1.1

De la part des linguistes travaillant sur l'empirie, les réactions sont venues du camp psycholinguistique.

1.2.2 Positionnement psycholinguistique

L'article de Croft provoque deux réponses : SANDRA (1998) et TUGGY (2001). La première est celle d'un psycholinguiste. Il n'y a rien de surprenant lorsque l'on sait que l'empirie est principalement expérimentale et que la psycholinguistique a depuis longtemps la main sur ce type d'études. Sandra affirme que la linguistique générative et la linguistique cognitive partagent les mêmes prémisses : pour comprendre les phénomènes langagiers, il faut les interpréter à l'aune de mécanismes cognitifs généraux. Ce qui importe, c'est de déterminer comment les représentations linguistiques sont stockées et traitées dans l'esprit humain. Sandra laisse à penser que la définition du mentalisme dans les années 50 marque l'apparition de la psycholinguistique telle que nous la connaissons aujourd'hui. À l'époque, la linguistique aurait dû se cantonner à l'analyse linguistique [7] et confier à la psycholinguistique la tâche de décrire les composantes de l'esprit humain. Il dresse le constat suivant :

7. Il écrit : « As far as linguists are concerned, their first job is to analyze language » (1998, p. 364)

Tableau 1.1: Les applications de l'empirie selon CROFT (1998)

problème de représentation mentale	description du problème	modèle rejeté	expériences psycholinguistiques	linguistique de corpus
usages distincts	deux emplois linguistiques sont considérés comme distincts par les locuteurs	modèle pragmatique	tâches de jugement sur la similitude de phrases ; les locuteurs distinguent différents degrés de séparation entre les prépositions (SANDRA et RICE, 1995)	deux sens de *eat* sont distingués en corpus : "consommer" et "dîner" (CROFT, 1995)
usages conventionnels	si l'usage d'une forme est conventionnel, aucune dérivation (grammaticale ou pragmatique) n'est possible	modèles pragmatique et grammatical	temps de réponse pour déterminer les réponses à des actes de langage indirects (GIBBS, 1994)	plus une forme est fréquente, plus elle est ancrée (BYBEE, 1985 ; LANGACKER, 1987)
usages affiliés	deux emplois linguistiquement affiliés sont perçus par les locuteurs comme non affiliés (et inversement)	modèle fondé sur la polysémie	des formes linguistiquement affiliées sont perçues comme totalement différentes par les locuteurs, par ex. l'emploi temporel vs. spatial des prépositions (SANDRA et RICE, 1995)	des syntagmes composés (nom-nom) inventés sont définis par des sujets ; les définitions sont influencées par les syntagmes composés les plus fréquents en corpus (RYDER, 1994)

> However, (...) up to the present day, linguists – both of the generative and cognitive tribes – have continued to relate language facts to aspects of the human mind. (1998, p. 362).

Tout comme Croft, Sandra se méfie des conclusions fondées uniquement sur l'intuition lorsqu'elles ont trait aux représentations linguistiques :

> All forms of scientific activity take place within a framework of a set of beliefs about how the object of inquiry should be studied and what constitutes a plausible (from the point of view of the scientist) working hypothesis. However, these (often) tacit assumptions strongly bias the way data are analyzed, which means that linguists must be aware that sometimes their analyses are merely restatements of their starting assumptions (in a way, they see in the data what they are looking for).

Les conclusions de SANDRA (1998) vis-à-vis des prétentions de la linguistique sont tellement pessimistes que l'on est en droit d'y voir une provocation délibérée. Que la tâche de la linguistique puisse ne se résumer qu'à l'analyse (comprendre : elle ne peut légitimement pas généraliser les résultats de ses recherches au-delà des unités linguistiques) revient à limiter considérablement son champs d'action. Par ailleurs, la distinction artificielle entre linguistique et psycholinguistique ne s'applique pas à la plupart des chercheurs de renom en linguistique cognitive, par exemple Dan Slobin (Université de Californie, Berkeley), Sally Rice

(Université d'Alberta, Canada), Michael Tomasello (Max Planck Institute for Evolutionary Anthropology, Leipzig) ou Adele Goldberg (Université de Princeton).

La critique de Sandra soulève un point qui me semble au moins partiellement justifié au vu du contexte de la publication. Premièrement, Dominiek Sandra collabore avec des linguistes cognitivistes et est présent à l'International Cognitive Linguistics Conference d'Amsterdam en 1997. Au moment de la publication de l'article, les colloques de l'ICLC sont des lieux de débat principalement théoriques. Hormis les travaux de nature psychologique sur les universaux du langage ou en acquisition du langage, la plupart des discussions se font au niveau intuitif. Or, la plupart des grandes thèses sur l'Intégration Conceptuelle, la Grammaire Cognitive, ou la Théorie de la Métaphore n'hésitent pas à prendre des positions tranchées sur le plan cognitif, c'est à dire des représentations mentales. La linguistique de corpus n'a pas encore percé et les protocoles d'expérimentation ne sont maîtrisés que par une minorité de linguistes. Aussi virulent soit-il, l'article de Sandra a au moins le mérite d'alerter la communauté des linguistes cognitivistes sur le besoin d'étayer empiriquement leurs conclusions, à un moment où les grandes thèses liminaires en linguistique cognitive (FAUCONNIER, 1985 ; LAKOFF, 1987 ; LANGACKER, 1987 ; LANGACKER, 1990 ; FAUCONNIER, 1997 ; GOLDBERG, 1995) sont en plein développement. Au-delà de la provocation, j'associe donc à cet article un second niveau de lecture visant à obtenir de la linguistique cognitive qu'elle s'outille à la hauteur de ses prétentions.

TUGGY (2001) intervient dans l'échange trois ans plus tard. Il reprend la critique de SANDRA (1998, p. 370) concernant la préférence a priori des linguistes cognitivistes pour le modèle fondé sur la polysémie (ce qu'il appelle "polysemy fallacy"). Selon lui, si ce modèle est effectivement privilégié, c'est parce qu'il est le plus prudent des quatre proposés par Croft. Cette prudence est le signe d'une modestie et ne justifie aucunement les conclusions pessimistes de Sandra. Comme le souligne GRIES (2013b, p. 96), la revue *Cognitive Linguistics* publie à cette époque ses premières contributions véritablement empiriques : TOMASELLO et BROOKS (1998), PALANCAR (1999) et GRIES (1999).

Quelques années plus tard, le psychologue Raymond Gibbs relance la discussion (GIBBS, 2007) [8]. De manière surprenante, du débat décrit plus haut il ne cite que SANDRA (1998). Son point de départ est le suivant : la linguistique cognitive est critiquée de l'extérieur. Pour être crédible, il lui faut prouver ses conclusions à l'aide de méthodes empiriques. Reconnaissant qu'il existe un fossé de compétences entre les psychologues, rompus aux protocoles expérimentaux, et les linguistes, qui se cantonneraient à l'introspection à défaut de savoir manipuler les méthodes empiriques, il propose aux linguistes d'adopter des méthodes "indirectes", qui sont à leur portée :

> Nonetheless, there are various empirical, experimental techniques that are part of the arsenal of « indirect methods » used in psycholinguistics that have proven to be quite useful in providing support for many of cognitive linguists' claims about mind and language. (2007, p. 2–3)

8. Gibbs est connu en linguistique cognitive pour avoir exploré le traitement cognitif de la distinction sens littéral/sens figuré et pour avoir cherché à étayer les principes de la Théorie de la Métaphore Conceptuelle de Lakoff.

Ces techniques sont les suivantes : l'imagerie mentale[9], des jugements sur le sens des métaphores en contexte ou des expériences amenant les sujets à décrire ce qu'ils ressentent physiquement lorsqu'ils entendent des métaphores. Étrangement, aucune référence n'est faite à la linguistique de corpus, l'empirie se résumant, pour Gibbs à l'expérimentation[10].

Au-delà de ces propositions, la véritable valeur de cet article est d'amener les linguistes à formuler des hypothèses falsifiables. L'assise empirique passe par la formulation d'hypothèses et par la falsification de celles qui invalident les prémisses de la théorie :

> Thus, each hypothesis must be stated in such a way that it can be experimentally/empirically examined and shown to be possibly false (and if not shown to be false, then one can reject the null hypothesis and conclude that there is evidence in support of the hypothesis). (GIBBS, 2007, p. 7)

Cette méthodologie est fondamentale mais elle n'est pas nouvelle en linguistique cognitive. Selon moi, deux raisons expliquent cette redite. La première raison est que Gibbs ne se concentre que sur une partie de la linguistique cognitive : la métaphore conceptuelle. Il est vrai que la grande majorité des travaux linguistiques porte sur l'application intuitive de la Théorie de la Métaphore Conceptuelle au discours, aux textes littéraires, voire à la gestualité, sans proposer de cadre véritablement expérimental. La seconde raison est liée à la première : Gibbs ne prend pas en compte l'essor de la linguistique quantitative depuis le début des années 2000 (GRIES, 2003a ; GRIES, 2003b ; GRIES et STEFANOWITSCH, 2006), y compris lorsque celle-ci croise ses méthodes avec l'expérimentation psycholinguistique (GRIES, HAMPE et al., 2005). D'un côté la prise de position de Gibbs est symptomatique d'une équation trop rapide entre empirie et expérimentation psychologique au détriment de la linguistique de corpus quantitative. D'un autre côté, elle est compréhensible au vu de la minorité des publications empiriques jusqu'à la fin des années 2000.

1.2.3 L'intuition n'est finalement pas évacuée

Le sens pose un défi en terme d'opérationnalisation et il peut sembler vain de vouloir le capturer empiriquement. Ceci explique en partie pourquoi la réaction au débat décrit plus haut vient des spécialiste de sémantique cognitive. TALMY (2007) appelle les linguistes à ne pas se départir de l'intuition et de l'introspection.

Si l'expérimentation fournit un examen détaillé du comportement linguistique d'un sujet unique, elle conduit l'expérimentateur à décontextualiser le facteur linguistique étudié. Selon Talmy, l'expérimentation conduit en effet à isoler un facteur linguistique des autres phénomènes qui lui sont consubstantiels. En linguistique cognitive, selon Talmy, c'est bien l'étude de l'intégration globale des facteurs linguistiques qui président à la manifestation d'un phénomène linguistique qui importe, et non l'étude de la décomposition.

Si les corpus fournissent une quantité de données textuelles permettant d'étudier l'ancrage cognitif ou la distribution des alternances, ils ne permettent pas selon lui d'extraire directe-

9. L'imagerie mentale (à ne pas confondre avec l'imagerie cérébrale) repose sur des questionnaires permettant par exemple à un sujet de décrire ce à quoi il pense lorsqu'il entend un idiome décontextualisé tel que *spill the beans*.

10. Il en va de même pour Sandra. On pourrait presque objecter à la psycholinguistique de fonctionner en vase clos.

ment des schémas linguistiques abstraits. En effet, les corpus composés d'échantillons de langue naturelle en conversation contiendraient une portion significative d'énoncés elliptiques. Par ailleurs, les corpus ne permettraient pas de faire la distinction entre les énoncés bien formés ou remplissant les conditions de félicité. Selon Talmy, c'est un problème dans la mesure où notre compétence linguistique s'articulerait sur des unités aux propriétés grammaticales et sémantiques idéales :

> Again, our linguistic cognition is organized so as to have abstracted out the ideal grammatical and semantic properties of particular lexical forms and constructions, ones that can emerge through introspection in a deliberative process like writing, but that are commonly breached in the kind of fluent speech recorded in corpora. (TALMY, 2007, p. xix)

Parce qu'elles contiennent des énoncés imparfaits, les données de corpus seraient en porte-à-faux vis-à-vis de la structure de notre compétence linguistique. Sur ce point, Talmy propose une vision de la linguistique cognitive très éloignée des préceptes de la linguistique de l'usage.

Plus intéressante à mon sens, car beaucoup plus au fait des techniques expérimentales et des méthodes d'exploitation de corpus, est la position de GEERAERTS (2010a). Selon lui, l'intuition est un moment décisif dans ce qu'il appelle « le cycle empirique ». Dans sa version originale, le cycle empirique comprend quatre moments :

— la théorie ;
— l'hypothèse ;
— l'expérimentation ;
— l'analyse.

Ces quatre moments sont reliés comme suit :

— la théorie amène à formuler une hypothèse ;
— l'hypothèse est testée par l'entremise d'une expérience ;
— l'expérience donne lieu à une analyse ;
— tant que l'analyse ne donne pas de résultats convaincants, le linguiste reformule l'hypothèse ou modifie le protocole expérimental ;
— si l'analyse donne des résultats convaincants, elle informe la théorie en retour.

L'intuition du linguiste intervient bien au départ de l'étude car elle conditionne le cheminement de la théorie à l'hypothèse. Elle intervient également à la fin car elle conditionne le cheminement de l'analyse à la théorie.

À la différence des remarques de Talmy, le modèle de Geeraerts est presque exploitable en tant que tel pour le chercheur. Dans ma monographie consacrée aux méthodes en linguistique de corpus quantitative (DESAGULIER, à paraître), je développe le cycle empirique de manière à fournir un organigramme directement exploitable pour les chercheurs (Figure 1.2). Chacune de ces étapes (y compris la collecte de données) porte la marque de la subjectivité du linguiste. La remise en cause de l'introspection conduit non pas à l'évacuer, mais à l'intégrer aux méthodes empiriques.

Les chapitres qui suivent reviennent plus en détail sur les étapes de mon parcours, depuis mes travaux théoriques sur les grammaires de constructions jusqu'aux travaux les plus récents.

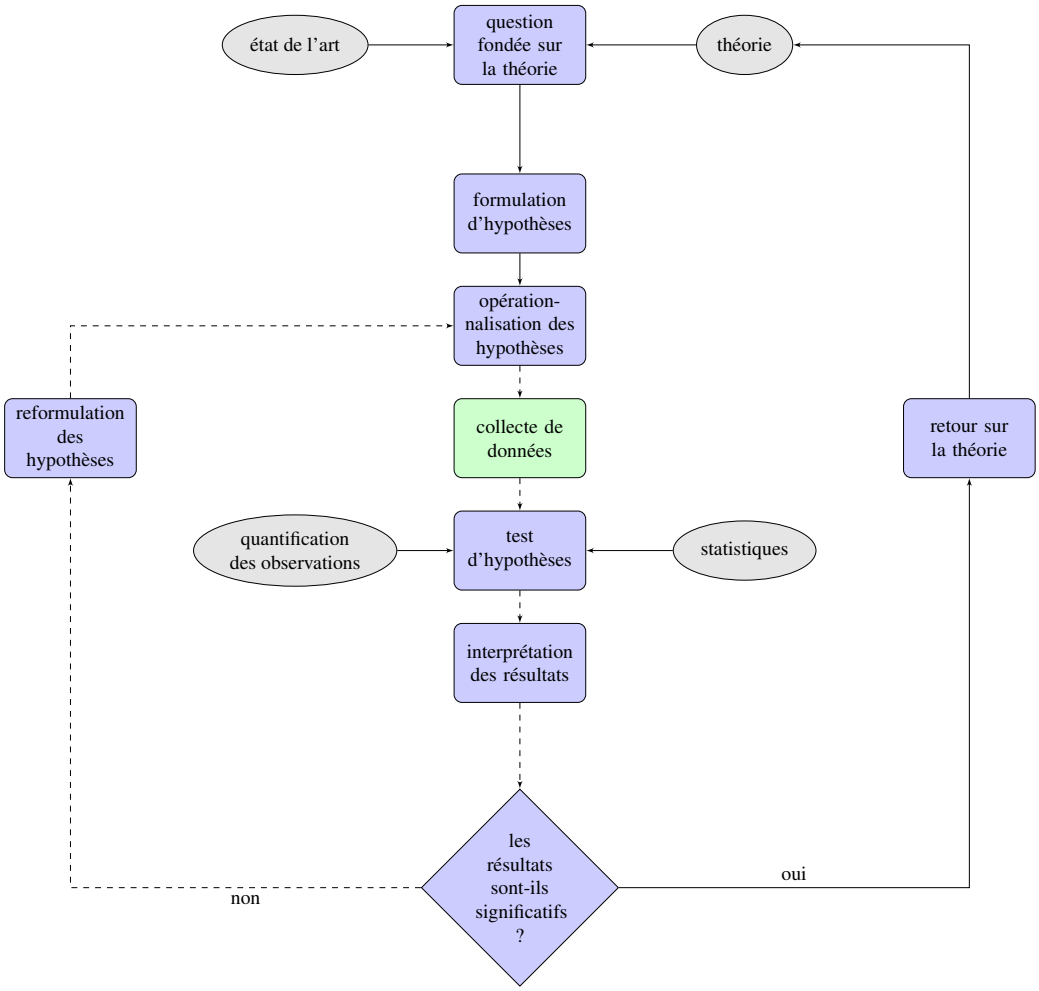


Figure 1.2: Le cycle empirique adapté à la linguistique de corpus

Ces derniers examinent les prétentions de la théorie à l'aune de méthodes empiriques et quantitatives.

2 Les grammaires de constructions : entre théories et grilles d'analyse

> *No one doubts that constructions exist – that is, that there is an active construction, a passive construction, etc. The only question is: what is their status?*
>
> **— Noam Chomsky**
> (« Beyond Linguistic Wars. An Interview with Noam Chomsky » 2011, p. 22)

2.1 Introduction

J'ai découvert les grammaires de constructions à l'occasion d'un séjour à l'Université de Berkeley, Californie, d'août 2001 à juin 2002. Mon premier contact avec ce paradigme fut la lecture de GOLDBERG (1995). Je cherchais alors un appareil théorique me permettant de rendre compte de blends lexico-syntaxiques dans la formation de modaux émergents en anglais américain. N'ayant alors que peu de recul vis-à-vis de ce que je découvrais, je concluais qu'il s'agissait d'une théorie homogène et unifiée que je pouvais traduire par « grammaire de construction ». J'ai très vite compris à la lecture d'autres travaux qu'il fallait plutôt parler de « grammaires de constructions » [1].

Les grammaires de constructions n'ont rien d'une théorie unifiée, hautement formalisée et regroupée autour d'un quelques travaux fondateurs (HOFFMANN et TROUSDALE, 2013). Pourtant, elles ont leur conférence internationale bisannuelle [2], une collection chez John Benjamins [3], une revue [4] et font l'objet de manuels (HILPERT, 2014) et autres publications trop nombreuses pour être citées ici.

Cette situation presque paradoxale attire les critiques, suscite les interrogations, voire l'incompréhension, tout comme elle attise l'intérêt [5]. De mon côté, c'est le fait qu'un paradigme puisse être le point de rencontre de la linguistique formelle et de la linguistique cognitive–fonctionnelle qui a fait des grammaires de constructions mon port d'attache.

1. La transition du singulier au pluriel est également attestée entre LEGALLOIS et FRANÇOIS (2006) et FRANÇOIS (2008)
2. `http://www.cognitivelinguistics.org/en/event/detail/conferences-on-construction-grammar`
3. `https://benjamins.com/#catalog/books/cal`
4. `https://benjamins.com/#catalog/journals/cf/main`
5. Le numéro 26 des Cahiers du CRISCO (FRANÇOIS, 2008) intitulé « Les grammaires de constructions : un bâtiment ouvert aux quatre vents » est révélateur de l'accueil réservé initalement en France aux Grammaires de Constructions.

Les grammaires de constructions sont donc d'une constellation d'approches regroupées autour d'un certain nombre de principes et diamétralement opposées sur d'autres. D'un côté, un consensus se dégage quant à l'hypothèse de départ : la grammaire est un inventaire structuré d'assemblages symboliques de forme et de sens. De l'autre, la question des unités pouvant prétendre au statut de construction diffère en fonction des approches, de même que l'appareil théorique et formel qui sous-tend l'analyse constructionnelle.

Dans ce chapitre, je rends compte de la géométrie variable des grammaires de constructions au travers d'une épistémologie succincte et d'une typologie des différentes approches. J'extrais de cette hétérogénéité ce sur quoi je me suis appuyé dans mes recherches. Je montre notamment que l'atout des grammaires de constructions est précisément de ne pas être une théorie unifiée mais un ensemble de grilles d'analyse permettant d'apporter un éclairage nouveau sur des problématiques anciennes.

2.2 Les idiomes : le grain de sable dans la mécanique générative

On sait depuis bien avant l'arrivée grammaires de constructions que le sens de certaines expressions complexes s'analyse à la manière de celui des mots, à savoir de manière plus conventionnelle que compositionnelle. SAUSSURE (1916) observe que les syntagmes ne se combinent pas tous aussi librement. Il y est d'autant plus sensible que cela l'amène à pondérer ce qui relève de la langue, aux règles strictement contraignantes, et ce qui relève de la parole, lié à la liberté du locuteur. Dans le sillage de Ferdinand Saussure, BALLY (1921, p. 66) développe le concept d' « unité phraséologique », à l'articulation non plus de la langue et de la parole individuelle, mais de la langue et du discours collectif :

> Dans la langue maternelle, l'assimilation des faits de langage se fait surtout par les associations et les groupements dans lesquels l'esprit fait entrer les mots. Ces groupements peuvent être passagers, mais, à force d'être répétés, ils arrivent à recevoir un caractère *usuel* et à former même des unités *indissolubles*. Il faut « penser » ces groupements comme le fait le sujet parlant sa langue maternelle. Entre les cas extrêmes (groupements passagers et unités indécomposables) se placent des groupes intermédiaires appelés *séries phraséologiques* (p. ex. les séries d'intensité et les périphrases verbales). (...)

Bally se fonde sur les manifestations du figement dans l'usage des locuteurs natifs pour mieux en cerner les enjeux dans l'acquisition d'une langue seconde. Cet extrait porte en germe plusieurs thèmes développés par la linguistique cognitive. Le premier est la prise en compte de l'activité conceptualisante du sujet parlant (LANGACKER, 1987 ; LANGACKER, 1999). Le second est une définiton avant l'heure de la linguistique de l'usage (LANGACKER, 1988) : c'est par la répétition en discours, au sein une communauté de locuteurs, que des formes linguistiques s'ancrent dans la grammaire mentale de chacun d'entre eux. Le troisième , peut-être le plus important, est l'existence d'un gradient dans la manifestation du figement, qui annonce la typologie des idiomes par FILLMORE, KAY et O'CONNOR (1988) et les problèmes que pose le principe de compositionnalité en sémantique cognitive (LANGACKER,

1987, p. 448). Au paragraphe suivant, Bally développe l'idée selon laquelle le principe de compositionnalité n'est pas systématique :

> Les unités phraséologiques se reconnaissent à certains indices extérieurs et intérieurs : les premiers se déduisent de la forme des groupes, les autres (seuls importants), de la manière dont les groupes sont conçus par l'esprit. Les principaux de ces indices sont : l'équivalence de la locution à un mot unique ; l'oubli du sens des éléments (notamment dans les locutions de forme analogue) ; la présence, dans la locution, d'archaïsmes de mots, de sens ou de syntaxe ; l'ellipse, etc.

Que le sens d'une expression complexe ne soit pas systématiquement décomposable sémantiquement en composants plus simples n'est devenu problématique qu'avec l'avènement de la linguistique générative.

Qu'une expression complexe puisse être intrinsèquement signifiante va à l'encontre de l'économie d'un modèle par essence minimaliste (CHOMSKY, 1995). Si l'on pose qu'une langue n'est pas un système de règles mais un ensemble de spécifications de paramètres dans un système invariant de principes de grammaire universelle, alors les expressions complexes sont des phénomènes irréguliers. Elles résultent de l'interaction de principes figés et de paramètres pré-établis. Si l'on pose également que les règles de syntaxe combinent les mots en des groupes plus larges et que les syntagmes dénotent des concepts complexes (des prédicats ou des propositions), ces règles sont purement combinatoires et n'ajoutent rien de conceptuel, contrairement aux mots qu'elles combinent. Afin de comprendre cette logique, au risque de caricaturer, appuyons-nous sur une analogie entre arithmétique et syntaxe [6]. Si l'on change l'association entre les nombres dans une séquence, en passant par exemple de $(10 \times 3) + 2$ à $10 \times (3 + 2)$, on change ce que la séquence arithmétique dénote (car $32 \neq 50$) mais pas ce que les nombres dénotent. En syntaxe, cela revient à dire que changer les associations syntaxiques dans une chaîne de mots ne change que le sens de la chaîne, pas celui des mots. Les grammaires de constructions rejettent cette analogie au motif que les mots lexicaux (les noms, les verbes, les adjectifs, etc.) n'entrent pas dans le même système de référence que les nombres. Le contexte syntaxique détermine ce que chaque mot dénote dans une chaîne. Il détermine par la même occasion le comportement distributionnel de ce mot (GOLDBERG, 1995 ; GOLDBERG et JACKENDOFF, 2004). C'est dans ce contexte qu'il faut comprendre la regain d'intérêt pour les constructions idiomatiques et plus généralement les expressions complexes à divers degrés de figement dans les années 80.

À ce titre, l'apport de FILLMORE, KAY et O'CONNOR (1988) est d'avoir montré que si les idiomes lexicalement figés (3–4) avaient eu toute l'attention des linguistes, les idiomes formels (5–6) avaient été injustement laissés de côté.

(3) In bocca al lupo (littéralement « dans la gueule du loup » ; « bonne chance »)

(4) Break a leg !

(5) The more I know about humans the more I love my dog.

(6) What is my best friend doing under the bed ?

6. Cette analogie m'a été proposée par plusieurs collègues générativistes au cours de discussions en marge de colloques.

La construction idiomatique *the X-er the Y-er* en (5) n'est pas réductible à une règle minimale de formation phrastique permettant de relier deux comparatifs dans un sens conditionnel. Aucune règle minimale n'existe non plus pour rendre compte de la construction *WXDY* en (6) qui se fonde sur une structure interrogative pour exprimer l'incongruité. Les constructions idiomatiques plus ou moins lexicalisées n'ont pas leur place dans le module grammatical de la grammaire générative. Nous savons qu'en grammaire générative, toute information spécifique, taxinomique ou non prédictible est placée dans le lexique.

Le renouveau des études sur les idiomes, appelé par FILLMORE, KAY et O'CONNOR (1988) et poursuivi par NUNBERG et al. (1994) et Peter W. CULICOVER et JACKENDOFF (1999) a poussé la linguistique générative à rendre compte d'aspects fondamentaux de la compétence grammaticale tout en posant les bases des grammaires de constructions telles que nous les connaissons aujourd'hui. Alors qu'un pan de la linguistique a rejoint les grammaires de constructions pour relever le défi des constructions idiomatiques sans quitter pour autant le cadre de la linguistique générative, le courant de la linguistique fonctionnelle s'est approprié ce courant émergent en le reliant au « cognitive commitment » (LAKOFF, 1990).

2.3 Une typologie à géométrie variable

Les principaux courants qui composent les grammaires de constructions s'accordent sur une définition axiomatique de la construction, à savoir un assemblage symbolique de forme et de sens, ce que résume la Figure 2.1 (CROFT et CRUSE, 2004) [7]. La composante formelle est pluristratale. Elle englobe la phonologie, la morphologie, le lexique et/ou la syntaxe. La composante sémantique n'est pas limitée au sens puisqu'elle peut inclure des propriétés pragmatiques et discursives (voir Section 2.4.1).

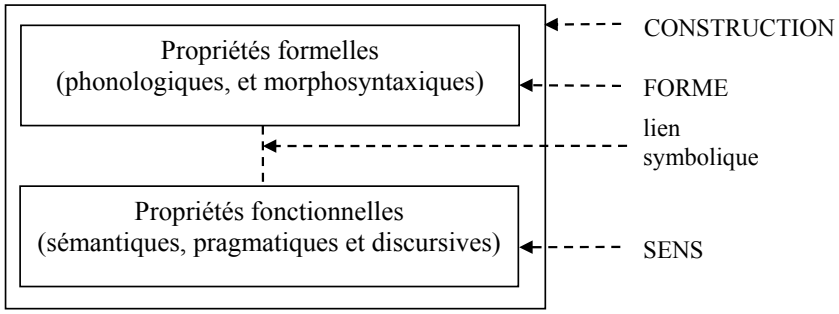


Figure 2.1: Anatomie d'un assemblage de forme et de sens

Si les constructions partagent des principes, elles se distinguent toutefois sur plusieurs points, ce qui a donné lieu à plusieurs typologies (CROFT et CRUSE, 2004 ; GOLDBERG, 2006 ; GRIES, 2013b). CROFT et CRUSE (2004, chapitre 10) distinguent quatre courants :

— Construction Grammar ;

— la grammaire de constructions de LAKOFF (1987) et GOLDBERG (1995) ;

— la Grammaire Cognitive ;

— la Grammaire de Constructions Radicale.

7. Voir également DESAGULIER (2007b)

2.3.1 *Berkeley Construction Grammar*

Le premier de ces courants, la Grammaire de Constructions, est l'approche originelle puisqu'elle se fonde sur les travaux inspirés de Charles Fillmore et Paul Kay (Fillmore, Kay et O'Connor, 1988 ; Fillmore et Kay, 1995 ; Fillmore, 1997 ; Kay et Fillmore, 1999). Goldberg (2006, p. 213–214) la nomme *Grammaire de Constructions d'Unification*, même s'il semble que l'appellation *Berkeley Construction Grammar* (BCG) soit à présent plus fréquente (Gries, 2013a ; Desagulier, 2015a). Kay (1995, p. 171) définit son approche comme étant « non modulaire, générative, non dérivationnelle, monostratale et unifiée ». Une construction est monostratale au sens où elle ne se constitue pas par la projection *ad hoc* d'une strate formelle sur une composante sémantique. Elle ne s'appuie pas sur un mécanisme de dérivation depuis une structure syntaxique profonde jusqu'à une structure de surface. Légitimée par l'usage, une séquence formelle est associée empiriquement à des effets principalement sémantiques et pragmatiques qui émergent en discours. Une fois ce couplage empirique ratifié dans et par l'usage, il est conventionnellement établi comme construction. Si la Grammaire de Constructions s'affranchit de l'hypothèse transformationnelle, elle demeure générative au sens où elle cherche à rendre compte des expressions infiniment productives d'une langue donnée de manière économique et exhaustive. Il s'agit bien d'une approche fondée sur l'usage, qui procède de l'empirie à la formalisation (approche de type « bottom-up ») et non inversement (approche de type « top-down ») car le « construit » n'est pas généré à partir d'une construction, il est ratifié par les contraintes formelles et sémantiques de la construction. La Grammaire de Constructions se fonde donc sur une générativité à rebours (Desagulier, 2011a, p. 103). La Grammaire de Constructions est l'approche la plus formelle. Elle est en cela très proche de la grammaire d'unification HPSG (Pollard et Sag, 1994). L'inventaire de constructions proposé par la BCG est redondant car l'information formelle ou sémantique n'est stockée qu'une seule fois. Il est également restrictif car n'y sont inclus que les schémas généraux et productifs tels que par exemple *let alone* (Fillmore, Kay et O'Connor, 1988), *WXDY* (Kay et Fillmore, 1999) ou les clivées en *all* (Kay, 2013).

2.3.2 La Grammaire de Constructions Cognitive

Le second courant est désormais connu sous le nom de Grammaire de Constructions Cognitive (GCC). Lorsqu'elle en pose les bases, Adele Goldberg fait aussi appel à la générativité : « Construction Grammar is generative in the sense that it tries to account for the infinite number of expressions that are allowed by the grammar while attempting to account for the fact that an infinite number of expressions are ruled out or disallowed » (1995, p. 7). Cette position a été remise en cause par Langacker (2009a, p. 169) au motif qu'elle va à l'encontre du principe selon lequel une construction est ratifiée dans et par l'usage. Goldberg s'est depuis distanciée de l'idée de générativité au profit du principe dit de « motivation maximisée » en vertu de laquelle les langues tendent à exploiter au maximum les correspondances sémantiques entre constructions formellement semblables (Goldberg, 2006, p. 218). L'idée force de la GCC est que la grammaire est un inventaire mental structuré, qu'Adele Goldberg nomme en anglais *constructicon*, mot-valise formé à partir de construction et lexicon.

À la différence de l'inventaire proposé en BCG, celui de la GCC est exhaustif et redondant [8]. L'exhaustivité de l'inventaire est résumé par la formule suivante : « it's constructions all the way down » (GOLDBERG, 2006, p. 18). La différence entre l'inventaire de la GCC et celui de la BCG est visible en comparant (7) et (8).

(7) Plus on est de fous, plus on rit.
   a. ?Plus nous sommes de fous, plus nous rions.
   b. ?Plus on est d'insensés, plus on rit.

(8) <plus/moins + PROP, plus/moins + PROP>
   a. Plus on lit, plus on s'instruit.
   b. Plus on lit de livres, moins on pense savoir de choses.

L'exemple (7) illustre une construction conditionnelle corrélative dans laquelle le premier segment (« plus on est de fous ») contient la condition de réalisation du deuxième segment (« plus on rit »). Le même phénomène est à l'œuvre en (8). Toutefois, à la différence de (8), la construction illustrée en (7) n'est pas productive. Alors qu'en (8) la construction admet un très grand nombre de propositions, il est impossible d'altérer quoi que ce soit en (7), au risque d'obtenir un énoncé sémantiquement et pragmatiquement incongru, comme illustré en (7a) et en (7b). Alors qu'en GCC les deux constructions ont leur place dans le *constructicon*, seule la construction illustrée en (8) a le statut de construction dans l'inventaire de la BCG.

Toutes les parties du discours peuvent ainsi prétendre au statut de construction dès lors qu'une partie au moins de l'assemblage forme/sens n'est pas prédictible à partir d'autres constructions :

> Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. (GOLDBERG, 2003, p. 219).

Figurent ainsi dans le *constructicon* des morphèmes, des lexèmes, des syntagmes, des interjections que des séquences syntaxiques partiellement lexicalisées, voire des schémas syntaxiques généraux (GOLDBERG, 2003 ; GOLDBERG, 2009).

2.3.3 La Grammaire Cognitive

La troisième approche est la Grammaire Cognitive. Ronald Langacker fait très tôt mention du concept de construction :

> Grammar involves the syntagmatic combination of morphemes and larger expressions to form progressively more elaborate symbolic structures. These structures are called grammatical constructions. (LANGACKER, 1987, p. 82)

La Grammaire Cognitive part du principe que la grammaire est un ensemble d'unités symboliques à des degrés divers de complexité (ce que les approches constructionnelles appellent

8. Il n'en a pas toujours été ainsi. Tel qu'il est décrit par GOLDBERG (1995), l'inventaire est non-redondant. Par ailleurs, les constructions qui y figurent ne sont pas compositionnelles. Tel qu'il est décrit par GOLDBERG (2006), l'inventaire devient redondant (l'information peut être dédoublée). Par ailleurs, les constructions qui y figurent peuvent être compositionnelles ou non.

des constructions). Trois types de structures sont nécessaires : les structures sémantiques, les structures phonologiques et les structures symboliques. Les unités linguistiques sont des structures symboliques. Une structure symbolique est bipolaire. C'est l'assemblage de deux pôles, à savoir un pôle phonologique et un pôle sémantique (LANGACKER, 1987, p. 76). Par exemple, le lexeme *car* est l'assemblage symbolique des propriétés conceptuelles formant le pôle sémantique ([CAR]) et du pôle phonologique (/kɑr/). L'assemblage est noté ainsi : [[CAR]/[kɑr]]. Le pôle phonologique est suffisamment large pour englober également une composante gestuelle.

En apparence, on est proche ici du signe saussurien, ce qui vaut aux grammaires de constructions d'être qualifiées d'approches néo-saussuriennes. Dans les faits, on est loin d'une conception structurale du sens. Celui-ci est résolument encyclopédique. Par exemple, le concept ne se réduit pas à une combinaison de traits tels que $[+véhicule]$, $[+moteur]$, etc. Il inclut le cadrage subjectif qui peut en être fait (la voiture comme mode de transport par opposition à la voiture comme signe de richesse). Deux conséquences importantes se dégagent de ce qui précède. Premièrement, le lexique et les règles forment un continuum. Pour reprendre les termes d'Adele Goldberg, « it's constructions all the way down ». Deuxièmement, toute la grammaire est porteuse de sens, y compris les structures symboliques les plus abstraites.

Langacker a mis du temps pour formaliser la Grammaire Cognitive en tant que grammaire de constructions. Il faut attendre le milieu des années 2000 et l'essor confirmé des grammaires de constructions pour que Langacker développe, précise et formalise ce qu'il entend par « construction » (LANGACKER, 2005 ; LANGACKER, 2008 ; LANGACKER, 2009b) au gré, parfois, de malentendus (LANGACKER, 2009a). Ayant soutenu ma thèse juste avant ces développements, ma compréhension des grammaires de constructions a été fortement colorée par la Grammaire Cognitive des origines. J'ai été particulièrement sensible à la caractérisation des constructions comme vecteurs d'interprétation (*construal)*, comme je vais l'illustrer plus bas (Section 2.4.1).

En Grammaire Cognitive, le sens d'une unité linguistique comprend tout à la fois un contenu conceptuel et la manière dont ce contenu conceptuel s'interprète (LANGACKER, 2008, p. 44). Le contenu conceptuel est un domaine, c'est-à-dire un ensemble cohérent de connaissances qui sert de point d'appui dans l'interprétation d'autres unités conceptuelles. Par exemple, des adjectifs tels que *moite*, *mouillé*, *détrempé*, *imbibé*, *imprégné*, etc. renvoient à des unités conceptuelles interprétables à l'aune du domaine HUMIDITÉ. Le contenu de ce domaine peut-être à son tour interprété à l'aune du domaine fondamental PROPRIÉTÉ. L'interprétation renvoie à la manière par laquelle un locuteur présente un contenu conceptuel spécifique à travers le choix d'une expression linguistique. S'il est vrai que les lexèmes ont un sens conventionnel, celui-ci est modifié par le contexte dans lequel l'unité linguistique est employée. Loin d'être un trait figé et assigné une fois pour toutes, le sens est négocié localement et socialement. Lorsque le locuteur adopte un point de vue spécifique, il influence la réception de la représentation conceptuelle par l'interlocuteur. L'interprétation est également au cœur de la GCC décrite plus haut. Une construction est donc tout à la fois un produit et un vecteur de conceptualisation. Par conséquent, deux formes linguistiques similaires ne présupposent pas la même façon de percevoir un événement, tant chez le locuteur que chez l'interlocuteur. Une construction peut être vue comme une grille de lecture d'expériences linguistiques et sociales, c'est-à-dire le produit de notre activité perceptuelle. À ce titre, la polysémie du

terme « construction » en anglais (« construction »/« interprétation ») est révélatrice du lien entre activité langagière et activité cognitive.

2.3.4 La Grammaire de Constructions Radicale

CROFT (2001) a formalisé la Grammaire de Constructions Radicale pour remettre en cause la thèse de l'universalité des catégories grammaticales. Il part du principe que les outils typologiques traditionnels tels que les catégories grammaticales (nom, verbe, adjectif, adverbe, etc.), et les fonctions (sujet, verbe, complément, etc.) et les schémas syntaxiques (transitif, intransitif, attribut, etc.) ne s'exportent pas aisément d'une famille de langues à l'autre. Abandonnant le projet d'une typologie universelle des unités linguistiques, Croft considère que la construction est la seule unité minimale descriptive fiable pour décrire la diversité des langues. De nature endogène, elle permet de s'abstraire d'outils descriptifs exogènes, ceux-là mêmes qui font courir le risque de fausser l'interprétation des phénomènes linguistiques.

La Grammaire de Constructions Radicale est une approche à part, j'en ai apprécié la formalisation sans pour autant pouvoir l'appliquer, mes recherches ne portant pas sur la typologie linguistique. Il n'en demeure pas moins que j'en ai retenu qu'une fois encore la grammaire de constructions se pose comme une grille d'analyse plutôt qu'une théorie.

2.3.5 Les autres approches

L'inventaire des grammaires de constructions ne saurait se réduire aux quatre courants mis en avant par CROFT et CRUSE (2004). D'autres approches existent :

— Embodied Construction Grammar ;

— Fluid Construction Grammar ;

— Sign-Based Construction Grammar

La Grammaire de Constructions Incarnée (BERGEN et CHANG, 2005) est très proche de la GCC et de la Grammaire Cognitive dans ses objectifs, mais s'engage dans une perspective expérimentale, computationnelle et prédictive vis-à-vis de la description des constructions. Elle considère que chaque construction est une hypothèse à valider par l'observation de ses réalisations dans des conditions naturelles et expérimentales. La Grammaire de Constructions Fluide est une application computationnelle des grammaires de constructions développée depuis les années 2000 (STEELS, 2004 ; STEELS, 2011 ; STEELS, 2012). Elle est associée principalement à la communication inter-agents.

La Grammaire de Constructions Fondée sur les Signes (SAG, 2008) se situe dans le prolongement de la BCG et est fortement influencée par l'approche HPSG. Rejetant le modularisme et la dérivation, elle postule que la grammaire est un inventaire de signes. Ces signes expriment des contraintes quant à la forme, le sens et l'usage. Dans cette approche, les constructions rendent compte de la combinaison de signes plus simples en signes plus complexes.

J'ai découvert ces trois approches lors de la troisième International Conference on Construction Grammar (ICCG3) à Marseille en 2004. Je cherchais alors une approche me permettant de modéliser les mécanismes interprétatifs à l'œuvre dans les actes de langage directifs.

L'accent mis sur le formalisme des trois approches que je viens de décrire m'en a éloigné au profit de la CCG et de la Grammaire Cognitive. Mes projets plus récents sur la détection semi-supervisée de réseaux de constructions (voir Chapitre 5) me fait revenir vers ces approches précisément pour ce qui m'en détachait auparavant : leur formalisme poussé.

2.4 Une approche intégrative de problèmes anciens

Si les grammaires de constructions s'organisent en un ensemble d'approches qui n'ont rien d'une théorie monolithique et unifiée, elles permettent toutefois de relire d'anciennes problématiques linguistiques avec un regard nouveau et souvent plus systématique que d'autres approches. Je présente ici deux études de manière chronologique : les constructions directives indirectes et la distinction massif-comptable.

2.4.1 Les constructions directives indirectes

La première étude est une exploration de la composante pragmatique des constructions. Je la considère comme étant le parent pauvre en grammaires de constructions, au profit de la sémantique (cf. Section 2.5.2 plus bas). Cette étude est représentative des recherches menées dans le sillage immédiat de ma thèse.

L'hypothèse du jeu constructionnel

Au sortir de ma thèse, encore fortement inspiré par la Grammaire Cognitive, j'ai cherché à développer le concept de *zone de développement potentiel* (DESAGULIER, 2007b ; DESAGULIER, 2008a), qui résulte de la concurrence synchronique entre plusieurs constructions pour une même famille de fonctions. J'ai décrit cette zone, rendue possible par un mécanisme de « jeu constructionnel », comme révélatrice d'une cognition sociale et comme moteur de la dynamique langagière, tant sur le plan synchronique que diachronique.

Le jeu constructionnel est tout d'abord lié au fait que chaque unité symbolique est potentiellement réanalysable, sans qu'il soit pour autant possible de prévoir exactement ce à quoi ressemble un assemblage émergent. D'autre part, il existe, pour certaines constructions, un degré d'incertitude chez les locuteurs quant à l'assemblage précis qui les caractérise, soit parce que ses constituants sont multifonctionnels (donc ambigus), soit parce que l'unité n'est pas suffisamment conventionnelle. Il ne suffit pas de reconnaître ce jeu : il faut lui accorder une place centrale dans toute entreprise typologique.

L'interaction langagière et les formes linguistiques qui y participent ne sont pas de l'ordre du donné, en dépit de leur nature largement conventionnelle. Les constructions les plus anodines recèlent des strates de sens sédimentées au cours de leur histoire (certaines plus accessibles que d'autres) qu'il est possible de réactiver synchroniquement au gré des attentes

sociales souvent conflictuelles qui conditionnent toute prise de parole. Ce décalage entre usage conventionnel et appropriation individuelle dans la synchronie crée un espace dans lequel les identités s'affirment, s'effacent, ou se (re)négocient.

Les constructions directives indirectes

Les constructions directives indirectes sont particulièrement éclairantes sur le jeu constructionnel et la subjectivité qui s'y déploie. Elles s'emploient notamment lorsqu'un locuteur pousse l'interlocuteur à agir sans pour autant donner l'impression d'empiéter sur son libre arbitre. Parmi les stratégies possibles, la plus subtile consiste à réunir dans une même construction hybride le marquage implicite d'une suggestion, d'un souhait ou d'un ordre émis par le locuteur et le codage explicite de la liberté de l'interlocuteur d'accéder ou non à cette demande. Cette stratégie est visible dans la construction anglaise *want to/wanna* en contextes directifs :

(9) You don't want to appear brash or pushy.
« Il faut éviter de se montrer impertinent ou arrogant. »

(10) You wanna be careful !
« Fais attention à toi ! » (= « tu as intérêt à faire attention à toi ! »)

Dans ce type de construction, l'acte illocutoire est le suivant : le locuteur L décrit la nécessité prétendument objective pour l'interlocuteur I d'accomplir l'action X, ce que traduit la séquence syntaxique <*you want X*>. Cet acte ne correspond toutefois pas exactement à la force illocutoire : L veut que I accomplisse X. Le locuteur fait ici semblant de reconnaître à l'interlocuteur la liberté d'accepter ou de refuser d'agir, d'où le choix du verbe *want*, qui est d'autant plus logique qu'il est naturel de vouloir ce qui est dans son propre intérêt. Mais le dépositaire ultime de l'autorité déontique est bel et bien le locuteur. Nous sommes donc en présence d'un acte de langage indirect. La construction déontique *want to/wanna* permet de brouiller les pistes quant à la source de contrainte : le locuteur, qui émet le jugement déontique, n'apparaît pas dans la syntaxe. Ce jeu sur l'identité perdrait toute sa subtilité si, au lieu d'adopter une stratégie compressive (c'est-à-dire sans mention du locuteur), on adoptait une stratégie expansive du type <*I want you to X*>, avec mention explicite de la source de contrainte (DESAGULIER, 2005).

La mitigation : une profusion de dénominations et de marqueurs

En intégrant le Centre de Linguistique Théorique (EHESS), j'ai trouvé un contexte favorable pour affiner les phénomènes pragmatiques à l'œuvre lorsqu'un locuteur cherche à atténuer l'impact d'un acte de langage directif. J'étais loin de me rendre compte que j'ouvrais alors une boîte de Pandore tant ce que j'ai réuni autour du concept de « mitigation », est un phénomène fuyant. Sa nature protéiforme a entraîné une profusion de dénominations. J'ai cherché à plusieurs reprises à en dégager le soubassement cognitif. Je résume ci-dessous le résultat de ces recherches, présentées lors de deux colloques (DESAGULIER, 2007a ; DESAGULIER, 2008b).

Les premiers travaux sur la mitigation remontent à LAKOFF (1973), qui étudie le phénomène au niveau de l'atténuation de l'assertion propositionnelle via le phénomène de *hedging*, comme en (11), où *a sort of* présente la prédication entre *penguin* et la catégorie *bird* sur le mode du quasiment vrai :

(11) A penguin is **a sort of** a bird.

FRASER (1975) étend le champ de la mitigation à l'étude des énoncés performatifs et à la force illocutoire. Dès lors, la mitigation est très fortement reliée à la caractérisation des actes de langage potentiellement conflictuels (FRASER, 1980 ; FRASER, 1990 ; BROWN et LEVINSON, 1987). Selon FRASER (1980, p. 341), la mitigation n'est pas un acte de langage en soi, mais une modification de celui-ci. Plus précisément, c'est un dispositif linguistique visant à atténuer l'impact négatif d'un acte de langage sur autrui. Des travaux plus récents reviennent sur cette intuition originale pour l'affiner (SBISÀ, 2001 ; THALER, 2012).

La mitigation apparaît sous plusieurs dénominations : *downgrading* « rétrogradation » ou « déclassement » (HOUSE et CASPER, 1981), *weakening* « affaiblissement » (BROWN et LEVINSON, 1987), *hedging* « modalisation » (HOLMES, 1995), *attenuating* « atténuation » (LEECH, 1983 ; HOLMES, 1984) ou *adoucissement* (KERBRAT-ORECCHIONI, 2001). On ne peut s'empêcher de penser à la réflexion suivante de BOLINGER (1981, p. 554) :

> One sign of immaturity [in a science] is the endless flow of terminology. The critical reader begins to wonder if some strange naming taboo attaches to the terms that a linguist uses, whereby when he dies they must be buried with him [9].

Ainsi, l'abondance terminologique qui frappe les travaux sur la mitigation est peut-être le signe que ce phénomène n'est pas encore bien cerné.

La recherche contemporaine, dont on trouvera un aperçu chez SCHNEIDER (2010), étudie la mitigation sous trois angles distincts : (a) les contextes motivant le recours à la mitigation, (b) les schémas linguistiques associés à la mitigation et (c) les effets des schémas de mitigation sur l'interaction. Depuis plus récemment, la mitigation est étudiée simultanément sous ces trois angles, suggérant qu'il s'agit d'un objet d'étude multidimensionnel (CAFFI, 2007 ; CZERWIONKA, 2012). En dépit de la démultiplication des dénominations et des questionnements, la difficulté n'est pas de définir la mitigation, mais d'en circonscrire les manifestations linguistiques et de les relier parallèlement au facteurs méta-, para- et extralinguistiques.

Les typologies de marqueurs abondent dans les publications sur la mitigation. Ces typologies ne se recoupent que partiellement. Parmi les publications les plus récentes, on trouve : CZERWIONKA (2012), ITAKURA (2013) et YANG (2013) et FLORES-FERRÁN et LOVEJOY (2015) [10]. Il n'est rien d'étonnant à cela dans la mesure où ces typologies se caractérisent par leur nature heuristique et non-limitative. Cette profusion typologique est à la mesure de l'ampleur du phénomène et à l'hétérogénéité des stratégies qu'il mobilise. Ces stratégies s'articulent notamment sur l'idée d'un évitement, d'une "communication indirecte" (HENGEVELD et KEIZER, 2011), d'une préparation ou d'une compensation. Selon FRASER (2010), l'échec de toute

9. Cité par CRYSTAL et SEARS (2008, p. vi).
10. Pour une tentative de typologie des typologies, voir FRASER (2010, p. 21).

délimitation exhaustive du phénomène de mitigation est dû au fait qu'aucun marqueur n'est intrinsèquement un marqueur de mitigation. Il s'agit avant tout d'une stratégie rhétorique générale qui recrute de manière *ad hoc* les marqueurs dont elle a besoin.

Plus systématique est l'approche de CAFFI (1999) et CAFFI (2007), qui distingue trois types de mitigateurs, définis en fonction de leur portée : les *bushes*, les *hedges* et les *shields*, illustrés en (12–14) respectivement :

(12) I'm going to give you a **little** shot to kill the pain. (COCA Bk :DoublePreyPosadas)

(13) **If you like**, we'll turn around your future. (COCA FantasySciFi)

(14) Well, **it says here that** you've never ridden before. (COCA Bk :RopedIn)

Ces trois étiquettes reposent sur un jeu de mot : les *bushes* ont un rôle dissimulateur (à l'image du buisson qui cache la forêt), les *hedges* jouent le rôle d'obstacle, et les *shields* ont un rôle protecteur. Les *bushes* atténuent la prise en charge du locuteur vis-à-vis du contenu propositionnel de son énoncé via une intervention sur sa valeur de vérité. Les *hedges* atténuent la prise en charge du locuteur vis-à-vis de la force illocutoire de son énoncé. Enfin, les *shields* portent sur l'origine déictique de l'énoncé, renvoyant la source énonciative en dehors de la triade *ego–hic–nunc*.

Les enjeux de la mitigation

Dès lors que l'on corrèle un acte illocutoire à sa dimension locutoire se pose la question de la correspondance entre forme linguistique et fonction. LEVINSON (1983, p. 264) aborde cette question dans le cadre de la distinction entre actes de langage direct et indirect. Ce qu'il nomme « l'hypothèse de la force littérale » consiste à poser une corrélation forme/fonction directe entre les actes de langage et leurs réalisations phrastiques. D'un point de vue structuraliste, aux trois actes de langage élémentaires (déclarer, ordonner, interroger) correspondent bien trois types de phrases reconnues dans la plupart des langues du monde, à savoir les formes déclarative, impérative et interrogative (SADOCK et ZWICKY, 1985, p. 160) :

(15) *Linda is reading a book.* → syntaxe déclarative – assertion

(16) *Is Linda reading a book ?* → syntaxe interrogative – question

(17) *Read a book !* → syntaxe impérative – ordre

Or, ces cas de correspondance entre forme et fonction ou plus précisément entre locution et illocution sont protypiques mais pas systématiques, voire moins fréquents qu'une situation de décalage. On sait par exemple qu'une ordre ou une requête peuvent prendre la forme d'une question :

(18) Why don't you read a book ?

Ce constat est étayé par des recherches expérimentales qui montrent que les locuteurs ont tendance à laisser de côté le sens littéral de requêtes indirectes conventionnelles au profit d'une interprétation non-littérale, y compris lorsque ces requêtes sont entendues dans des contextes invitant une interprétation littérale (GIBBS, 1983 ; PAPAFRAGOU, 2000).

Il n'en demeure pas moins que si l'interprétation littérale de la composante locutoire est rare, on est en droit de penser que la composante littérale influence le cheminement du mécanisme inférentiel. Ainsi, sachant que la mitigation, en tant que stratégie énonciative indirecte, se caractérise par un surcroît de forme linguistique, plus cette forme est conséquente, plus l'acte illocutoire est indirect et plus la mitigation est forte. C'est la logique sous-jacente au gradient de médiation (« indirectness scale ») à l'œuvre entre un acte illocutoire et son accomplissement chez LEECH (1983, p. 123–124). Ainsi, la complexité du cheminement inférentiel serait à la mesure de la complexité formelle du marqueur de mitigation. En poussant ce raisonnement à son terme, on voit poindre la possibilité de mesurer objectivement la nature indirecte des manifestations de mitigation.

Cet idéal de mesure objective est présent dans les études sur la politesse. BROWN et LEVINSON (1987) proposent une formule censée mesurer le poids W d'un acte de langage menaçant (« face threatening act ») :

$$W_x = D(S, H) + P(H, S) + R_x, \tag{2.1}$$

où x est l'acte de langage menaçant, D la distance sociale entre les interlocuteurs, P le pouvoir relatif que l'interlocuteur H a sur le locuteur S et R le degré d'imposition de l'acte de langage. Toutefois, cette formule n'est qu'une abstraction posant autant de problèmes qu'elle entend en résoudre dans la mesure où il est tout aussi illusoire de penser pouvoir mesurer D, P et R.

Si la mesure objective de la mitigation est une tâche difficilement réalisable, cerner le phénomène sur la base de sa dimension locutoire n'est toutefois pas impossible dans la mesure où l'on perçoit intuitivement que, dans la pratique, la mitigation se situe entre un défaut de forme, comme en (19), et un excès de forme, comme en (20) :

(19) Don't answer your phone in class.

(20) Would you think it an imposition on my part if I were to ask you not to answer your mobile phone in class ?

Sur la base de ce qui précède, j'ai interrogé le ratio entre quantité de forme et acte illocutoire, sans laisser de côté la nature qualitative des marqueurs de mitigation. Pour cela, j'ai choisi d'étudier le *hedging* caténatif.

Le cas du *hedging* caténatif

Le *hedging* caténatif est une forme de mitigation reposant sur au moins deux *hedges*. Pour en cerner les enjeux, commençons par en décrire un usage en contexte. Improv Everywhere est

une troupe d'acteurs new yorkais spécialisée dans les improvisations humoristiques en lieu public grâce à l'appel à des volontaires recrutés sur internet. Chacune des missions est filmée en caméra cachée, afin de capter la réaction des victimes et des badauds. La mission « Best Buy » a eu lieu le 23 avril 2006 [11]. Son principe est le suivant : 80 participants reçoivent la consigne de se présenter au magasin d'électronique de la firme Best Buy sur la 23^{e} rue à Manhanttan à 15h30. Ils doivent être vêtus d'un pantalon de toile beige doté d'une ceinture, d'un polo bleu roi et de chaussures noires. Ce code vestimentaire est la réplique exacte de celui des employés du magasin. À l'heure dite, les 80 participants entrent l'un après l'autre dans le magasin. Il leur est demandé de ne pas interagir et de ne pas se faire passer pour des employés du magasin. Cette mission frappe par son caractère incongru et émergent : ni les participants ni leurs victimes n'ont une idée de la finalité de l'événement. Gênés par l'intrusion, qui intrigue les clients, les gérants du magasin font appel aux forces de l'ordre qui rétorquent que rien ne peut être retenu contre les participants. Les gérants se trouvent dès lors face à un dilemme : convaincre les membres de la troupe de quitter le magasin sans motif valable. S'ensuit l'échange suivant, retranscrit par l'un des participants de la mission :

(21) Gérant – **I'm going to have to ask you to leave.**
Participant – You're kicking me out ?
G – No, I'm not saying that.
P – Ok, then I can stay ? (. . .)
G – **I'm asking you to leave.**
P – Are you kicking me out ?
G – No.

La transcription du dialogue est certes tronquée mais elle a le mérite de mettre en avant les stratégies discursives des deux interlocuteurs. Ce dernier encadre sa première requête *I . . . ask you to leave* de deux *hedges* : *going to* et *have to*. Cette double mitigation est indéniablement indirecte. Elle est immédiatement suivie par une requête d'explicitation de la part du participant : *You're kicking me out ?*. De manière à ne pas enfreindre les lois anti-discrimination en vigueur aux États-Unis, le gérant réfute l'interprétation que le participant souhaite lui faire expliciter : *No, I'm not saying that*. Souhaitant poursuivre ce jeu de dupes, le participant provoque une seconde requête d'éviction de la part du gérant : *I'm asking you to leave*. Les deux *hedges* utilisées lors de la première tentative sont absentes de la seconde, laissant penser que cette dernière est plus directe et donc plus franche. En dépit de cet ajustement, le participant campe sur sa position et poursuit sa stratégie d'explicitation, au grand dam du gérant.

L'emploi directif à l'œuvre dans la construction caténative *I am going to have to ask you to X* (littéralement : « je vais être dans l'obligation de vous demander de X ») amène à considérer l'acte illocutoire (le locuteur L veut que l'interlocuteur I accomplisse X) comme imminent (ce que traduit *going to*) [12] et indépendant de la volonté du locuteur (ce que traduit *have to*). Cette stratégie composite se trouve mise en place au sein d'une seule et même construction caractérisée par un haut degré de coalescence morphosyntaxique. En d'autres termes, elle

11. `http://improveverywhere.com/2006/04/23/best-buy/`
12. Si l'imminence est à l'œuvre ici, on ne peut toutefois pas l'associer à *going to* en général. Avec *going to*, le futur est présenté comme entièrement déterminé par la situation présente.

a valeur d'unité dans l'esprit de qui l'emploie, comme l'atteste sa fréquence d'occurrence en corpus. Elle n'est pas pour autant figée. Une fois amputée d'un semi-auxiliaire (*have to* ou *going to*), la demande se fait plus marquée. Ôter une ou plusieurs de ces formes non finies revient à augmenter proportionnellement le degré d'imposition de la construction. Tout se passe comme si la distance du locuteur vis-à-vis de l'acte de langage était codée iconiquement. Ainsi, en théorie du moins, plus la chaîne de marqueurs de forme non finie est longue, moins l'acte de langage est contraignant car ce sont tout autant de barrières qui font obstacle à la réalisation de l'événement et libèrent une marge de manœuvre pour l'interlocuteur.

Ce travail partiellement conscient sur les assemblages de forme et de sens tend à prouver qu'une construction porte en elle bien plus que des règles abstraites visant à la bonne formation des énoncés. Elle sous-tend aussi un processus complexe d'émergence de la subjectivité.

La mitigation *via* le *hedging* caténatif décrite ci-dessus se situe au carrefour de trois interfaces : linguistique, métalinguistique et cognitive. L'interface linguistique a déjà été abordée à la Section 2.4.1 et peut être résumée de la manière suivante : plus la construction comporte de forme linguistique, plus la mitigation est marquée. Il y aurait ainsi une gradation croissante de la mitigation de la requête depuis (22a) jusqu'à (22d) :

(22) a. Leave !
b. I am asking you to leave.
c. I have to ask you to leave.
d. I am going to have to ask you to leave.

La composante métalinguistique est décrite par LEECH (1983, p. 139–142) en référence au ratio coût/bénéfice qu'implique un acte de langage. L'auteur d'une requête cherche généralement à obtenir la coopération de l'interlocuteur en fournissant des précisions sur les motivations de ladite requête ou sur les bénéfices que l'interlocuteur peut tirer de son exécution. Lorsque la requête est perçue par le locuteur comme un empiètement sur la sphère privée de l'interlocuteur, le premier peut assortir son illocution d'une description métalinguistique qui a valeur de *hedge*. En (21), deux stratégies sont à l'œuvre, illustrant une tendance déjà observée par LEECH (1983, p. 140) : « Bringers of bad tidings may find it advisable to express both the distasteful and the unavoidable nature of their task (. . .). ». Cette stratégie est rendue d'autant plus nécessaire lorsque l'interlocuteur est un inconnu. En (22b), l'objectif du locuteur (*you leave*) est présenté métalinguistiquement sous l'angle de la requête (*I am asking*). En (22c), ce même objectif est présenté métalinguistiquement à la fois sous l'angle de la requête et celui de l'obligation externe (*have to*). Enfin, en (22d), la requête est présentée en plus comme étant imminente (*going to*). En somme, la composante métalinguistique rejoint la composante linguistique en indexant la quantité forme sur la recherche d'une oblicité de la mitigation.

Les interfaces linguistique et métalinguistique ont un corrélat cognitif. En adoptant le formalisme de la Grammaire Cognitive (LANGACKER, 1999), on peut représenter schématiquement en Figure 2.2 la construction *<I/ask you to leave>* illustrée en (22b). Dans ce schéma

complexe, l'ellipse de gauche représente la composante locutoire. Le locuteur exerce une force modale subjectivisée sur l'allocutaire. Cette force modale est représentée par une double flèche en pointillés. La composante locutoire renvoie à un acte de langage accompli sur le plan du réel. C'est à l'allocutaire qu'il revient d'accomplir l'action dénotée par le verbe complément *leave* (flèche ondulée). Il s'agit de la composante illocutoire de la construction, représentée dans le rectangle de droite. Celle-ci projette l'allocutaire sur le plan de l'irréel.

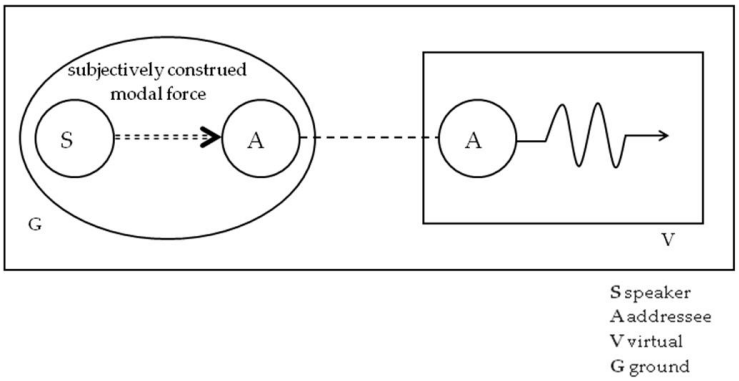


Figure 2.2: Soubassement cognitif du schéma constructionnel <*I/ask you to X*>

La Figure 2.3 représente le *hedging* marqué par *I have to* en (22c). Sur le plan du réel, La force modale d'origine externe est subjectivisée par le locuteur qui la conçoit comme applicable à lui-même. La Figure 2.4 représente la conceptualisation subjective du défilement chronologique marqué par *be going to* en (22d) [13]. La combinaison de ces marqueurs caténatifs est visible en Figure 2.5.

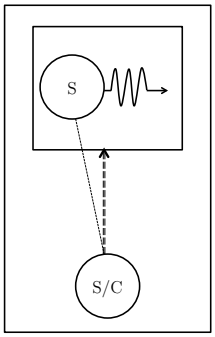


Figure 2.3: Soubassement cognitif du schéma constructionnel <*I have to X*>

Deux logiques sont à l'œuvre ici. D'un côté, *I am going to have to ask you to X* est une construction conventionnelle, presque figée, puisque employée sous cette forme exacte dans un contexte bien identifié : un locuteur dépositaire d'une autorité (plus ou moins limitée) invite l'interlocuteur à faire une action dont il n'est pas à l'origine sur la base d'une contrainte externe. Symboliquement, cette construction signale à la fois un contenu propositionnel, une intention pragmatique et une posture subjective au croisement de la linguistique, de la métalinguistique et de la cognition. D'un autre côté, cette construction s'adapte au degré de mitigation souhaité. L'exemple (22b), schématisé par la Figure 2.6, montre une proximité plus grande entre le locuteur et l'acte attendu, d'où une mitigation plus faible.

13. La projection dans le futur n'est pas effective mais subjectivisée par le locuteur, ce que dénotent les flèches en pointillés.

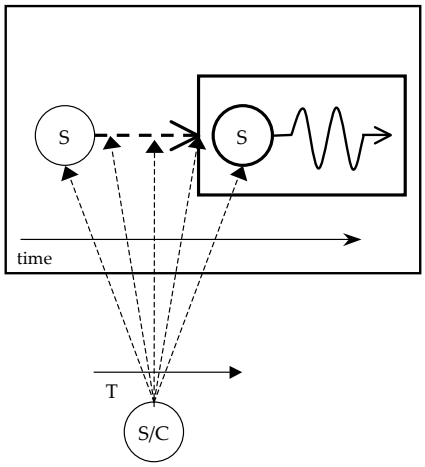


Figure 2.4: Soubassement cognitif du schéma constructionnel *<I am going to X>*

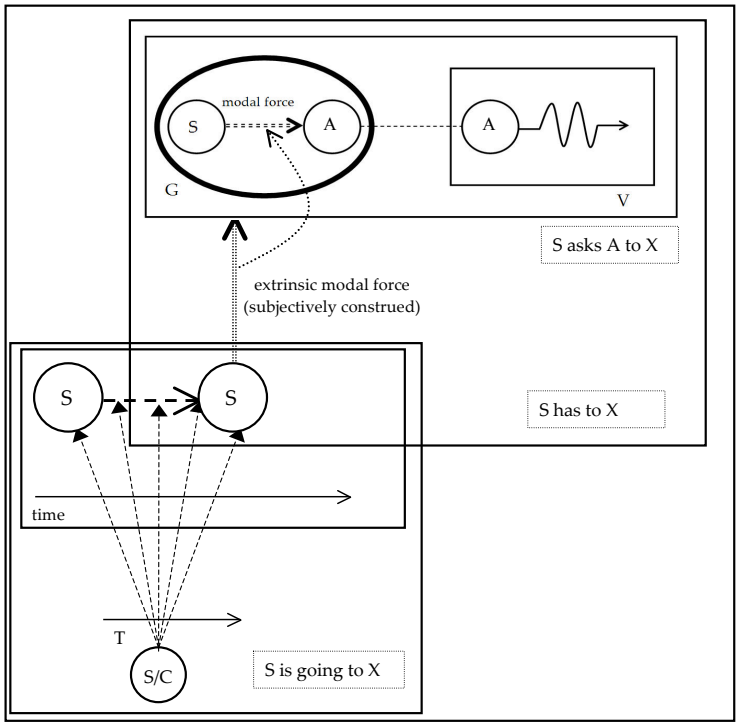


Figure 2.5: Soubassement cognitif du schéma constructionnel *<I am going to have to ask you to X>*

La géométrie variable de cette construction directive indirecte est l'effet d'un ajustement intersubjectif. Plutôt que d'assigner à chaque construction une dénomination, au risque de tomber dans une démultiplication terminologique confuse, il semble plus productif de partir d'un patron générique qui s'adapte aux coordonnées de l'interaction, ces coordonnées pouvant changer au sein d'une même séquence dialogique.

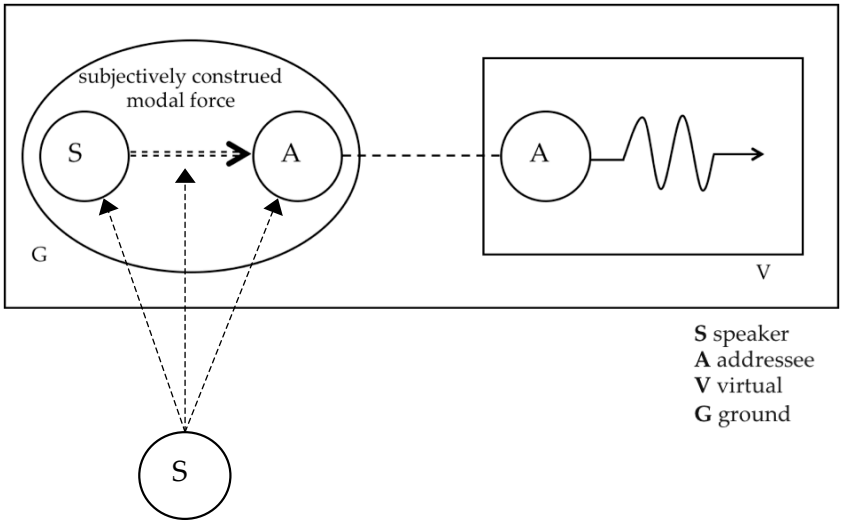


Figure 2.6: Soubassement cognitif du schéma constructionnel *<I am asking you to X>*

2.4.2 La distinction massif-comptable

La seconde étude porte sur l'intensification au sens large. Elle est représentative de mes recherches plus récentes. Dans un volume collectif consacré à une approche constructionnelle du français, j'aborde la conversion des noms dénombrables en noms massifs en français contemporain DESAGULIER (2012a). Il s'agit d'un point récurrent de la linguistique du français. DAMOURETTE et PICHON (1930, p. 414) écrivent à son sujet :

> Ce qui fait la beauté du système de quantitude de la langue française de nos jours, c'est son extrême souplesse et son absolue généralité. C'est après avoir bien médité et posé la question, et après avoir observé la langue tant parlée qu'écrite, que nous osons affirmer ici qu'il n'est pas de substance nominale qu'il soit interdit à un locuteur plus ou moins hardi de concevoir soit comme numérative, soit comme massive.

Si la distinction massif/comptable est tributaire de l'arbitraire du locuteur, l'intérêt de cette étude de cas est la difficulté qu'éprouve la linguistique traditionnelle pour la formaliser, l'expliquer et la prédire. Parce que la linguistique de l'usage pose cet arbitraire au cœur de ses prémisses, le défi est intéressant à relever. C'est ce que j'ai cherché à faire.

La distinction entre noms dénombrables (*un lapin*, *un chêne*) et noms massifs (*du lapin*, *du chêne*) a souvent été décrite en termes ontologiques pour rendre compte du fait que certains noms ont un fonctionnement prioritairement comptable (animal, arbre) et d'autres un fonctionnement prioritairement massif (viande, bois). QUINE (1960), CHENG (1973) et TER MEULEN (1981) et BUNT (1985) ont mis en avant des critères ontologiques visant à expliquer le fonctionnement des noms par la caractérisation de l'essence même du référent nominal. Ce type d'approche pose des problèmes d'un point de vue centré sur l'usage, car il n'est pas dit que les locuteurs aient accès à ces connaissances ontologiques. De plus, ces principes peuvent difficilement prévoir le fonctionnement d'un nom en contexte.

En philosophie du langage, on fait appel à deux machines hypothétiques pour rendre compte de la conversion du comptable au massif : le « broyeur universel » et le « multiplicateur universel ». Ces machines sont des méthodes de pensée permettant d'expliquer respectivement le passage de *un lapin* (animal dénombrable) à *du lapin* (viande « broyée ») et, dans un registre plus familier, de *une meuf* (individu dénombrable) à *de la meuf* (individu démultiplié) (PELLETIER, 1975). GALMICHE (1989) a montré qu'aucune de ces machines n'était universelle. La linguistique traditionnelle reconnaît l'influence du contexte et le besoin de se distancier des propriétés intrinsèques du GN mais ne propose pas de véritable formalisation (GALMICHE, 1989 ; KLEIBER, 1997 ; NICOLAS, 2002).

La Berkeley Construction Grammar emploie le terme de « coercition » (*coercion*) en lieu et place de « conversion » (DE SWART, 1998). Ce concept rend compte du phénomène de « glissement de type » (*type shifting*) à l'œuvre dans les conversions du dénombrable au massif et inversement. Ainsi, l'exemple (23) serait un cas de coercition de massif en dénombrable et (24) de dénombrable en massif (FILLMORE et KAY, 1995).

(23) Give me a tea.

(24) Give me some blanket.

Selon MICHAELIS (2004), les deux exemples ci-dessus illustrent le phénomène de coercition exocentrique : un mot qui n'est pas une tête fonctionne sémantiquement comme une tête. En l'occurrence, le déterminant contraint les noms qu'il détermine.

Une autre approche, de type BCG et compatible avec la Sign-Based Construction Grammar, consiste à postuler des constructions dérivées et un mécanisme de domination pour modéliser la coercition nominale. Ainsi, en (23), la construction Dénombrable-vers-Massif autorise l'emploi d'un type de lexème nominal portant le trait $[borné -]$ (*tea*), celui-ci étant unifié avec le trait $[borné +]$ requis par la construction Détermination-Indéfinie. En (24), la construction Massif-vers-Dénombrable autorise l'emploi de *blanket*, prototypiquement $[borné +]$ en l'unifiant avec le trait $[borné -]$ requis par la construction Détermination-Partitive. La BCG ne postule pas l'existence d'un conflit entre ce que requiert la construction et les traits sémantiques des lexemes qui interviennent. Au lieu de cela, les constructions dérivées génèrent de nouveaux types de lexemes dont les valeurs $[borné \pm]$ se conforment au contexte coercitif. Ce contexte est l'article indéfini en (23) et l'article partitif en (24).

Tout en reconnaissant la valeur des explications précédentes, mon point de départ est différent. J'examine l'hypothèse de WEINREICH (1966) et K. ALLAN (1980) selon laquelle le trait $[\pm comptable]$ est assignable non pas au seul nom, mais à l'ensemble du groupe nominal. J'étends cette hypothèse au-delà du groupe nominal et prends en compte l'ensemble de la construction dans laquelle s'insère le groupe nominal. En français oral contemporain, il apparaît que lorsqu'un locuteur souhaite émettre un commentaire d'ordre qualitatif sur le référent d'un groupe nominal dans une construction copulative du type *<c'est du/de la* + N>, le nom aura un fonctionnement massif même s'il est prioritairement comptable en dehors de la construction.

(25) a. c'est une bagnole (identification)

b. ça c'est de la bagnole ! (conversion + commentaire qualitatif)

(26) a. ce film, c'est une bombe (prédication par comparaison)

b. ce film, c'est de la bombe ! (conversion + commentaire qualitatif)

Le trait $[+massif]$ est assigné à chaque fois par la construction et non par l'essence du référent nominal. Dans ce cas précis, l'approche constructionnelle a une valeur prédictive plus grande que l'approche ontologique. Mon approche trouve un écho favorable auprès de KLEIBER (2014) :

> Il faut donc prendre en considération le type d'emploi qui favorise tel ou tel type de transfert. M. Galmiche est conscient de cette nécessité, puisqu'il avertit dans sa conclusion que « l'interprétation ne doit pas être restreinte au seul SN » (1989 : 75), mais les exemples qu'il utilise, pour démontrer la possibilité ou l'impossibilité de tel ou tel transfert, ne sont principalement que des SN (cf. par exemple, p. 68 : **du kilo / *du litre / *du chapitre*, etc.), ce qui justifie la critique de G. Desagulier (2012 : 228) contre les « NP-centered approaches » et ouvre la voie à la linguistique de corpus et aux approches « constructionnelles », qui lient l'apparition d'un transfert à sa place dans un ou dans des types de construction.

Selon Georges Kleiber, les conversions ne peuvent se manifester qu'en « emploi » et l'interprétation de celles-ci ne doivent pas être limitées au syntagme nominal mais prendre en compte un contexte plus large. Le passage cité ci-dessus est encourageant quant à la pertinence des grammaires de constructions comme grille d'analyse et aux possibilités offertes par les méthodes de la linguistique fondée sur l'usage.

Sur la base de ce qui précède, j'ai décrit un réseau de trois constructions : la construction copulative sujet-prédicat (CSPC), la construction CCDN et la construction CDN. Elles sont illustrées respectivement en (27), (28) et (29).

(27) a. CSPC : [(ça) [c'est [un(e) (GA) N]$_{GN}$]$_S$]$_{(S)}$

b. (Ça), c'est une (belle) voiture.

(28) a. CCDN : [ça, [c'est [du/de la/de l' (GA) GN]$_{GN}$]$_S$]$_S$

b. Ça, c'est de la voiture !

(29) a. CDN : [([ce/cette/ces GN$_i$]$_{GN}$), [c'est [du/de la/de l' (GA) GN$_j$]$_{GN}$]$_S$]$_{(S)}$

b. (Cette voiture,) c'est de la bombe !

Il y a lieu de distinguer deux constructions proches dans le constructicon [14] lorsqu'elles ont à la fois un ou plusieurs points formels et/ou sémantiques en commun et d'autres caractéristiques distinctives. Les constructions CSPC et CCDN, dont la forme est rappelée en (27a) et (28a) ont des caractéristiques formelles communes :

— elles ont deux constituants principaux : la fusion du démonstratif et de la copule (*c'est)*) ainsi qu'un GN en position attribut ;

— elles s'appuient sur des pronoms anaphoriques (*c'* et de manière optionnelle *ça*) pour connecter le GN à un référent du contexte (extra-)linguistique par identification et prédication ;

14. c'est à dire reliées au même nœud

— elles acceptent une prédication interne au GN par le biais d'un adjectif ;

— elles supposent que le référent des pronoms démonstratifs est pragmatiquement accessible.

La construction CCDN se distingue toutefois par les points suivants :

— un article partitif détermine le GN attribut ;

— le GN attribut est interprété comme massif même s'il est prototypiquement dénombrable en dehors de la construction ;

— la conversion du comptable au massif invite une interprétation qualitative et non quantitative du GN.

La CCDN et la CDN partagent les traits suivants :

— une trame syntaxique héritée la CSPC, à savoir le schéma copulatif sujet-prédicat, un article partitif et un GN interprété comme massif ;

— l'interprétation massive du GN a une fonction qualitative (l'expression d'un haut degré dans un jugement de valeur).

Ces deux constructions sont toutefois distinctes. Si l'on insère un focus prosodique, celui-ci ne s'applique pas sur le même élément.

(30) ÇA, c'est de la voiture !

(31) Cette voiture, c'est de la BOMBE !

Dans la construction CCDN en (30), l'accent emphatique est attendu sur le pronom disloqué à gauche *ça* car la catégorie dénotée par le nom a déjà été évoquée et est pragmatiquement accessible. Dans la construction CDN en (31), l'accent emphatique est attendu sur *bombe* car on suppose que la voiture a déjà été évoquée. Ce qui est nouveau, c'est l'association de *voiture* à un haut degré qualitatif par l'entremise du GN *bombe*. Une autre différence concerne la fonction du GN. En (30), le nom *voiture* prédique une qualité du référent pragmatiquement accessible du pronom déictique *c'*, en l'occurrence la voiture dont il est spécifiquement question dans le contexte, La construction CCDN compare cette occurrence spécifique de la voiture au prototype de la catégorie. L'identification et la prédication sont intrinsèques. Par contraste, la construction CDN prédique une qualité de manière extrinsèque car *bombe* n'appartient pas à la même catégorie que *voiture*. Puisque le nom *bombe* est employé au sens figuré, il peut qualifier tout ce que le locuteur considère excellent (de la musique, des personnes, des vêtements, un film, etc.). *Bombe* n'est d'ailleurs pas le seul GN a remplir cette fonction en français contemporain. On trouve à cette position les GN suivants :

— prédication négative : *n'importe quoi*, *pipeau*, *vent*, *merde*, *foutaise*, *roupie de sansonnet*, etc.

— prédication positive : *or*, *balle*, *bonne came*, *grand art*, *haut vol*, *pain béni*, *tout cuit*, etc.

Dans les constructions CCDN et CDN, l'expression de l'identification et de la prédication repose sur la conversion du comptable au massif. La première identifie le référent du pronom sujet *c'* au prototype de la catégorie dénotée par l'attribut nominal. La seconde procède à une identification plus lâche. Dans les deux cas, la conversion n'implique pas de quantité. La conversion massive, systématique dans ces deux cas, a une fonction qualitative prédictible.

2.5 Quel bilan pour les grammaires de constructions ?

Le bilan que je dresse ici est, à l'évidence, subjectif. Il se limite à l'expérience que j'ai pu avoir des grammaires de constructions tant au niveau de la théorie que de la pratique. Dresser un bilan global est d'ailleurs illusoire tant les approches sont, une fois encore, multiples.

2.5.1 La démultiplication des approches est-elle un point faible ?

À l'occasion d'une candidature à un poste de maître de conférences, il m'a été suggéré par un membre du comité de sélection que la démultiplication des approches en était un point faible des grammaires de constructions. L'argument du collègue était le suivant : puisque les grammaires de constructions réunissent des générativistes et des cognitivistes-fonctionnalistes, chacun allant dans une direction opposée, le paradigme était contradictoire et ne pouvait pas fonctionner.

Ma réponse fut celle que je formule lorsque je défends que le générativisme et la linguistique cognitive sont deux approches complémentaires (plutôt qu'opposées) du point de vue de la méthode. L'une procède du haut (les règles abstraites) vers le bas (une infinité de phrases) tandis que l'autre procède du bas (les énoncés en contexte) vers le haut (les généralisations issues de l'observation des énoncés). Les prémisses théoriques ont beau être contradictoires, deux paradigmes peuvent se rejoindre sur certains points de la méthode. Il est par exemple possible de travailler sur des problématiques linguistiques posées en termes générativistes à partir des corpus (voir par exemple Bermúdez-Otero et al., 2000), en dépit de l'argument traditionnel selon lesquels les corpus ne peuvent pas apporter de preuves négatives (je reviens sur ce point au Chapitre 3). Inversement, l'exigence de générativité existe chez certains linguistes cognitivistes, notamment (Kay, 2013).

Une deuxième réponse, qu'il m'est possible de formuler avec le recul, consiste à dire qu'il existe des disparités au sein même de la linguistique générative et de la linguistique cognitive. Je considère cet état de fait comme un signe de bonne santé théorique. La BCG et la Sign-Based Grammar sont des points de contact entre les deux paradigmes. Ces deux approches rejettent la stricte séparation entre syntaxe et sémantique et entre règles et listes, de même qu'elles rejettent l'idée de dérivation, mais elles n'excluent pas totalement dans un premier temps l'idée d'une projection (Pinker, 1989 ; Jackendoff, 1997) et d'une architecture fondée sur les contraintes.

Pour résumer, l'hétérogénéité des grammaires de constructions est constitutive de son originalité et de sa vitalité. Elle invite à ne laisser aucune composante constructionnelle de côté, que ce soit d'un point de vue formel ou fonctionnel.

2.5.2 La non-distinction entre sémantique et pragmatique

Plus problématique est selon moi la non-distinction entre sémantique et pragmatique dans les GCC et la Grammaire Cognitive. J'ai développé ce point dans un article épistémologique (DESAGULIER, 2011a). Je l'ai également traité lors d'une journée d'étude de l'Association Française de Linguistique Cognitive à Bordeaux au printemps 2010 dont j'avais suggéré le thème : « What type of cognition for Cognitive Linguistics » (DESAGULIER, 2010).

Cette non-distinction est fondatrice et se retrouve dans la plupart des grammaires de constructions (DESAGULIER, 2011a, Section 2.1). Il existe un continuum entre les propriétés sémantiques d'une construction et ses propriétés pragmatiques.

> Among current non-modular approaches to grammar, CG places great emphasis on the fact that probably any of the kinds of information that have been called 'pragmatic' by linguists may be conventionally associated with a particular linguistic form and therefore constitute part of a rule (construction) of a grammar. (KAY, 1995)

La composante pragmatique d'une construction comprend l'information dépendante du contexte. En BCG, c'est cette composante qui rend compte du décalage entre la forme (interrogative) et le sens (incongruité) dans la construction WXDY en (6) et c'est la scalarité qui est au coeur de la construction *let alone* en (32).

(32) His friends have no idea what he looks like nowadays, let alone what he's up to.

Si la pragmatique a sa place en BCG, il s'agit toutefois d'une vision étroite, par souci d'économie .

> (...) pragmatic force and effect have been recognized primarily as conveyed through conventions of language, not in terms of conversational reasoning or socio-cultural constraints and possibilities. (FRIED et ÖSTMAN, 2004, p. 24).

La composante pragmatique a une place plus centrale en GCC. Contre les thèses universalistes, GOLDBERG (2006, p. 184) écrit :

> What we find is that the "universals" are only tendencies, and each tendency is argued to be a result of general cognitive, pragmatic, or processing attributes of human cognition.

À titre d'illustration, GOLDBERG (2006, p. 190) fait appel aux « Pragmatic Mapping Generalizations » pour rendre compte du fait que la structure argumentale d'une construction est très souvent soumise à des pressions pragmatiques et omet des arguments lorsque les conditions contextuelles et situationnelles sont réunies (DESAGULIER, 2011a, p. 111). Les causes du phénomène d'omission sont hors de portée des règles de projection universelles, qui n'admettent pas d'exception. Par exemple, dans certaines constructions agentives, le patient est fréquemment omis lorsque l'emphase est mise sur l'action, comme en (33).

(33) Some people just give and give, whereas others just take and take.

On trouve une vision plus développée, car mentionnant l'interaction, chez HILPERT (2008, p. 15) :

> There is evidence that many aspects of grammatical form emerge from the practice of actual conversation, in which speakers interact and convey meanings to each other. This does not endorse the claim that the primary function of language is the exchange of factual information. Rather, many conveyed meanings are purely social. The main idea is that interaction, for whatever purpose, shapes grammar.

CROFT (2009) va plus loin en rappellant aux linguistes cognitivistes que le débat sur la cognition ne se limite pas à l'examen des structures conceptuelles de la langue mais intègre des considérations sociales, précisément pour maintenir un accès privilégié à la cognition. Il s'inspire des travaux du philosophe (BRATMAN, 1992 ; BRATMAN, 1993 ; BRATMAN, 1997), à qui il emprunte le concept d'« activité coopérative partagée » pour illustrer la manière dont les processus cognitifs sont partagés, négociés et modifés par la communauté des locuteurs. Il s'inspire également du sociologue (WENGER, 1998) pour élargir la communauté linguistique à la « communauté de pratiques ». Croft définit ainsi un modèle macrosociologique dans lequel les unités linguistiques font partie d'une activité communicative. Cette activité s'insère à son tour dans une communauté de pratiques qui suppose l'implication mutuelle de ses membres.

Dans DESAGULIER (2011a, p. 112–118), je présente ce que pourrait être une grammaire de constructions sur la base d'une contribution de CROFT (2009), à savoir le produit d'une activité de langage qui se fonde moins sur l'échange de représentations individuelles prédéfinies que sur une activité conjointe et coopérative. Paradoxalement, alors que ce travail aurait dû m'orienter sur la voie d'une micro-analyse des constructions, mettant ainsi à profit les méthodes que j'ai apprises auprès des sociolinguistes, sociologues et anthropologues de l'EHESS, il m'a orienté vers les statistiques et notamment l'étude des phénomènes de fréquence (cf. Chapitres 3 et 4). J'ai trouvé dans les corpus et dans les statistiques exploratoires (certaines étant empruntées à la sociologie quantitative) un moyen d'étudier les constructions comme produits d'activités coopératives dans le cadre de communautés de pratiques.

2.5.3 Qu'est-ce qu'une construction ?

La critique la plus sérieuse qu'un détracteur est en droit d'adresser à l'encontre des grammaires de constructions est le manque de consensus quant à la définition de son unité de base : la construction. Nous avons vu en Section 2.3 que la BCG et la GCC s'opposaient quant à la nature de leurs taxinomies respectives : non-redondante pour la première et redondante pour la seconde. Une taxinomie non-redondante signifie que l'information linguistique n'est stockée qu'une seule fois dans l'inventaire des constructions, au niveau le plus haut. Une taxinomie redondante signifie que l'information linguistique peut être stockée à plusieurs niveaux. Cette différence taxinomique n'est pas un problème en soi, car elle peut très bien s'appliquer sur la base d'une même définition de la construction.

Beaucoup plus dommageable est la contribution récente de Paul Kay (KAY, 2013). Ce dernier a accentué les différences entre la BCG et les approches redondantes en affirmant qu'il existe une séparation nette entre la « grammaire » et la « méta-grammaire ». N'auraient leur place dans la grammaire que les schémas généraux et infiniment productifs. Les schémas

partiellement productifs seraient relégués à la périphérie de la grammaire, dans la méta-grammaire. La clivée en *all* illustrée en (34) est un exemple de schéma général et productif.

(34) *All-cleft*

a. All I want is to stand and look at you, dear boy ! (BNC-FPU)

b. All she reads is textbooks.

Les clivées en *all* héritent leur syntaxe d'une construction plus abstraite, commune aux autres clivées (par exemple les clivées en *wh-*). Ce qui les distingue, c'est l'interprétation scalaire qu'elles invitent, ce que Kay nomme « 'below-expectation' reading ». Les clivées en *all* sont pleinement productives car elles ne sont contraintes qu'au niveau lexical, « with respect to the filler constituent of the subject phrase » (2013, p. 37) et parce qu'elles offrent un moyen de répérer les énoncés qui relèvent de la construction sans avoir à en faire une liste. Selon Kay, la clivée en *all* est, à ce titre, une construction.

À l'inverse, le schéma *A as NP* illustré en (35) n'est que partiellement productif.

(35) *A as NP*

a. stiff as a board

b. cool as a cucumber

c. flat as a pancake

Il est vrai que, si l'on connaît le sens de *cool* et le sens de *cucumber*, on ne peut prédire ni l'association des deux, ni le sens de l'ensemble. De plus, ce schéma ne peut pas être étendu librement.

(36) a. ??cool as a tomato

b. ??hot as a zucchini

Selon Kay, *A as NP* est un schéma non-productif. Ce n'est pas une construction mais un « patron d'innovation » (*pattern of coining)*.

Ma réponse à l'article de Kay (Desagulier, 2015a) comporte deux moments. Une réponse théorique (que je résume ci-dessous) et une réponse empirique, que je présente au Chapitre 4. La réponse théorique se décompose en trois points. Premièrement, la BCG vue par Kay ne définit pas la productivité de manière univoque. Comme le souligne Barðdal (2008), il est difficile de savoir si la productivité est graduelle ou binaire. Kay et Fillmore (1999) semblent admettre plusieurs degrés de productivité :

> (...) a construction-based approach appears to provide promise of accounting both for the relatively idiomatic and for the abstract and more fully productive aspects of a language.

Cette conception de la productivité est en porte-à-faux vis-à-vis de la définition binaire proposée par Kay (2013). Pourtant, Fillmore (2002) reconnaît que « productivity is a

notion of degree ». S'il reconnaît que la productivité (l'application de règles générales sans exception) est bien distincte de l'invention (*coining*, l'usage de patrons improductifs pour créer de nouvelles expressions), il ne relègue pas l'invention à la périphérie de la grammaire.

> There is a view of grammar according to which the grammar proper will identify only the productive processes. Since the ability to create new words, using non-productive processes, is clearly a linguistic ability, it is my opinion that a grammar of a language needs to identify constructions that exist for "coining" purposes as well.

En cela, la conception de la productivité de Fillmore est identique à celle de la GCC et diamétralement opposée à celle proposée par Kay dans le cadre de la BCG.

Deuxièmement, la corrélation établie par Kay entre la nature schématique d'une construction et la productivité est problématique. Bybee (2010) montre que la construction *V someone A* illustrée en (37) est productive au niveau de l'adjectif, mais peu productive au niveau du verbe (elle n'admet que peu de types : *drive*, *send* ou *make*).

(37) *V someone A*

a. She drove me **crazy/mad/insane/nuts/etc.**

b. She **drove/made** me crazy.

Une même construction peut avoir différents degrés de productivité. Croft et Clausner (1997) observent également qu'un seul et même schéma peut être productif à divers degrés : « The instanciation of one schema may function as an intermediate schema for other more specific instantiations (1997, p. 251) ». En d'autres termes, parce qu'un patron n'est pas pleinement productif au niveau de schématisme le plus haut n'implique pas qu'il soit improductif aux niveaux de schématisme inférieurs.

Enfin, il ne semble pas pertinent d'indexer la productivité d'une construction entièrement sur la fréquence de type (V). D'autres facteurs contribuent à la notion de productivité : la similitude sémantique entre les occurrences (Croft et Clausner, 1997) ou la cohérence sémantique (Barðdal, 2008).

La linguistique cognitive-fonctionnelle postule une distinction nette a priori entre les fréquences hautes et les fréquences faibles (qu'il s'agisse de fréquence de type ou de fréquence de token). (Langacker, 1987 ; Bybee, 1985 ; Bybee, 2001 ; Croft et Clausner, 1997). Pourtant, dès que l'on se confronte aux chiffres, on se rend compte que le seuil entre les fréquences hautes et les fréquences faibles est relatif. J'y reviens au Chapitre 4.

On pourrait conclure à ce stade que le choix d'une grammaire de constructions radicalement réductrice – « a severe view of grammar » selon Kay (2013, p. 33) – ou redondante est une affaire de préférence théorique. Je pense au contraire que ce choix doit être fait sur une base empirique (Desagulier, 2015a, p. 43). La frontière entre ce qui relève d'une construction et ce qui relève d'un patron d'innovation ne peut être décidée qu'en corpus, à condition d'admettre un degré de productivité partielle (voir également Goldberg, 2016). L'empirie est selon moi la seule base d'une réconciliation possible.

3 Quelles données de corpus ?

> *Usage-based linguistics is essentially a distributional science in the sense that linguists explore the distribution of linguistic elements at every level of linguistic analysis : phonology, morphology, syntax, semantics, pragmatics and text linguistics etc. Corpus linguistics is no exception to this.*
>
> — **Stefan Th. Gries**
> (2014, p. 365)

3.1 Introduction

Nous avons vu au Chapitre 1 que les données pouvaient être de deux types : observationnelles et expérimentales. Dans mon parcours de recherche, j'ai largement privilégié les premières par l'entremise de la linguistique de corpus. Néanmoins, ces cinq dernières années, j'ai été amené à travailler à partir de données expérimentales au gré de directions de Masters 2. Entre ma soutenance de thèse et la rédaction de cette synthèse, j'ai eu l'occasion de croiser de nombreux collègues se déclarant linguistes de corpus. Autant j'ai pu témoigner d'un engouement croissant envers ce type de données, autant les linguistes de corpus ont un profil aussi varié que les objets qu'ils manipulent.

Dans sa définition contemporaine la plus simple, un corpus est un ensemble de textes produits dans un cadre naturel (c'est-à-dire sans intention de les compiler dans un corpus). Les textes sont rassemblés de manière à être représentatifs d'une ou plusieurs variétés, d'un ou plusieurs registres ou genres, et dans l'optique de faire de l'analyse linguistique (Gries, 2009, p. 7). L'expérience nous confronte d'emblée à une définition plus complexe parce qu'un corpus se définit tout autant par ses caractéristiques propres que par l'usage qui en est fait.

Dans ce chapitre, je présente une réflexion sur les types de données utilisées pour capturer l'usage en grammaires de constructions. La question sous-jacente est la suivante : les données utilisées par les linguistes cognitivistes sont-elles en mesure de capturer l'usage ?

3.2 Typologie des corpus

Proposer une typologie complète des corpus est rendu vite caduque parce que ce domaine de la linguistique est en constante mutation. Un novice aura certainement du mal à faire le lien entre le `Survey Corpus` (originellement le Survey of English Usage corpus) commencé en 1959 par Randolph Quirk, achevé trente ans plus tard et archivé sous forme de fiches de 6×4 pouces dans un gigantesque meuble classeur à University College London et les très grands corpus compilés depuis moins d'une dizaine d'année à partir de la Toile (BARONI et al., 2009). La nature protéiforme des corpus est l'effet d'une multitude de critères définitoires, que je présente dans mon ouvrage (DESAGULIER, à paraître, Section 1.2.1) et résume ci-dessous.

3.2.1 Un échantillon représentatif équilibré d'échantillons

Si l'on ne s'en tient qu'à la définition simpliste proposée plus haut, tout ensemble de textes compte comme corpus. Toutefois, les conclusions que le linguiste peut tirer de l'exploitation d'un tel corpus ne sont pas toutes équivalentes. Je présente ci-dessous ce que je considère être les propriétés essentielles communes aux corpus prototypiques, à savoir l'échantillonnage, la représentativité et l'équilibre. Ces critères garantissent un minimum d'exigence scientifique dans l'exploitation des corpus.

Imaginons un instant qu'un linguiste angliciste dispose de tous les énoncés parlés ou écrits produits depuis l'apparition de la langue. Quand bien même un linguiste aurait accès à cette gigantesque base de données, l'exploiter demanderait des ressources démesurées et le chercheur serait vraisemblablement perdu au milieu de cette masse d'information. Cette base de données n'existant pas, il faut pouvoir travailler à partir d'un échantillon de cette langue que l'on souhaite étudier. En Figure 3.1, les cercles bleus représentent les énoncés d'une langue ou d'une variété de cette langue et les cercles rouges l'échantillon. Un corpus est un échantillon de ce type.

L'échantillonnage peut être vu comme un pis-aller dans la mesure où l'on sait que certains phénomènes sont nécessairement absents de l'échantillon. L'échantillonnage est en fait un atout, à condition de bien connaître le schéma d'échantillonnage et de n'émettre des conclusions sur des résultats obtenus qu'à l'échelle du corpus. L'échantillonnage est d'autant plus important qu'il conditionne les deux autres critères définitoires.

Le deuxième critère définitoire d'un corpus est sa représentativité. Un corpus est représentatif lorsque son schéma d'échantillonnage est fidèle à la variation interne à la langue visée (ou un dialecte de cette langue) (BIBER, 1993, p. 244). Par exemple, un corpus constitué pour étudier la langue des adolescents parisiens ne pourra pas se contenter de conversations entre pairs. Il faudra également inclure des transcriptions de conversations avec des adultes (parents, professeurs, éducateurs) ainsi qu'une part de textes écrits (courriels, réseaux sociaux, etc.). Pour estimer la variabilité, on s'appuie sur des paramètres linguistiques et para-linguistiques. Les paramètres para-linguistiques incluent par exemple le mode (écrit/parlé), le format (publié/non publié), le cadre (institutionnel, privé, publique), des renseignements sur le locuteur (genre, âge, profession, etc.), sur l'interlocuteur (présent, absent, actif, passif,

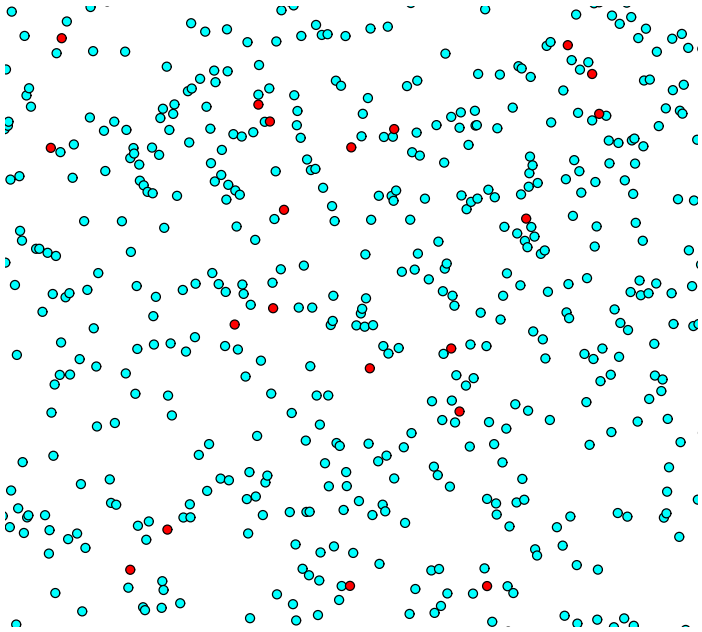


Figure 3.1: Une population et un échantillon (DESAGULIER, à paraître, Figure 8.1)

unique, multiple, etc.), la factualité de l'extrait (information, fiction, etc.), les thèmes abordés (religion, politique, éducation, etc.), etc. Le choix des paramètres linguistiques a un impact sur le type d'étude que l'on souhaite mener. BIBER (1993, p. 249) souligne à juste titre que les classes grammaticales ne sont pas toutes représentées également dans les textes. Par exemple, les pronoms et les contractions sont plus fréquents dans les textes de nature communicative, tandis que les relatives en *WH* sont typiques des textes à visée informative.

Le troisième critère définitoire d'un corpus est son équilibre. Un corpus est équilibré lorsque la proportion des échantillons qui font sa représentativité est conforme à la proportion des mêmes éléments dans la langue (ou le dialecte) cible. Par exemple, le corpus `Brown` (W. N. FRANCIS et KUČERA, 1979) et son équivalent britannique `LOB` (Lancaster-Oslo-Bergen) (JOHANSSON et al., 1978) sont cités comme exemples de bonne pratique en matière d'équilibre. Les créateurs de `Brown` ont fait en sorte de représenter chacun des genres et sous-genres de la collection de livres et de périodiques des bibliothèques de l'université Brown et de Providence Athenæum. Comme il y avait 13 fois plus de textes en sciences que de textes en science-fiction, 80 textes scientifiques furent inclus dans le corpus contre 6 de science-fiction. Les corpus `Brown` et `LOB` étant de taille modeste et fondés sur des modèles d'échantillonnage maîtrisés, il ne fut pas très difficile de les rendre représentatifs. Pour un corpus de référence censé représenter un dialecte ou une langue, l'objectif est beaucoup plus difficile à atteindre.

Les critères exposés ci-dessus sont un idéal. Aucun corpus ne les respecte tous intégralement, ce qui n'a rien d'étonnant puisqu'en l'état actuel des connaissances, les linguistes n'ont qu'une idée approximative de la répartition des compositions exactes des langues naturelles en matière de registres, de genres, de mode, etc. On sait que les langues naturelles sont principalement parlées. Pourtant, le `British National Corpus` (`BNC`) comporte 90% de textes écrits, pour 10% de conversations transcrites. La solution pour le linguiste consiste à bien connaître la composition du corpus et à examiner les résultats obtenus à la lumière de ces chiffres.

3.2.2 Les corpus de l'anglais

Pour bien prendre conscience de la diversité structurelle des corpus (avant même de parler de différences d'usage), j'ai décrit 17 des corpus les plus utilisés en linguistique anglaise en fonction de 7 critères. Ces critères sont décrits dans mon ouvrage (DESAGULIER, à paraître, Chapitres 1 et 3). Les données ont été rassemblées dans le Tableau 3.1.

Tableau 3.1: Description comparative de 17 corpus de l'anglais à l'aide de 7 variables (DESAGULIER, à paraître, Tableau 3.1)

corpus	variety	general vs. specific	static vs. dynamic	synchronic vs. diachronic	stored data format	mode	size
Bank of English	var : international	general	dynamic	synchronic	text	spoken + written	largest
BNC	var : GB	general	static	synchronic	text	spoken + written	large
BROWN	var : US	general	static	synchronic	text	written	small
CHILDES-English	var : international	specific	dynamic	synchronic	text + audio + video	spoken + written	large
COCA	var : US	general	static	synchronic	text	spoken + written	large
COHA	var : US	specific	static	diachronic	text	written	large
COLT	var : GB	specific	static	synchronic	text	spoken	very small
enTenTen12	var : international	general	static	synchronic	text	written	very large
FLOB	var : GB	general	static	synchronic	text	written	small
Frown	var : US	general	static	synchronic	text	written	small
GloWbE	var : international	general	static	synchronic	text	written	very large
Helsinki	var : GB	specific	static	diachronic	text	written	small
ICE-GB	var : GB	general	static	synchronic	text + audio	spoken + written	small
Lampeter	var : GB	specific	static	diachronic	text	written	small
LOB	var : GB	general	static	synchronic	text	written	small
London Lund Corpus	var : GB	general	static	synchronic	text	spoken	very small
SBCSAE	var : US	specific	static	synchronic	text + audio	spoken	very small

Le tableau a été résumé à l'aide d'une méthode exploratoire, l'analyse des correspondances multiples, que je décris à la Section 4.6.4 du Chapitre 4. Le graphe de la Figure 3.2 a été produit avec cette méthode. Plus deux variables sont proches dans l'espace à deux dimensions, plus on peut dire qu'elles ont des profils similaires.

On obtient, en synthétisant, deux grands profils. Chacun se subdivise en deux sous-profils :

- les corpus de taille relativement modeste se répartissent dans la partie gauche du graphe ;
  - dans la partie supérieure, on trouve les corpus diachroniques et spécifiques ;
  - dans la partie inférieure, on trouve les corpus synchroniques ;
- les corpus de taille relativement grande se répartissent dans la partie droite du graphe ;
  - dans la partie supérieure, on trouve les corpus dynamiques d'anglais représentant des dialectes de plusieurs pays ;
  - dans la partie inférieure, on trouve les corpus statiques généraux.

Dans mes recherches, j'ai privilégié le recours au `BNC` et au `Corpus of Contemporary American English` (`COCA`). Ces corpus étant très utilisés par la communauté des linguistes de corpus spécialisés dans l'anglais, ils permettent une réplication aisée des recherches.

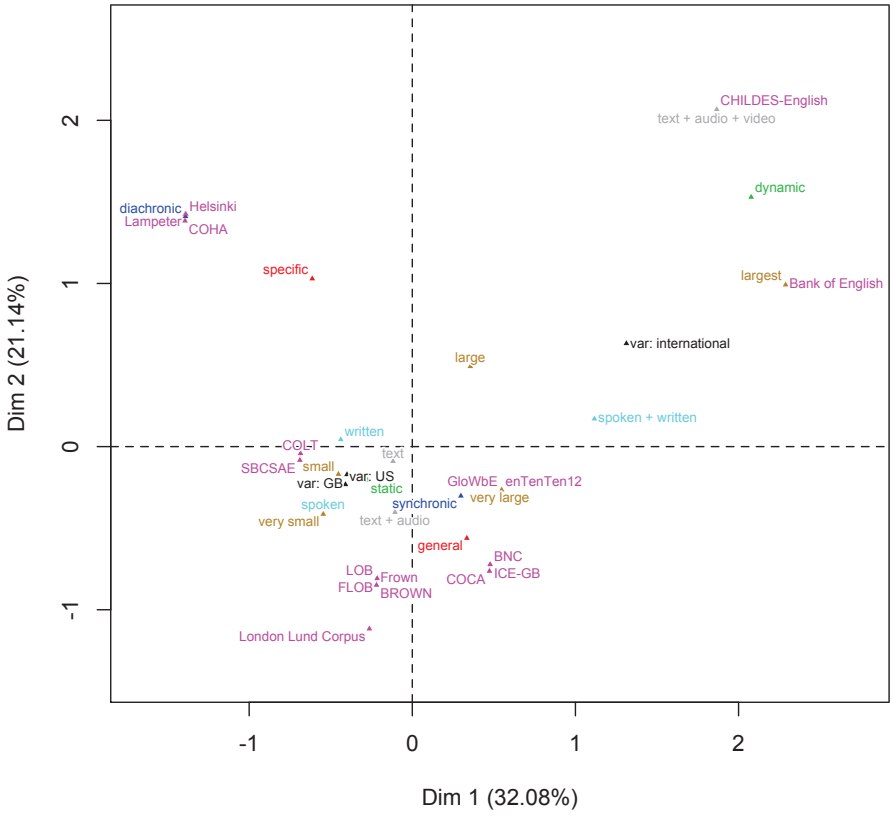


Figure 3.2: Graphe issu d'une analyse des correspondances multiples présentant les profils relatifs des grands corpus de l'anglais (DESAGULIER, à paraître, Figure 3.1)

3.3 Typologie des usages

De même qu'il existe une grande variété de corpus, il existe une pluralité d'usages. Ces usages se placent dans un continuum en fonction de la place du corpus dans l'organigramme méthodologique du linguiste. Je distingue trois grandes familles d'usage : le corpus comme source ou recueil d'exemples, le corpus comme outil en vue d'une quantification et le corpus comme modélisation de la grammaire mentale.

3.3.1 Une source pour un recueil d'exemples

Au niveau le plus élémentaire, le corpus est une source à partir de laquelle on compile un recueil d'exemples. Cette pratique est connue sous le nom de « corpus-illustrated research » (TUMMERS et al., 2005). Elle est purement qualitative car il s'agit d'illustrer un phénomène linguistique ou, a contrario, de trouver des contre-exemples. Ses détracteurs au sein de linguistique de corpus comparent cette pratique à la « chasse aux papillons », comme le déplore RENOUF (2015), au sens où la recherche du bel exemple l'emporte sur l'exigence de généralisation. Cette métaphore est injuste dans la mesure où, à ma connaissance, les

linguistes puisant des exemples dans des corpus ne se réclament pas de la linguistique de corpus.

Il est néanmoins possible de faire de la linguistique de corpus qualitative, à condition de pouvoir comparer les résultats à ceux obtenus à partir d'un autre corpus. Lorsque j'étudie la conversion du comptable au massif sous l'angle des grammaires de constructions (DESAGULIER, 2012a), je ne dispose pas d'un corpus de français parlé suffisamment grand et adapté au registre familier nécessaire à l'extraction d'une masse critique d'exemples.

L'outil de prédilection de la linguistique de corpus qualitative est le concordancier, un logiciel permettant de faire la recherche d'une occurence cible (un mot, une construction) accompagnée de son contexte. J'ai donc écrit un script pour R (R CORE TEAM, 2016) pour extraire des données à partir de serveurs de blogs francophones identifiés et de Usenet sans passer par la constitution préalable d'un corpus (DESAGULIER, 2012a, p. 205–206). J'ai bien entendu veillé à restreindre mon interprétation des résultats à la provenance des exemples sans pour autant généraliser au français dans son ensemble. J'ai dépassé pour cette étude le simple recueil d'exemples. J'ai constitué, pour chaque construction (CCDN et CDN) un jeu de données selon les mêmes critères et ai procédé à un classement des occurrences sur la base du schéma syntaxique et des déterminants. Enfin, j'ai procédé à un calcul de fréquences brutes (DESAGULIER, 2012a, Tableaux 1, 2 et 3).

L'article ayant été rédigé bien avant sa publication, à une période où je m'initiais encore aux techniques de corpus, je n'ai pas utilisé de mesure d'association pour quantifier le degré d'attraction entre les constructions et les lexèmes qui y interviennent. Pour ce faire, il aurait fallu que je procède à des extractions sur un corpus fermé afin de disposer des fréquences marginales (cf Chapitre 4). Sans être purement qualitative, cette première étude sur corpus n'est donc pas encore quantitative au sens où je le décris plus loin (DESAGULIER, 2012a, p. 228–229).

3.3.2 Une étape dans le cercle empirique

À un degré supérieur de systématisation, le corpus est une étape dans le cercle empirique décrit au Chapitre 1 (Figure 1.2). La linguistique de corpus vient en appui de la théorie. À ce titre, elle est qualifiée de « corpus-based linguistics » (par opposition à « corpus-driven ») par TOGNINI-BONELLI (2001). C'est l'approche que j'adopte à présent. Elle va de pair avec une quantification des résultats, rendue possible par le recours à des corpus fermés.

GRIES (2014) considère que la linguistique de corpus telle qu'elle est pratiquée en linguistique de l'usage est une « science distributionnelle ». Des considérations de sens sont inférées à partir de la distribution, de la dispersion et/ou de la co-occurrence des unités linguistiques. Pour cette raison, la linguistique « corpus-based » a recours à des listes de fréquence, des mesures de dispersion et des fréquences de co-occurrences.

Les linguistes qui adoptent l'approche « corpus-based » cherchent une forme de généralisation à partir de leurs observations. Cette généralisation s'obtient le plus souvent par une forme de quantification, par le recours aux fréquences brutes, aux fréquences relatives ou à des

tests statistiques. La quantification est au cœur de mes publications après 2012. J'y reviens en détail au chapitre suivant.

3.3.3 Une modélisation de la grammaire mentale

Certains linguistes considèrent le corpus comme une modélisation de la grammaire mentale parce qu'ils considèrent que la grammaire mentale est schématisable sous la forme d'un corpus. On est très proche ici de la définition de la grammaire proposée en Grammaire Cognitive (LANGACKER, 1987), à savoir un répertoire d'unités linguistiques mémorisées à partir de l'usage. Chaque expérience de l'usage laisse une trace dans la mémoire du locuteur et interagit avec les unités stockées précédemment. Ce répertoire dynamique est ce que TAYLOR (2012) nomme le corpus mental [1]. Tel qu'il est décrit par John Taylor, le corpus mental reste une abstraction conceptuelle.

Les linguistes de corpus qui se réclament de la mouvance « corpus-driven » étendent et radicalisent les principes de l'approche « corpus-based » : le corpus n'est pas une étape dans le cercle empirique mais l'alpha et l'oméga car il est considéré comme le seul accès possible à la compétence. L'approche « corpus-driven » se rattache historiquement à la tradition firthienne et de ce fait accorde une place centrale à l'étude des collocations, que FIRTH (1957) résume ainsi : « You shall know a word by the company it keeps ». Considérons un mot au sens apparemment vague : *something*. Ce même mot acquiert un sens majoritairement négatif dès lors qu'il est le sujet grammatical de *happen*. L'exemple (38) illustre ce phénomène sémantique à l'aide de cinq occurrences extraites au hasard depuis le BNC.

(38) a. (…) I was drunk, I think **something happened** (…) (KBN)
b. (…) but Amanda explained it to me, about Joe. **Something happened**, he got accused of something (…) (KD9)
c. (…) **Has something happened** here ? (singing) Right here ! Right now ! (KP6)
d. I think I was rude to Marina or something then **something happened**. (KP6)
e. (…) I don't think he (pause) he had to go out or **something happened**, and I thought, oh blow this for being (…) (KST)

Chez Firth, une collocation n'est pas le simple fruit d'une juxtaposition mais d'une attente mutuelle entre un mot cible (le « nœud ») et son contexte lexical. Il y a entre les deux une attente mutuelle (1957, p. 181). Firth n'avait pas connaissance des corpus. Ses intuitions ont été traduites en méthodologie « corpus-driven » par les « néo firthiens », notamment SINCLAIR (1966), SINCLAIR (1987), SINCLAIR (1991), SINCLAIR et CARTER (2004), STUBBS (2001) et TEUBERT (2005).

Dans mon parcours de linguiste de corpus, je n'ai pas adopté le point de vue théorique de l'approche « corpus driven » car je ne pense pas que la grammaire se réduise à un corpus mental. Si tant est que cela soit démontrable, il me semble que les locuteurs n'ont pas un égal accès à toutes les expériences linguistiques auxquelles ils ont été confrontés.

1. La Grammaire de Constructions Cognitive (cf. Chapitre 2) se distancie de cette vision de la grammaire comme corpus mental mais partage l'idée que la grammaire est un répertoire d'unités linguistiques structurées en réseau.

Par contraste, les requêtes formulées en corpus présupposent un égal accès à toutes les parties du corpus. Cette différence sépare à elle seule le corpus comme objet créé du corpus mental comme abstraction. Si l'équation entre corpus mental et corpus linguistique peut être maintenue, c'est peut-être en pensant ce dernier comme la trace collective d'une conscience linguistique.

3.4 Les défis de la linguistique de corpus

La linguistique de corpus n'étant pas l'apanage de la linguistique de l'usage, on est en droit de se demander si le recours à la linguistique de corpus permet de capturer l'usage. Je décris ici trois défis à relever pour une linguistique de corpus de l'usage : le sens, la variation et l'interaction.

3.4.1 Ancrage cognitif et catégorisation

La linguistique de l'usage repose sur un vaste ensemble de principes (LANGACKER, 1987 ; LANGACKER, 1988 ; BARLOW et KEMMER, 2000 ; BYBEE, 2010). Deux s'en détachent : l'ancrage cognitif et la catégorisation.

L'ancrage cognitif est défini par LANGACKER (1987, p. 59) comme suit :

> Every use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit ; moreover, units are variably entrenched depending on the frequency of their occurrence (*driven*, for example, is more entrenched than *thriven*).

Selon Langacker, dès qu'une unité linguistique (un phonème, un morphème, un schéma morphosyntaxique, etc.) est utilisée ou perçue, elle active un nœud ou un ensemble de nœuds dans la grammaire du locuteur. Plus l'unité est fréquente, plus elle a de chances d'être stockée de manière indépendante dans la grammaire. En d'autres termes, la répétition joue un rôle clé dans l'établissement de conventions. On retrouve cette idée dans la linguistique exemplariste (BYBEE, 1985 ; BYBEE, 2006 ; BYBEE, 2007 ; BYBEE, 2010). Il s'agit ici de fréquence perçue et non de fréquence mesurée. La question est de savoir si les fréquences mesurées en corpus rendent compte de la fréquence perçue (DESAGULIER, 2012b).

La catégorisation renvoie à une faculté cognitive par laquelle nous appréhendons la diversité du sensible à travers des catégories conceptuelles. Selon LAKOFF (1987), les êtres humains catégorisent en émettant des jugements de différence et de similitude sur la base de leurs observations. Ils relient ainsi les concepts sont ainsi reliés de manière encyclopédique (et non dictionnairique).

3.4.2 Opérationnaliser le sens, la variation et l'interaction

La linguistique « corpus based » fait la part belle à deux grandes méthodes : l'analyse collocationnelle et l'analyse en traits distinctifs. Nous avons déjà vu plus haut que l'analyse

collocationnelle déduisait le sens à partir d'une distribution conjointe [2]. Le Tableau 3.2 est issu d'un article comparant quatre intensifieurs de l'anglais considérés par les grammaires traditionnelles comme quasi-synonymes : *rather*, *quite*, *fairly* et *pretty* (DESAGULIER, 2014, Tableau 1).

Tableau 3.2: Extrait d'un exemple de collocations entre intensifieurs et adjectifs dans COCA

rather			*quite*		
adjective	freq. in corpus	freq. in construction	adjective	freq. in corpus	freq. in construction
large	119 992	260	*different*	17 021	2 247
different	17 021	231	*sure*	137 372	1 347
small	165 348	189	*clear*	81 553	805
difficult	6 672	129	*good*	378 826	578
unusual	17 736	116	*right*	4 558	548
simple	48 134	96	*possible*	88 919	458
limited	3 533	95	*similar*	60 967	328
good	378 826	89	*ready*	55 338	300
high	191 591	85	*common*	63 239	278
strange	23 457	80	*simple*	48 134	271
...	...	...	...	...	...
fairly			*pretty*		
adjective	freq. in corpus	freq. in construction	adjective	freq. in corpus	freq. in construction
good	378 826	346	*good*	378 826	7 492
easy	59 914	337	*sure*	137 372	1 241
large	119 992	281	*bad*	90 297	743
common	63 239	278	*clear*	81 553	729
high	191 591	247	*big*	187 641	583
simple	48 134	243	*tough*	33 746	486
new	64 824	202	*cool*	3 235	481
certain	71 149	201	*close*	94 845	471
small	165 348	190	*hard*	123 894	436
typical	20 483	154	*strong*	69 137	383
...	...	...	...	...	...

En première analyse, ce tableau permet d'entrevoir le domaine conceptuel sur lequel intervient chaque intensifieur [3]. Prenons un exemple. L'adjectif *good* est commun aux quatre adverbes. Cependant, en comparant la fréquence constructionnelle de l'adjectif à sa fréquence dans le corpus, on voit qu'il est ici spécifique à *pretty*. Il en va de même pour l'antonyme *bad*. On peut en conclure que, dans le corpus, l'adverbe joue à la fois le rôle d'intensifieur vis-à-vis de *good* et *bad* (on parlera d'effet modérateur) et qu'il est donc spécialisé dans la modération des jugements de valeur.

L'analyse collocationnelle est une manière d'opérationnaliser le sens à la lumière de ce qu'entend la linguistique de l'usage, sous l'angle double de l'ancrage et de la catégorisation. Pour revenir à notre exemple, plus la collocation entre *pretty* et *good/bad* est fréquente [4], plus les expressions complexes *pretty good* et *pretty bad* sont susceptibles d'être ancrées en tant qu'unités indissociables dans l'esprit du locuteur. Ces nœuds constructionnels sont par ailleurs classés en fonction du domaine conceptuel dans lequel ils interviennent. Les

2. HOEY (2005) va plus loin et fait de l'analyse des collocations le cœur de la théorie de l'amorçage lexical (*lexical priming*). Dans cette théorie, le mot est le principe organisateur du langage. Les locuteurs internalisent des schémas de cooccurrences possibles. Employer un mot déclenche une série d'attentes conceptuelles lexicales que nous exploitons pour construire du discours. La lexis est donc au coeur de notre faculté langagière, car les choix grammaticaux et lexicaux sont conditionnés par l'amorçage lexical.

3. Pour une analyse détaillée, voir DESAGULIER (2014).

4. Nous verrons au Chapitre 4 qu'il est simpliste de faire reposer l'ancrage sur la seule fréquence brute.

nœuds intervenant dans des domaines identiques ou similaires seront proches dans le réseau d'unités symboliques tandis que les nœuds intervenant dans des domaines différents seront éloignés. Ce dualisme entre similitudes et différences peut être vu comme une déclinaison des mécanismes de catégorisation.

L'analyse en traits distinctifs (*feature-based analysis*) repose également sur l'idée de cooccurrence. Contrairement à l'analyse collocationnelle, l'analyse en traits distinctifs (*feature-based analysis*) ne déduit pas le sens d'une cooccurrence formelle. Elle prétend y avoir accès plus directement, par la description de chaque observation (GLYNN, 2014). Les traits dont il est question peuvent être d'ordre formel (morphologique, syntaxique) et sémantique. Les traits sémantiques n'ont que peu à voir avec leurs homologues en linguistique structurale puisqu'ils sont de nature encyclopédique et non dictionnairique.

Lorsque *quite* et *rather* modifient des adjectifs épithètes, ils alternent entre une position pré-adjectivale (39a –39c) et une position pré-déterminantale (39b–39d). Cette propriété les distingue des autres intensifieurs (39e–39f).

(39) a. That has proved to be **a quite difficult** question to answer. (position pré-adjectivale)

b. That has proved to be **quite a difficult** question to answer. (position pré-déterminantale)

c. That is **a rather difficult** question to answer. (position pré-adjectivale)

d. That is **rather a difficult** question to answer. (position pré-déterminantale)

e. I know it is **a fairly difficult** question. (position pré-adjectivale)

f. ??I know it's **fairly a difficult** question. (position pré-déterminantale)

Le Tableau 3.3 sert typiquement de base à l'analyse en traits distinctifs. Il ne contient qu'un extrait des données utilisées dans mon étude comparative de *quite* et *rather* DESAGULIER (2015b).

Tableau 3.3: Extrait d'un exemple de tableau servant de base à l'analyse en traits distinctifs (DESAGULIER, 2015b, Tableau 7)

construction	intensifier	text_mode	text_type	text_info	sem_class
PREDETERMINER	QUITE	SPOKEN	OTHERSP	S pub debate	psych_stim_good
PREADJECTIVAL	QUITE	WRITTEN	FICTION	W fict prose	factual
PREADJECTIVAL	RATHER	WRITTEN	FICTION	W fict prose	dullness
PREADJECTIVAL	RATHER	WRITTEN	FICTION	W fict prose	atypicality_odd
PREADJECTIVAL	QUITE	SPOKEN	OTHERSP	S meeting	importance
PREADJECTIVAL	RATHER	WRITTEN	NONAC	W religion	difficulty_complexity
PREADJECTIVAL	RATHER	WRITTEN	NEWS	W newsp other : social	singularity
PREADJECTIVAL	QUITE	WRITTEN	NONAC	W biography	factual
PREDETERMINER	QUITE	SPOKEN	OTHERSP	S lect soc science	age_young
PREADJECTIVAL	RATHER	WRITTEN	NONAC	W nonAc : nat science	simplicity
...	...	...	...	...	...

Chaque ligne du tableau représente une observation. Chaque observation est caractérisée par autant de descripteurs que nécessaire, placés par convention en colonnes. Dans

l'article en question, je cherche à décrire l'alternance entre les schémas préadjectival et prédéterminantal pour *quite* et *rather*. Le jeu de données comprend plus de 3000 observations. Chaque trait (chaque variable) a plusieurs modalités. La variable `intensifier` comprend 2 modalités (`quite`/`rather`). La variable `construction` est de nature syntaxique et comprend également 2 modalités (`predeterminer` et `preadjectival`). Les variables contextuelles `text_mode`, `text_type` et `text_info` sont extraites des métadonnées XML du corpus. Elles comprennent respectivement 2 modalités (*spoken* et *written*), 8 modalités (`NONAC` non-academic writing, `FICTION`, `NEWS`, `OTHERSP` other spoken) et 18 modalités (`S parliament`, `S interview`, `W pop lore`, etc.). La variable `sem_class` décrit sémantiquement les adjectifs intensifiés. Elle comprend 59 modalités choisies au moment de l'annotation manuelle du jeu de données.

Les tableaux servant de base à l'analyse en traits distinctifs peuvent contenir d'autres types de variables, décrivant par exemple des phénomènes sociolinguistiques et interactionnels. Ils sont ensuite résumés à l'aide de méthodes exploratoires, comme par exemple l'analyse des correspondances multiples (DESAGULIER, 2015b, Section 4). Ces méthodes catégorisent les observations en regroupant celles qui ont des profils similaires et en distinguant celles qui ont des profils différents. Ce faisant, elles fournissent un point d'accès au phénomène de catégorisation. Elles dégagent également les profils de variables prototypiques et les profils atypiques, fournissant ainsi un point d'accès privilégié au phénomène d'ancrage cognitif.

Les méthodes de linguistique de corpus présentées succinctement ci-dessus sont grandement compatibles avec l'agenda de recherche de la linguistique de l'usage. C'est tout naturellement que la linguistique de l'usage les a adoptées. Cela explique en partie mon revirement empirique au début des années 2010.

3.5 Les limites de la linguistique de corpus viennent-elles des corpus ?

La linguistique de corpus est d'autant plus un atout pour le linguiste que ce dernier en connaît les limites. Je présente ci-dessous les critiques adressées à la linguistique de corpus par des chercheurs extérieurs au domaine. Je formule quelques réponses.

3.5.1 La question des preuves négatives

Historiquement, la première critique formulée à l'encontre de la linguistique de corpus provient de la linguistique générative (CHOMSKY, 1957, 15 et suivantes) : les corpus ne fourniraient pas de preuves négatives. Indéniablement, les corpus ne permettent pas d'affirmer qu'une expression, un énoncé ou tout autre usage est impossible sur la seule base de son absence dans un corpus.

Dans une approche introspective de type « top-down », allant des règles abstraites aux phrases effectives, l'absence de preuves négatives pose problème. Qu'un linguiste générativiste puisse se prononcer sur la possibilité ou l'impossibilité d'un exemple est fondamental car le

jugement de grammaticalité est intuitif. Dans une approche de type « bottom up » telle que la linguistique de l'usage, l'absence de preuves négative n'est pas rédhibitoire.

De l'absence de (40) ou (41) dans le BNC, la linguistique de l'usage aura soin de ne rien conclure quant à l'agrammaticalité de ces constructions.

(40) He be working.

(41) He working.

Ces deux constructions étant parfaitement attestées en African American Vernacular of English, leur absence dans le BNC n'indique rien quant à leur inacceptabilité.

Selon STEFANOWITSCH (2006), les corpus peuvent apporter des preuves négatives, à condition de savoir les faire apparaître à l'aide d'un outillage statistique approprié. Il s'avère que la différence qualitative entre une forme non attestée et une forme rare est mince. Elle n'est pas binaire, mais progressive, à l'image du gradient d'acceptabilité en linguistique introspective allant de ? (improbable) à ?? (fortement improbable) et * (impossible). Montrer qu'une forme est rare/absente n'est que la première étape, la seconde étant d'en déterminer la raison. Pour rendre compte de l'inacceptabilité, il faut en rechercher les causes, l'agrammaticalité n'en étant qu'une parmi d'autres. Enfin, poser la question de la rareté d'une forme est biaisé si l'on ne se fie qu'aux fréquences brutes. L'emploi de fréquences relatives ou de mesures d'association permet de capturer la sous-représentation d'une forme en corpus (je reviens sur ce point au Chapitre 4). La conclusion de Stefanowitsch est radicale : les linguistes de corpus peuvent se départir des jugements introspectifs.

Le positionnement que j'adopte est plus consensuel (CHAMBAZ et DESAGULIER, 2016, p. 3). Je ne pense pas que l'introspection soit incompatible avec la linguistique de l'usage ou la linguistique de corpus, comme je l'ai montré en Section 1.2.3. Par ailleurs, la grammaire étant une généralisation à partir de l'observation de l'usage dans sa diversité, les preuves négatives n'ont qu'une importance relative (GLYNN et FISCHER, 2010). Le recours à ce type de preuves peut être utile à un niveau d'analyse plus local que celui de la généralisation.

3.5.2 Généraliser à partir d'un échantillon

La taille d'un corpus est relative. À sa sortie sous forme de CD-ROM dans les années 90, le BNC, qui comptait près de 100 millions de mots, était considéré comme un très grand corpus. Grâce aux progrès techniques, qui ont rendu les aspirateurs de donnée et les algorithmes d'annotation et les serveurs très performants, plusieurs corpus récents dépassent à présent le milliard de mots (*tokens*), comme par exemple :

— `Hansard Corpus` (DAVIES, 2016a) : 1,6 milliards de mots ;

— `Wikipedia Corpus` (DAVIES, 2015) : 1,9 milliards de mots ;

— `Global Web-Based English` (DAVIES, 2013) : 1,9 millards de mots ;

— `ukWaC British English web corpus` (FERRARESI et al., 2008) : 2,25 milliards de mots ;

— NOW Corpus (DAVIES, 2016b) : 2,8 milliards de mots (comptage au 31 août 2016) ;

— enTenTen12 (THE SKETCH ENGINE, 2012) : 13 milliards de mots.

Comparé à ces corpus plus récents, le BNC ne semble plus aussi grand qu'à sa sortie.

D'un côté, cette course à la taille critique est une bonne chose. Plus un corpus est grand, plus un linguiste spécialisé dans la recherche d'une forme rare a de chances de l'y trouver. Les corpus de langue naturelle ont une distribution zipfienne (ZIPF, 1949), ce qui veut dire qu'un très grand nombre de types rares coexiste avec un petit nombre de types fréquents (Figure 3.3). En augmentant la taille du corpus, on étend la queue de la courbe et la probabilité d'observer des phénomènes rares.

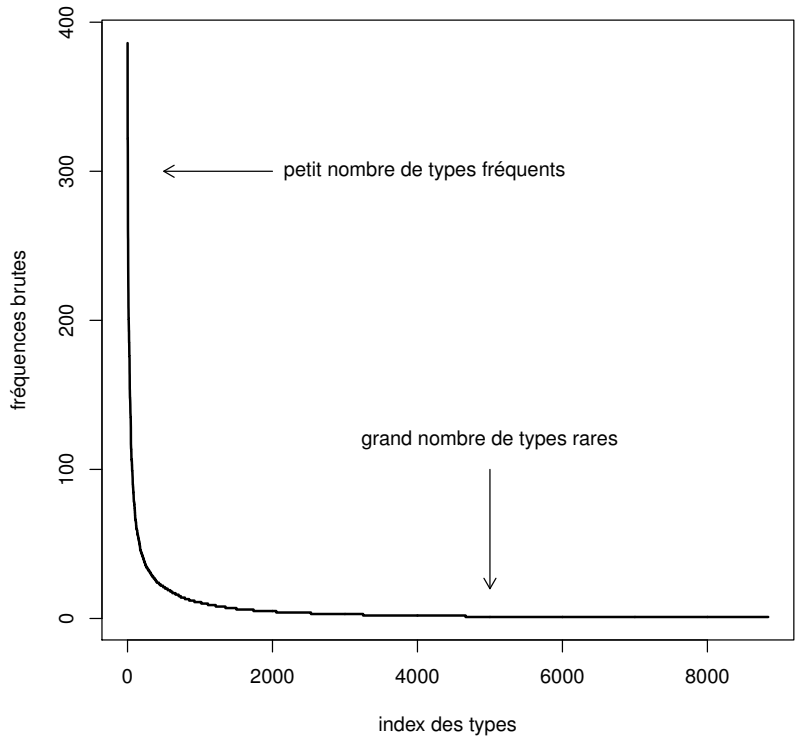


Figure 3.3: Un exemple de distribution zipfienne (DESAGULIER, à paraître, Section 6.2)

D'un autre côté, en augmentant la taille d'un corpus, on augmente par la même occasion le bruit. Lorsque l'on effectue une requête en corpus, le bruit correspond aux données extraites mais non voulues. Admettons par exemple que l'on souhaite extraire les occurrences de la construction *A as NP* décrite au chapitre précédent. Une requête syntaxique impliquant un adjectif suivi de *as* et d'un GN retournera indistinctement les exemples (42), (43) et (44), sachant que seul le premier est correspond à l'objet de la recherche.

(42) She is happy as a lark.

(43) They are happy as a couple.

(44) She was happy as a child.

Pour filtrer les occurrences voulues, il faut pouvoir assigner un sens à *as*, ce qui reste délicat. Une solution commune consiste à formuler une requête assez générale engendrant du bruit, et à nettoyer les données manuellement par la suite, ce qui a pour avantage de prendre connaissance des données. Le problème avec un très grand corpus, c'est que le bruit est souvent présent dans des proportions défiant tout nettoyage manuel par la suite.

Travailler à partir d'un corpus de taille modeste n'est pas en soi une mauvaise chose (voir à ce titre le plaidoyer pour ce type de corpus par GRIES et DIVJAK (2010)) [5]. Les recherches sur les dialectes du Lancashire menées sur un corpus réduit par HOLLMANN et SIEWIERSKA (2007) sont un bon exemple. Selon moi, à l'heure actuelle, le seul domaine dans lequel le recours à un grand corpus est décisif est l'apprentissage de vecteurs lexicaux, comme nous le verrons au Chapitre 6. On pourrait inclure la linguistique de type « corpus driven », si ce n'est qu'il n'est pas certain qu'un corpus de très grande taille soit à la mesure du corpus mental des locuteurs, vraisemblablement plus réduit.

Plus donc que la taille, c'est l'échantillonnage qui fait la qualité d'un corpus. C'est sur la base de l'échantillon que le linguiste peut espérer généraliser les résultats obtenus à un registre, un dialecte, voire une langue. Dans un entretien mené par ANDOR (2004, p. 97), Chomsky met à mal la généralisation des résultats fondés sur les investigations en corpus :

> Corpus linguistics doesn't mean anything. It's like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. Well, you know, sciences don't do this.

Pour bien comprendre cette critique, il faut la replacer dans son contexte. Pour Chomsky, la linguistique a vocation à rendre compte du langage (*I-language*) et non de la langue, ce qui justifie sa comparaison avec les sciences physiques et la chimie. Ces sciences font référence à des lois immuables aisément exportables à l'ensemble de l'univers. La première erreur de Chomsky est de penser que la linguistique de corpus a pour but de rendre compte de la compétence, par opposition à la performance. Pas plus qu'elle n'a vocation à rendre compte du langage, la linguistique de corpus ne cherche pas à rendre compte d'une langue dans son intégralité [6]. Une étude (fictive) qui comparerait deux quasi-synonymes (par exemple *sort of* et *kind of*) sur la base d'un corpus d'anglais britannique représentatif et équilibré vis-à-vis des genres et des registres du Royaume-Uni ne pourrait pas permettre à son auteur de conclure quoi que ce soit quant à l'anglais britannique. Les conclusions devraient être limitées au corpus. Dans le cas contraire, l'auteur commettrait une généralisation fallacieuse. À moins d'assortir l'étude d'un volant quantitatif et de méthodes statististiques inférentielles, la généralisation des résultats à la l'anglais britannique relève de l'acte de foi, c'est-à-dire d'un saut qualitatif depuis les conclusions effectuées sur un ensemble fini jusqu'aux lois inconnues d'une langue. Pour résumer, la linguistique de corpus n'est qu'un outil dans la description de la langue et du langage.

5. De nouveau, le sens de « taille modeste » est relatif et dépend de l'étude de cas. Pour ce qui est de la majorité des études de cas que j'ai menées, le `BNC` m'est apparu comme un corpus modeste, par conséquent aisément exploitable.

6. Cette idée fausse est, il est vrai, partagée par des novices en linguistique de corpus, encouragés en cela par la publicité qui est faite des corpus généraux tels que le `BNC`, `COCA`, `GloWbE`, etc. Cette publicité fait croire aux utilisateurs qu'ils disposent d'une copie miniature de la langue qu'ils étudient. En guise de copie, j'ai tendance à croire qu'il s'agit, au mieux, d'une photographie subjective.

La seconde erreur de Chomsky est de n'aborder qu'un aspect des sciences physiques et de la chimie : l'expérimentation. Les corpus ne relèvent pas de l'expérimentation mais de l'observation. L'observation est une composante essentielle des sciences dures. L'univers étant par essence infini, on ne peut en observer que des échantillons à une échelle infinitésimale. La force des sciences dures vient de l'utilisation de méthodes inférentielles. Celles-ci permettent de généraliser les phénomènes observés dans un échantillon à un ensemble infiniment plus grand. Ces méthodes inférentielles sont parfaitement applicables aux corpus, Je montrerai au Chapitre 4 que les méthodes semi-paramétriques les plus récentes permettent de réaliser le saut depuis l'échantillon jusqu'à une langue en estimant la loi inconnue de cette langue et par déduction la loi du phénomène estimé.

3.5.3 Quantifier l'inquantifiable

La linguistique de corpus quantitative est souvent critiquée par rapport aux prétentions de la quantification. On lui reproche notamment d'employer des méthodes trop objectives pour capturer ce qui n'est à la portée que du regard expert du linguiste, par exemple le sens. Ce reproche est partiellement justifié car les corpus ne révèlent aucunement l'intention des locuteurs (WIDDOWSON, 2000). Par ailleurs, une fois décontextualisés de leur source et recontextualisés dans un ensemble souvent hétérogène, les éléments de discours perdent de leur cohésion discursive. Toutefois, cette vision de la linguistique de corpus est erronée.

En effet, dans le cadre des linguistiques « corpus-based » et « corpus-driven », il est inconcevable de se lancer dans l'exploitation d'un corpus (a fortiori d'effectuer des quantifications) avant d'avoir défini une hypothèse de recherche et une idée claire quant à l'opérationnalisation de cette question. Ce n'est qu'une fois l'hypothèse de recherche opérationnalisée qu'il devient possible de formuler une requête et d'en interpréter les résultats.

3.6 Quels outils pour la linguistique de corpus ?

Faire de la linguistique de corpus ne requiert plus des connaissances pointues en ingénierie linguistique ou un équipement informatique ultra performant. Les linguistes spécialisés en anglais ou en espagnol peuvent depuis quelques années bénéficier de la plateforme créée par Mark Davies (`http://corpus.byu.edu/`). Cette plateforme regroupe tout à la fois des corpus de tailles différentes et une interface de requête, ce qui semble convenir à près de 11000 chercheurs inscrits [7]. Plusieurs annuaires permettent d'accéder à des corpus constitués (annotés ou non), le plus connu étant la base de données *Oxford Text Archive* (`https://ota.ox.ac.uk/catalogue/index.html`). Pour qui souhaite constituer ses propres corpus, il existe à présent une pléthore d'outils libres et/ou gratuits pour générer des concordances, annoter des parties de discours faire de l'annotation sémantique [8]. J'ai pourtant fait le choix d'une autre plateforme.

7. Au 31 août 2016, 10571 chercheurs sont inscrits (`http://corpus.byu.edu/researchers.asp`).
8. Voir notamment les logiciels réalisés par Lawrence Anthony : *Antconc*, *FireAnt*, *TagAnt* et *AntCLAWSGUI* (`http://www.laurenceanthony.net/software.html`).

Mes recherches en corpus impliquant des requêtes complexes ainsi que la quantification post-extraction, j'ai cherché un outil me permettant dans le même environnement de :

— traiter des chaînes de caractères ;

— effectuer des requêtes à l'aide d'expressions régulières ;

— tabuler les données ;

— quantifier les données tabulées ;

— réaliser des tests statistiques ;

— résumer les résultats graphiquement.

Ces fonctionnalités sont réunies dans R (R Core Team, 2016). Très populaire auprès des statisticiens et des mathématiciens, c'est avant tout un langage interprété : l'utilisateur entre du code dans une console et celui-ci est interprété par la machine. R n'est pas un logiciel interactif. Pour l'utiliser, il faut apprendre les bases du code, ce qui représente un investissement en temps et en énergie non négligeable. Après avoir pris en considération le ratio entre une courbe d'apprentissage assez rude au départ et le gain en productivité d'une interface certes exigente mais flexible et puissante, j'ai pris la décision de me lancer dans R, encouragé en cela par la popularité croissante de cette interface auprès des linguistes sous l'impulsion de Stefan Th. Gries et Harald Baayen.

L'ouvrage que je présente dans le cadre de mon habilitation, *Corpus Linguistics and Statistics with R* [9], fait l'objet d'un contrat de publication depuis le 9 juillet 2015 chez Springer. Il s'agit de la première version dont la publication effective est prévue pour juillet 2017 [10]. Je prévois une phase de test lors d'un séminaire de linguistique de corpus au sein du département de Sciences du Langage à l'Université Paris 8. Je souhaite m'assurer que le code fonctionne à la fois sur Windows, Mac OS et Linux.

D'apparence technique, ce n'est pas qu'un livre d'ingénierie de corpus, mais le produit de cinq années d'apprentissage et de mise en pratique de R dans le cadre de mes recherches. Il est donc écrit du point de vue d'un linguiste. J'y propose des réflexions épistémologiques sur la linguistique de corpus et sur les méthodes quantitatives. L'ouvrage comporte deux parties. La première concerne les techniques d'exploitation des corpus et la seconde les méthodes quantitatives et statistiques. Il s'agit, en somme, de l'ouvrage dont j'aurais aimé disposer pour me lancer dans la linguistique de corpus quantitative.

Le premier volet reprend et met à jour certaines des techniques présentées par Gries (2009). Il s'étend sur les cinq premiers chapitres. L'introduction présente le statut théorique des corpus et leur matérialité. Elle invite également à réfléchir au ratio coût/bénéfice mentionné plus haut. Le chapitre 2 présente les bases du langage de programmation propre à R ainsi que les objets propres à l'environnement (les vecteurs, les listes, les matrices les data frames, etc.), les boucles, les assertions conditionnelles, etc. Au troisième chapitre, je dresse une typologie des corpus et montre comment constituer et exploiter des corpus en texte brut et des corpus annotés. Le quatrième chapitre aborde le traitement des chaînes de caractères. C'est un chapitre central au sens où j'exploite le langage de programmation pour manipuler les données textuelles des corpus [11]. À l'ère des corpus numériques, j'estime qu'il est fondamental

9. Le titre est provisoire.
10. Ceci explique l'absence d'index, de préface, et de remerciements.
11. À ce titre, les linguistes exploitent pleinement une facette de R que les sciences dures ont utilisé de manière secondaire pour l'étiquetage de leurs données.

pour un linguiste de corpus de savoir comment la machine interprète du texte. Lorsque l'on utilise une interface utilisateur graphique – par exemple le moteur de recherche des corpus mis à disposition par Mark Davies sur `http://corpus.byu.edu/` – ou *AntConc*, on est souvent loin de se douter des mécanismes impliqués en arrière-plan et l'on risque de ne pas obtenir ce que l'on cherche précisément. Ce chapitre invite le lecteur à anticiper le fonctionnement de la machine pour maximiser la précision des requêtes. Le cinquième chapitre applique les principes de traitement des chaînes de caractères pour réaliser trois objets indispensables à la linguistique « corpus-based », à savoir des concordances (Tableau 3.4), des jeux de données annotés (Tableaux 3.3 et 3.5) et des listes de fréquences (Tableaux 3.2 plus haut et 3.6 ci-dessous). Ces trois objets sont une base pour la quantification dont fait l'objet le second volet du livre, que j'aborde au chapitre suivant.

Figure 3.4: Extrait d'un exemple de concordance (*blood* dans *Dracula* de Bram Stoker) (DESAGULIER, à paraître, Figure 5.1)

	A	B	C
1	LEFT CONTEXT	NODE	RIGHT CONTEXT
2	all this region that has not been enriched by the	blood	of men patriots or invaders In old days there were
3	saw that the cut had bled a little and the	blood	was trickling over my chin I laid down the razor
4	right to be proud for in our veins flows the	blood	of many brave races who fought as the lion fights
5	the dying peoples held that in their veins ran the	blood	of those old witches who expelled from Scythia had mated
6	or what witch was ever so great as Attila whose	blood	is in these veins He held up his arms Is
7	more gladly than we throughout the Four Nations received the	bloody	sword or at its warlike call flocked quicker to the
8	and again though he had to come alone from the	bloody	field where his troops were being slaughtered since he knew
9	we threw off the Hungarian yoke we of the Dracula	blood	were amongst their leaders for our spirit would not brook
10	sir the Szekelys and the Dracula as their heart s	blood	their brains and their swords can boast a record that
11	underlying the sweet a bitter offensiveness as one smells in	blood	I was afraid to raise my eyelids but looked out

Figure 3.5: Extrait d'un exemple de jeu de données (*each/every* + GN dans le `BNC Baby`) (DESAGULIER, à paraître, Figure 5.4)

corpus file	info	mode	type	exact match	determiner	NP	NP_tag
A1E.xml	W newsp brdsht nat: commerce	wtext	NEWS	each nation	each	nation	NN1
A1E.xml	W newsp brdsht nat: commerce	wtext	NEWS	each other	each	other	NN1
A1E.xml	W newsp brdsht nat: commerce	wtext	NEWS	each other	each	other	NN1
A1E.xml	W newsp brdsht nat: commerce	wtext	NEWS	each country	each	country	NN1
A1E.xml	W newsp brdsht nat: commerce	wtext	NEWS	each type	each	type	NN1
A1E.xml	W newsp brdsht nat: commerce	wtext	NEWS	every problem	every	problem	NN1
A1E.xml	W newsp brdsht nat: commerce	wtext	NEWS	each double	each	double	NN1
A1E.xml	W newsp brdsht nat: commerce	wtext	NEWS	each jurisdiction	each	jurisdiction	NN1
A1E.xml	W newsp brdsht nat: commerce	wtext	NEWS	every share	every	share	NN1
A1F.xml	W newsp brdsht nat: editorial	wtext	NEWS	every day	every	day	NN1

Figure 3.6: Extrait d'un exemple de liste de fréquences des noms, verbes et adjectifs dans le `BNC Baby` (DESAGULIER, à paraître, Figure 5.5)

WORD	FREQUENCY
said	12704
know	10202
got	8825
get	6756
go	6427
think	5961
time	5827
see	5551
other	5448
want	4285
going	4144
way	3979
people	3846
good	3594
put	3575

La linguistique de corpus doit demeurer un outil au service du linguiste. Les techniques nécessaires à l'exploitation de corpus avec R ont quelque chose d'aliénant au premier abord. Je suis toutefois convaincu qu'en acceptant de se confronter à la matérialité des corpus, en s'astreignant à intégrer la logique d'un code de programmation et des algorithmes de traitement des chaînes de caractères, on arrive très vite à gravir la courbe d'apprentissage. Ce faisant, on acquiert une plus grande autonomie et une plus grande liberté dans l'exercice du métier de linguiste.

4 Vers une statistique de l'usage en grammaires de constructions

> *Language is never, ever, ever, random.*
>
> — **Adam Kilgarriff**
> (Kilgarriff, 2005)

4.1 De l'usage des statistiques en linguistique

L'usage des statistiques n'a pas bénéficié de la même diffusion que la linguistique de corpus. Cela s'explique en partie par le fait que l'on peut faire de la linguistique de corpus sans s'engager dans aucune forme de quantification [1]. Cela s'explique également par le fait que la pertinence des statistiques en linguistique n'est pas immédiatement évidente pour qui n'a jamais été formé à cette discipline. À ma connaissance, lorsqu'ils sont proposés, les cours d'initiation aux méthodes statistiques s'adressent principalement aux psycholinguistes, qui travaillent sur des données continues (issues principalement de mesures). Beaucoup plus rares sont les cours dispensés aux linguistes travaillant sur des données catégorielles (issues principalement de comptages).

Qu'on le veuille ou non, la linguistique contemporaine comporte désormais un minimum de quantification. Cette tendance s'accélère à l'ère de l'analyse des données dans une optique pluridisciplinaire. Travaillant à partir de données de corpus dans le cadre de la sémantique cognitive et des grammaires de constructions, je n'ai jamais eu de cours de statistiques lors de ma formation initiale. C'est un mal pour un bien au sens où j'ai amorcé l'apprentissage des statistiques au moment où elles me semblaient nécessaires. Ce faisant, je n'ai jamais séparé ces méthodes de mes objectifs de recherche.

Je présente ci-dessous les méthodes statistiques que j'estime pertinentes en linguistique de l'usage. Comme nous l'avons vu au sujet de la linguistique de corpus au chapitre précédent, le terme « statistiques » recouvre des réalités différentes en fonction de l'utilisation qui en est faite.

1. L'utilisation des corpus est attestée dès les années 80 chez de futurs acteurs de la linguistique cognitive (Dirven, Goosens et al., 1982 ; Dirven et Taylor, 1988). Cet emploi des corpus n'implique toutefois pas de quantification.

4.2 Le statut de la fréquence

Les données catégorielles font la part belle au comptage, autrement dit au phénomène de fréquence. Cette même fréquence est l'étalon à l'aune duquel l'ancrage cognitif des unités symboliques est évalué dans la linguistique cognitive de première génération. La question est de savoir si la fréquence mesurée de la linguistique cognitive de seconde génération, beaucoup plus empirique que la première, est de même nature que la fréquence intuitive (DESAGULIER, 2012b ; DESAGULIER, 2015c).

4.2.1 Les limites théoriques des fréquences brutes

Si la fréquence est considérée comme le principe organisateur de la linguistique de l'usage, elle pose deux types de problèmes sur le plan théorique. Premièrement, sa définition évolue entre la linguistique cognitive de première génération (introspective) et la linguistique cognitive de deuxième génération, beaucoup plus quantitative. Deuxièmement, son rôle central dans l'ancrage cognitif est remis en cause depuis quelques années au profit notamment la saillance.

La linguistique cognitive de première génération ne propose aucune méthode empirique pour appréhender la fréquence. La fréquence n'est donc pas un concept opérationnalisé mais d'une phénomène perçu. Pour décider si une unité est ancrée, LANGACKER (1987, p. 59) fait intervenir l'idée de fréquence mais, selon lui, il est vain de vouloir l'appréhender quantitativement :

> [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.

Néanmoins, LANGACKER (2008, p. 238) admet plus tard que l'ancrage est graduel et opérationnalisable sous l'angle de la fréquence mesurée :

> (...) in principle the degree of entrenchment can be determined empirically. Observed frequency provides one basis for estimating it.

De subjective, la fréquence devient empirique car mesurable. Tout en demeurant fidèle à son orientation introspective, LANGACKER (2008, p. 86) ouvre la voie aux méthodes 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*.

Ce revirement intervient sous l'impulsion de la linguistique de corpus quantitative (GEERAERTS, 2010b, p. 263–264).

Le deuxième problème est le suivant : la linguistique de l'usage est influencée par l'équation entre fréquences hautes et ancrage cognitif. Cette équation est incomplète. Un phénomène

complémentaire entre en jeu : la saillance (SCHMID, 2010) [2]. La saillance et l'ancrage peuvent être en résonance ou 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. (DESAGULIER, 2015c, p. 103)

J'illustre ci-dessous une situation de dissonance possible entre ancrage cognitif et fréquence :

> Orson Welles's *Citizen Kane* provides a good illustration of how salience works. The film opens on the main character, Charles Foster Kane, at the end of his life. Kane dies calling for "Rosebud". It is the job of a reporter to discover who Rosebud is. It appears that Rosebud is (spoiler alert) the trademark of eight-year-old Kane's sled. Although the word is never uttered by Kane in his lifetime as a newspaper magnate, it is heavily entrenched in his mind. (DESAGULIER, à paraître, Section 9.1).

Pour résumer, une unité linguistique peut être ancrée cognitivement sans nécessairement être fréquente.

4.2.2 Les limites empiriques des fréquences brutes

Parler de « fréquence mesurée » est une simplification, la mesure n'allant pas de soi. Avant de s'engager dans un comptage, il faut savoir que compter. La définition « humaine » d'une unité linguistique n'est pas immédiatement exportable à son traitement par la machine. Une phase de réflexion est nécessaire avant tout calcul de fréquence (DESAGULIER, à paraître, Chapitre 5).

La linguistique exemplariste distingue la fréquence de *type* de la fréquence de *token*. (DESAGULIER, à paraître, Section 9.4.2). La fréquence de token est une fréquence textuelle. Elle dénombre le nombre d'occurrences d'une unité linguistique dans un texte et prend en compte les répétitions de cette unité. La fréquence de type ne dénombre que les occurrences uniques. Elle dresse en quelque sorte une typologie des unités dans un texte.

En soi, la fréquence de token est trompeuse. Il ne suffit pas de dénombrer les occurrences d'une unité pour conclure quoi que ce soit quant à sa haute fréquence ou sa rareté. Il faut, pour cela, avoir recours à des fréquences relatives. Le Tableau 3.2 au chapitre précédent permet d'estimer partiellement la fréquence relative d'un adjectif dans une construction. Savoir que l'adjectif *different* est intensifié 2247 fois par *quite* dans le BNC n'informe en rien sur la fréquence ou la rareté de l'adjectif dans la construction. Savoir que ce même adjectif apparaît 17021 fois dans le corpus permet de calculer une fréquence relative et de l'exprimer par exemple sous forme de pourcentage. Ainsi, on trouve $\frac{2247}{17021} \times 100 = 13.2\%$ des

2. Pour une typologie complète de la saillance, voir SCHMID (2007) et GEERAERTS (2000).

occurrences de *different* dans le contexte de la construction avec *quite*. Nous verrons plus bas que, calculée de la sorte, une fréquence relative ne résume qu'imparfaitement l'attraction entre *quite* et l'adjectif. Il faut également prendre en compte la fréquence totale de *quite* dans la constructions ainsi que dans le corpus.

Au vu de ce qui précède on est en droit de se poser deux questions. Tout d'abord, que cherche-t-on lorque l'on relève des fréquences d'occurrence ? Deux réponses viennent à l'esprit : soit on cherche à quantifier un phénomène en examinant son étendue à l'aune de la taille de l'échantillon, soit on cherche à déterminer si le phénomène ainsi mesuré apparaît plus fréquemment que ne le laisse supposer le hasard. Dans le premier cas, on fait appel à des statistiques descriptives. Dans le second cas, on fait appel à des statistiques analytiques. Ensuite, ces statistiques sont-elles en accord avec l'usage ? En d'autre termes, reflètent-elles le mécanisme probabiliste à l'œuvre dans la grammaire des locuteurs (GRIES et ELLIS, 2015) ?

4.3 Quelles statistiques pour les fréquences ?

Je distingue ici deux types de statistiques : les statistiques descriptives et les statistiques analytiques (ou inférentielles). Je les décris aux Chapitres 6, 7 et 8 de mon livre (DESAGULIER, à paraître). J'en résume les enjeux ci-dessous de manière à mettre en avant la valeur ajoutée des statistiques analytiques.

4.3.1 Les statistiques descriptives

À leur niveau le plus élémentaire, les données de fréquences sont appréhendées à l'aide de statistiques descriptives. Celles-ci se divisent en mesures de centralité et en mesures de dispersion (DESAGULIER, à paraître, Chapitre 7). À l'image des pourcentages, les mesures de centralité (moyenne, médiane et mode) résument un ensemble de mesures à l'aide d'un nombre unique. Les mesures de dispersion, souvent ignorées, indiquent à quelle distance d'une mesure de centralité donnée se trouvent les observations. On a recours pour cela à l'écart interquartile, la variance ou l'écart type. Si l'ancrage cognitif est bien indexé sur la fréquence, il faut également rendre compte de sa longitudinalité et des contextes dans lesquels une forme est fréquente.

4.3.2 Les statistiques analytiques

À un niveau plus avancé, les fréquences sont appréhendées à l'aide de statistiques analytiques. Celles-ci sont par nature inférentielles. Comme nous l'avons vu au chapitre précédent, la linguistique de corpus consiste à tirer des conclusions sur les énoncés d'une langue à partir d'un échantillon de cette langue (Section 3.2.1). Les statistiques analytiques reposent sur le concept de variable aléatoire. Les valeurs que prend une variable aléatoire sont soumises au hasard (DESAGULIER, à paraître, Section 8.4). L'exemple typique est le lancer de dé.

Lorsque l'on lance un dé une fois, chacun des six résultats possible a une chance sur six de se réaliser.

Contrairement aux mathématiciens, qui n'ont aucun problème pour concevoir et modéliser l'aléatoire, les linguistes ne trouvent pas naturel de réduire l'occurrence d'un énoncé à un tirage au hasard, comme le souligne KILGARRIFF (2005, p. 264) :

> The problem for empirical linguistics is that language is not random (...). Language is not random because we speak or write with purposes. We do not, indeed, without computational help are not capable of, producing words or sounds or sentences or documents randomly.

Aucun linguiste ne peut sérieusement penser qu'un échantillon est comparable à une urne de laquelle on tirerait des mots au hasard ou qu'une phrase est la somme aléatoire des n mots qui la composent. L'ordre des unités linguistique est primordial et la linguistique étudie précisément les contraintes qui président à cet ordre.

En apparence, les statistiques analytiques n'ont pas leur mot à dire. En effet, elles postulent des modèles probabilistes qui reposent sur de l'aléatoire. Or, ce postulat n'est jamais vérifié dans les langues. La linguistique de corpus quantitative argue que ces modèles sont pourtant utiles parce que c'est précisément en comparant la distribution de l'échantillon à ce que postule un modèle probabiliste théorique que l'on peut mesurer l'effet des contraintes linguistiques à l'œuvre dans l'échantillon.

4.4 La cooccurrence

Une grande partie du travail du linguiste de corpus consiste à étudier des cooccurrences au sens très large du terme. Sur la base de la description structurale d'un fait de langue, on va rendre compte de sa distribution à la lumière de phénomènes connexes de nature linguistique, para-linguistique ou extra-linguistique. Au sens très restreint, les cooccurrences sont lexicales.

4.4.1 Dépendance et indépendance

Dans mon livre, j'ai souligné que les fréquences brutes ne répondaient pas toujours aux questions que se posent les linguistes. Pour cela, je me suis appuyé non seulement sur les données de mes propres recherches mais aussi, ponctuellement, sur des recherches déjà publiées. Le Tableau 4.1 est adapté de TAGLIAMONTE et HUDSON (1999, p. 158). Les auteurs comparent les systèmes des verbes introducteurs de discours rapporté (*quotatives*) chez des jeunes locuteurs en anglais britannique et en anglais canadien [3].

La question sous-jacente à ce relevé de fréquences est double :

1. la sélection des verbes dépend-elle de la variété d'anglais ?
2. si c'est le cas, quels verbes dépendent le plus de chaque variété d'anglais ?

3. Les détails de l'étude sont résumés dans DESAGULIER (à paraître, Section 8.9.1)

Tableau 4.1: Distribution des marqueurs de discours rapporté dans deux corpus (anglais britannique et canadien)

verbes	anglais britannique	anglais canadien
say	209	219
go	120	135
be like	120	79
think	123	27
zero	66	123
be (just)	11	5
misc	16	24

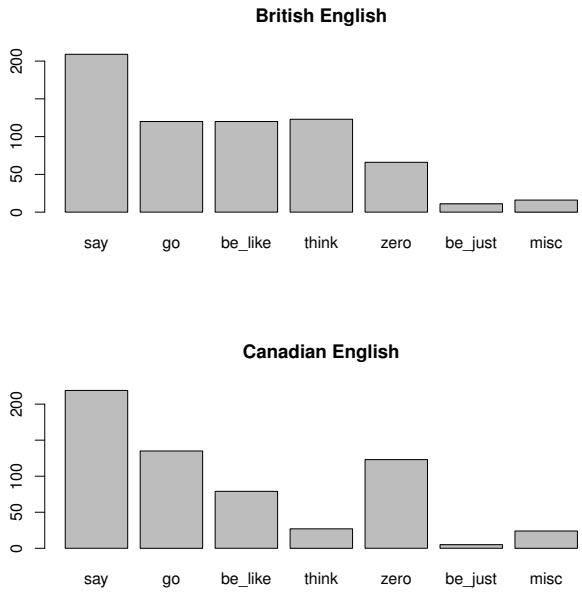


Figure 4.1: Deux diagrammes en barres montrant la distribution des verbes en fonction de la variété d'anglais

Le Tableau 4.1 est résumé par deux diagrammes en barres en Figure 4.1. D'un côté, les fréquences brutes nous renseignent partiellement quant au lien entre l'emploi des verbes et la variété de l'anglais [4]. Les variables *say*, *go*, *be just* et *misc* [5] ont des distributions relatives similaires, contrairement à *be like*, *think* et *zero*.

D'un autre côté, ces fréquences brutes ne disent rien quant à l'interdépendance entre le choix des verbes et la variété d'anglais. De plus, le tableau ne disposant pas de totaux marginaux (la somme des lignes et des colonnes), il ne nous permet pas d'estimer la déviation de chaque fréquence vis-à-vis d'une distribution aléatoire. Le recours aux statistiques analytiques permet de répondre à ces interrogations. Elles s'appuient sur le test d'hypothèses (DESAGULIER, à paraître, Section 8.6).

Pour déterminer si le choix des verbes introducteurs dépend de la variété de l'anglais, nous commençons par formuler deux hypothèses : H_0 et H_1. La première est l'hypothèse nulle.

4. Dans l'article original, les fréquences brutes sont assorties de pourcentages.
5. Un ensemble varié de verbes introducteurs.

Elle postule une relation d'indépendance entre les deux variables. La seconde est l'hypothèse alternative. Elle postule une relation de dépendance entre les deux variables.

H_0 : le choix des verbes et le choix de la variété d'anglais sont indépendants ;

H_1 : le choix des verbes et le choix de la variété d'anglais sont dépendants.

Lorsque l'on teste des hypothèses, on ne cherche pas tant à prouver que l'une des deux est vraie mais que l'autre hypothèse doit être rejetée.

Le test d'indépendance du χ^2 de Pearson est typiquement appliqué dans ce cas de figure. Il compare les effectifs réels (les fréquences brutes observées dans le corpus) aux effectifs théoriques (les fréquences auxquelles on s'attend si les deux variables sont indépendantes en conservant les mêmes totaux marginaux) [6]. En l'occurrence, nous obtenons un score de χ^2 d'environ 89 avec une probabilité associée proche de 0 ($2.2 \cdot 10^{-16}$). Cette « p-valeur » correspond à la probabilité d'obtenir un score de χ^2 aussi élevée sous l'hypothèse nulle d'indépendance. En l'occurrence, cette probabilité est quasiment nulle. Cela signifie que l'hypothèse nulle peut être rejetée et que les deux variables sont dépendantes. Le test du χ^2 ne renseigne pas pour autant sur l'intensité de la relation. Il faut pour cela faire appel à une autre mesure, la plus courante étant le V de Cramér. Cette mesure produit un score sur une échelle continue allant de 0 (il n'y a pas d'association entre les deux variables) à 1 (il y a une association parfaite). Pour ce qui est des variables du Tableau 4.1, $V \approx 0.27$. L'association entre le choix du verbe introducteur de discours rapporté et la variété de l'anglais est non-négligeable, sans être intense [7].

La réponse à la deuxième question ci-dessus (quels verbes dépendent le plus de chaque variété d'anglais ?) s'obtient en comparant chaque fréquence réelle du Tableau 4.1 à sa fréquence attendue correspondante (Tableau 4.2).

Tableau 4.2: Fréquences théoriques des marqueurs de discours en fonction de la variété d'anglais

verbes	anglais britannique	anglais canadien
say	222,881754	205,118246
go	132,791699	122,208301
be like	103,629601	95,370399
think	78,112764	71,887236
zero	98,422083	90,577917
be (just)	8,332028	7,667972
misc	20,830070	19,169930

Plutôt que d'effectuer une comparaison manuelle, peut calculer pour tout le tableau les résidus de Pearson (Tableau 4.3). Les résidus de Pearson sont positifs si la fréquence observée est supérieure à la fréquence attendue. Inversement, ils sont négatifs si la fréquence observée est inférieure à la fréquence attendue. Plus un résidu est éloigné de 0, plus la surprise est grande.

En pratique, il est plus rapide de visualiser les résidus directement à l'aide d'un diagramme de Cohen-Friendly (Figure 4.2).

6. Je décris la logique de ce test important dans mon ouvrage.
7. Cette intensité modérée est d'autant moins surprenante que l'on suppose une association plus forte entre le choix du verbe et des variables linguistiques et sociolinguistiques d'une granularité beaucoup plus fine.

Tableau 4.3: Résidus de Pearson

verbes	anglais britannique	anglais canadien
say	-0,9298376	0,9692643
go	-1,1100506	1,1571186
be like	1,608116	-1,6763028
think	5,0788087	-5,2941589
zero	-3,2680947	3,4066675
be (just)	0,9242849	-0,9634762
misc	-1,0582983	1,1031719

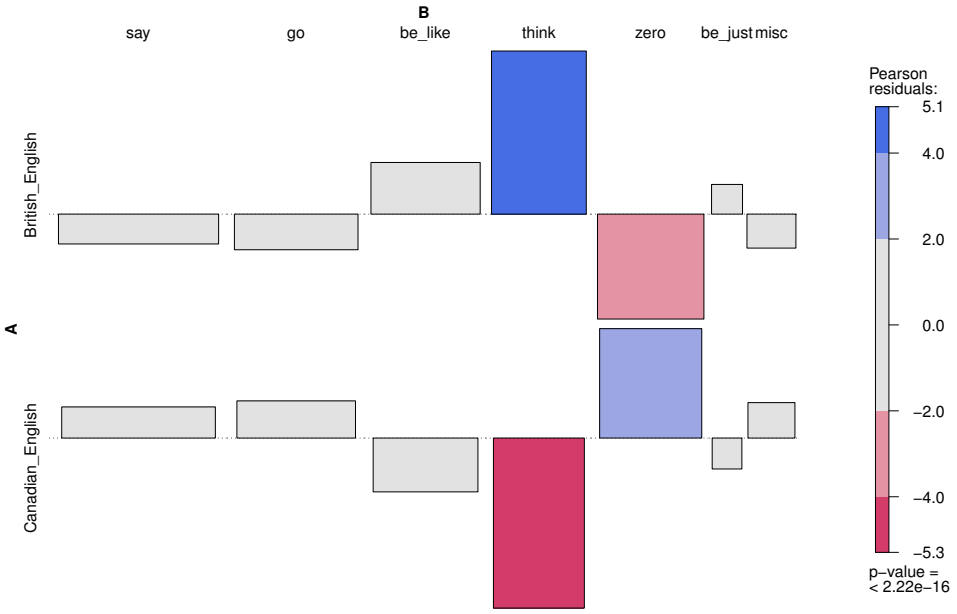


Figure 4.2: Diagramme d'association de Cohen-Friendly

Chaque cellule du Tableau 4.1 est représenté par un rectangle dont la hauteur est proportionnelle au résidu de Pearson et dont la largeur est proportionnelle à la racine carrée de la fréquence attendue. L'aire du rectangle correspond ainsi à la différence entre fréquence observée et fréquence théorique. La ligne en pointillés représente l'indépendance entre les variables. Si la fréquence observée d'une cellule est supérieure à la fréquence attendue, le rectangle apparaît au dessus de cette ligne. Si la fréquence observée d'une cellule est inférieure à la fréquence attendue, le rectangle apparaît au dessous. Les verbes introducteurs qui contribuent le plus à la dépendance entre les deux variables sont *think* et *zero*. En anglais britannique, *think* est sur-représenté tandis que *zero* est sous-représenté. En anglais canadien, *think* est sous-représenté tandis que *zero* est sur-représenté.

Il existe évidemment bien d'autres méthodes, mais si j'ai choisi cet exemple [8], c'est qu'il est représentatif de la démarche en statistiques analytiques. Tout en s'appuyant sur des fréquences brutes, ces statistiques mesurent le degré de surprise (ou d'absence de surprise) associé à un jeu de données.

8. En simplifiant toutefois ce qui figure plus en détail dans mon livre.

4.5 Les mesures d'association

Les mesures d'association capturent le phénomène de cooccurrence aux niveaux lexical et lexico-grammatical. Elles sont traditionnellement employées dans deux contextes :

— la recherche du degré d'attraction ou de répulsion entre des lexèmes (on parle de *collocation* (DESAGULIER, à paraître, p. 9.2.1)) ;

— la recherche du degré d'attraction ou de répulsion entre une forme linguistique et un phénomène grammatical (on parle de *colligation* (DESAGULIER, à paraître, p. 9.2.2)).

Plus récemment, au début des années 2000, Stefan Th. Gries et Anatol Stefanowitsch ont adapté l'inventaire des mesures d'association existantes au cadre des grammaires de constructions en proposant une famille de méthodes connue sous le nom d'analyse collostructionnelle. L'analyse collostructionnelle (dont je détaille les méthodes ci-dessous) vise à mesurer le degré d'attraction ou de répulsion des lexèmes vis-à-vis des constructions dans lesquelles ils sont observés. L'accueil de l'analyse collostructionnelle en France est mitigé pour une raison principalement historique. Au moment où l'analyse collostructionnelle commence à se diffuser, l'école française de lexicométrie est déjà bien implantée depuis au moins deux décennies (LAFON, 1980 ; LAFON, 1981 ; LAFON, 1984 ; LEBART et SALEM, 1994 ; MULLER, 1964 ; MULLER, 1973 ; MULLER, 1977). On reproche principalement à l'analyse collostructionnelle d'empiéter sur le calcul des spécificités de Lafon ou de faire doublon avec les études sur la colligation. Deux raisons m'ont convaincu du bien fondé de cette approche :

> L'originalité de l'analyse collostructionnelle ne vient donc ni de son principe de fonctionnement, ni des mesures d'association qu'elle fait intervenir. 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. (DESAGULIER, 2015c, p. 106).

L'analyse collostructionnelle est particulièrement adaptée à l'étude distributionnelle du sens. Je l'ai donc exploitée systématiquement dans l'étude conjointe de la quasi-synonymie et de l'intensification des adjectifs en anglais contemporain (DESAGULIER, 2014 ; DESAGULIER, 2015c ; DESAGULIER, 2015b ; DESAGULIER, 2015a).

4.5.1 La logique des mesures d'association

Les mesures d'association mesurent le degré d'attraction ou de répulsion entre deux unités linguistiques (simples ou complexes). Même si elles sont calculées à partir de fréquences brutes, les mesures d'association font intervenir des tests statistiques de manière à s'assurer que les unités en collocation sont significatives, c'est-à-dire plus fréquentes que ce que l'on est en droit d'attendre de leurs distributions aléatoires.

Pour ce faire, la plupart des mesures d'associations s'appuient sur un tableau de contingence tel que le Tableau 4.4. Celui-ci mesure la collocation entre deux mots : W_1 et W_2. Quatre cellules sont nécessaires dans ce cas :

— a dénote le nombre de cooccurrences entre W_1 et W_2 ;

— b dénote le nombre d'occurrences de W_1 sans W_2 ;

— c dénote le nombre d'occurrences de W_2 sans W_1 ;

— d dénote le nombre de mots dans le corpus qui ne sont ni W_1 ni W_2.

Tableau 4.4: Un tableau de contingence générique impliquant deux mots (¬ : "autre que")

	W_2	¬ W_2	sommes des lignes
W_1	a	b	a+b
¬ W_1	c	d	c+d
sommes des colonnes	a+c	b+d	a+b+c+d

Les mesures d'association dégagent les collocations statistiquement significatives à partir des fréquences contenues dans le tableau de contingence. Il existe une multitude de mesures d'associations, dont on trouve un inventaire détaillé chez CHURCH et al. (1991), EVERT (2005) et EVERT (2009) et PECINA (2010). Les quatre mesures les plus utilisées sont sans doute l'information mutuelle, le test exact de Fisher, le test du χ^2 et le rapport de log-vraisemblance (DESAGULIER, à paraître, Section 9.3.3). Ces mesures sont aussi exploitées en analyse collostructionnelle.

4.5.2 Associations symétriques

L'analyse collostructionnelle regroupe quatre méthodes, résumées dans le Tableau 4.5.

Tableau 4.5: Les méthodes de l'analyse collostructionnelle

Méthode	Mesures d'association	références	mes applications
analyse collexémique	test exact de Fisher, rapport de log-vraisemblance, information mutuelle, test du χ^2, odds ratio	STEFANOWITSCH et GRIES (2003)	DESAGULIER (2014)
analyse collexémique distinctive	test exact de Fisher, rapport de log-vraisemblance	GRIES et STEFANOWITSCH (2004b)	
analyse collexémique distinctive multiple	test multinomial	HILPERT (2006) et GILQUIN (2013)	DESAGULIER (2014) et DESAGULIER (2015b)
analyse collexémique covariante	test exact de Fisher, rapport de log-vraisemblance, odds ratio	GRIES et STEFANOWITSCH (2004a) et STEFANOWITSCH et GRIES (2005)	DESAGULIER (2015c) et DESAGULIER (2015a)

Je résume ci-dessous l'apport de chacune de ces méthodes.

L'analyse collexémique

L'analyse collexémique (AC) mesure le degré de répulsion ou d'attraction entre une construction et les lexèmes apparaissant dans cette même construction. Elle se fonde sur un tableau de contingence très semblable au Tableau 4.4, si ce n'est que l'on étudie l'association d'un lexème et d'une construction et non plus de deux lexèmes (Tableau 4.6).

Tableau 4.6: Tableau d'entrée pour une analyse collexémique (L : lexème, C : construction)

	L_j	$\neg L_j$	sommes des lignes
C_i	a	b	a+b
$\neg C_i$	c	d	c+d
sommes des colonnes	a+c	b+d	a+b+c+d

Je décris l'AC dans DESAGULIER (à paraître, Section 9.3.6.1). Je l'applique à l'étude de la quasi-synonymie de *rather*, *quite*, *fairly* et *pretty* dans DESAGULIER (2014, Section 4.1). Si ces quatre adverbes sont quasi-synonymes lorsqu'ils fonctionnent comme intensifieurs d'adjectifs (DESAGULIER, 2014, p. 153), on suppose que leurs distributions se caractérisent par des similitudes (parce les adverbes interviennent sur des domaines conceptuels en commun, ces domaines étant convoqués par les adjectifs) et des différences (parce que chaque adverbe intervient parallèlement dans des domaines conceptuels spécifiques). Il s'agit de montrer que les quatre adverbes sont des « synonymes cognitifs », selon la terminologie de PARADIS (1997, p. 71).

J'ai procédé à quatre extractions (une pour chaque adverbe) depuis le COCA et ai ignoré à ce stade la syntaxe de l'adverbe (pré-adjectival ou pré-déterminental) et de l'adjectif (épithète ou attribut), contrairement à DESAGULIER (2015b). La mesure d'association choisie est le rapport de log-vraisemblance, ou G^2 (DUNNING, 1993). L'analyse génère un tableau de sortie (Tableau 4.7) qui permet de classer les adjectifs par ordre décroissant d'attraction vis-à-vis des quatre adverbes.

Tableau 4.7: Les 10 collexèmes les plus spécifiques de *rather*, *quite*, *fairly* et *pretty* dans COCA

rather		*quite*		*fairly*		*pretty*	
adjectif	coll. strength	adjectif	coll. strength	adjectif	coll. strength	adjectif	coll. strength
large	1025.19	*different*	15082.43	*easy*	2686.55	*good*	61820.34
different	709.26	*sure*	8231.32	*common*	2078.88	*sure*	8148.39
unusual	705.49	*clear*	4923.96	*simple*	1883.49	*clear*	4764.46
small	521.34	*possible*	2216.18	*large*	1751.18	*bad*	4734.36
difficult	480.6	*similar*	1614.27	*good*	1523.3	*tough*	3635.13
odd	436.5	*good*	1482.02	*straightforward*	1433.43	*cool*	3628.34
remarkable	415.88	*ready*	1480.81	*certain*	1326.03	*amazing*	2848.64
limited	413.37	*simple*	1357.46	*typical*	1315.09	*big*	2604.22
vague	400.68	*remarkable*	1297.76	*high*	1250.31	*close*	2531.19
strange	384.69	*common*	1259.48	*consistent*	1112.44	*strong*	2139.59

Je renvoie le lecteur à l'interprétation détaillée que je fais de cette AC dans DESAGULIER (2014, p. 161–162). Deux tendances se dégagent :

— pour chaque adverbe, il est possible de regrouper les adjectifs modifiés à l'aide de quelques classes sémantiques ; par exemple, les collexèmes de *rather* se classent dans les catégories suivantes : dimension (*large*, *small*, *limited*), atypicalité (*unusual*, *odd*, *vague*, *strange*, *remarkable*), différence (*different*) et difficulté (*difficult*) ;

— la synonymie cognitive est avérée au sens où :

  — les adverbes intensifient en commun des domaines sémantiques identiques ; par exemple *quite* et *rather* interviennent tous deux dans le domaine de la typicalité tandis que l'intensification de l'extension spatiale est commune à *fairly* et *pretty* ;

  — les adverbes interviennent distinctivement dans certains domaines ; par exemple, l'atypicalité est le domaine distinctif de *rather*.

Tableau 4.8: Tableau d'entrée pour une ACD (L : lexème, C : construction)

	L_j	$\neg L_j$	row totals
C_1	a	b	a+b
C_2	c	d	c+d
column totals	a+c	b+d	a+b+c+d

Il est peu aisé de distinguer les adjectifs distinctifs de chaque adverbe sur la base du tableau de sortie. La distinctivité relève de l'analyse collexémique distinctive.

L'analyse collexémique distinctive (multiple)

L'analyse collexémique distinctive (ACD) mesure la préférence d'un lemme pour une construction par rapport à une autre construction équivalente. Je décris cette méthode dans DESAGULIER (à paraître, Section 9.3.6.2). Le Tableau (4.8) est un tableau de contingence générique utilisé comme entrée pour cette analyse.

Lorsqu'il y a plus de deux constructions alternantes, l'ACD est multiple (ACDM). J'ai particulièrement exploité l'ACDM dans DESAGULIER (2014) et DESAGULIER (2015b). Le Tableau 4.9 est extrait du tableau de sortie de l'ACDM que j'ai réalisée dans DESAGULIER (2015b, Section 4) afin de comparer les emplois de *quite* et *rather*. Les données sont extraites du BNC et prennent en compte la syntaxe de l'adjectif et de l'adverbe (DESAGULIER, 2015b, Tableau 1).

Tableau 4.9: Extrait d'un tableau de sortie de l'ACDM (DESAGULIER, 2015b, Tableau 3)

Coll_Word	obs freq	exp freq	pbin	SumAbsDev	LargestDev
sure	716	402.84	173.42	313.52	0_quite_adj_0
long	227	24.35	168.36	289.18	quite_det_adj_np
clear	570	325.52	124.56	227.21	0_quite_adj_0
happy	543	313.2	111.76	208.21	0_quite_adj_0
right	407	242.04	70.89	143.63	0_quite_adj_0
possible	235	132.23	57.11	103.59	0_quite_adj_0
different	290	100.84	56.94	154.42	0_quite_adj_np
good	205	61.88	51.45	97	quite_det_adj_np
big	77	12.55	39.93	62.73	quite_det_adj_np
large	99	22.31	36.6	91	quite_det_adj_np
distinct	41	8.08	17.53	39.86	0_quite_adj_np
separate	33	7.09	13.08	36.09	0_quite_adj_np
other	15	1.31	13.03	26.94	0_quite_adj_np
tired	28	6.5	12.94	21.42	0_rather_adj_0
extraordinary	19	2.03	12.87	24.83	det_quite_adj_np
surprised	45	14.71	12.56	24.22	0_rather_adj_0
substantial	23	3.89	12.09	32.65	quite_det_adj_np
unusual	29	6.77	11.18	28.18	det_rather_adj_np
slow	29	7.74	11.14	24.35	0_rather_adj_0
specific	21	3.82	10.07	25.07	0_quite_adj_np
remarkable	18	2.66	9.73	16.96	det_quite_adj_np
limited	32	10.68	8.85	24.53	0_rather_adj_0
gloomy	11	1.45	7.76	18.59	det_rather_adj_np
negative	10	1.37	6.88	15.49	det_rather_adj_np
eccentric	10	1.37	6.88	14.23	det_rather_adj_np
new	13	2.16	6.61	19.27	det_quite_adj_np
similar	40	19.05	5.69	22.14	0_rather_adj_0
grim	8	1.13	5.47	13.24	det_rather_adj_np
loud	5	0.57	3.77	7.66	0_rather_adj_np
inherent	2	0.03	3.62	4.72	rather_det_adj_np
recent	5	0.63	3.46	10.58	det_quite_adj_np
conservative	4	0.41	3.32	7.91	0_rather_adj_np
formal	5	0.74	3.2	9.49	0_rather_adj_np
fundamental	4	0.48	2.94	10.54	det_quite_adj_np
stronger	3	0.25	2.9	6.99	0_rather_adj_np
whole	2	0.06	2.85	6.14	rather_det_adj_np
low	12	4.38	2.83	10.72	0_rather_adj_np
essential	2	0.08	2.64	4.88	rather_det_adj_np
old	5	1.39	1.88	6.6	rather_det_adj_np
…	…	…	…	…	…

Si l'on fait abstraction des huit configurations syntaxiques, plusieurs tendances émergent (DESAGULIER, 2015b, §31), la principale étant que *quite* et *rather* se distinguent quant à

Tableau 4.10: Tableau d'entrée pour une ACC

	L_{slot1}	$\neg L_{slot1}$	row totals
L_{slot2}	a	b	a+b
$\neg L_{slot2}$	c	d	c+d
column totals	a+c	b+d	a+b+c+d

l'expression des jugements de valeur. Il existe une attraction entre *quite* et les adjectifs connotés positivement (par ex. *happy*, *good*, *extraordinary*, *remarkable*). Parallèlement, il existe une attraction entre *rather* et les adjectifs connotés négativement (par ex. *tired*, *slow*, *limited*, *gloomy*, *negative*, etc.).

Le Tableau 4.9 devient rapidement difficile à interpréter, a fortiori si l'on démultiplie les contextes distinctifs. J'aborde ce point plus bas (Section 4.5.2) et propose une solution en Section 4.6.

L'analyse collexémique covariante

L'analyse collexémique covariante (ACC) mesure le degré d'attraction ou de répulsion de lemmes dans différentes positions d'une même construction. Je décris cette méthode dans DESAGULIER (2015c, Section 3.3) DESAGULIER (à paraître, Section 9.3.6.3). Le Tableau 4.10 est un tableau de contingence générique utilisé comme entrée pour l'ACC.

J'ai recours à l'ACC dans deux études : DESAGULIER (2015c) et DESAGULIER (2015a). Ces deux articles traitent de la même construction (*A as GN*) sur deux corpus différents (le COCA dans le premier cas et le BNC dans le second) [9]. L'ACC est utilisée pour mesurer l'association entre les lexèmes qui instancient la position de l'adjectif et du groupe nominal. La méthode détermine pour chaque adjectif quels GN apparaissent dans la même construction plus souvent que ce à quoi on peut s'attendre. Les Tableaux 4.11 et 4.12 sont extraits des tableaux de sortie de l'ACC dans DESAGULIER (2015c) et DESAGULIER (2015a) respectivement [10].

Dans les deux corpus, on observe deux types de relations sémantiques entre les adjectifs et les GN parmi les appariements les plus associés (DESAGULIER, 2015c, p. 112) :

— une relation assez littérale, l'adjectif dénotant une propriété distinctive immédiate du GN (*cold as ice*, *smooth as silk*, *quick as a flash*) ;

— une relation plus lâche, plus imagée et parfois fantasque fondée sur une connotation du GN (*clear as a bell*, *old as the hills*, *right as rain*, *happy as a clam*) ou sur une correspondance phonétique (*good as gold*, *bold as brass*).

Dans tous les cas, les appariements les plus associés sont des occurrences conventionnalisées de *A as GN*. Ces appariements sont sujets à une très faible variation morpho-lexicale. Ils sont en soi peu productifs, au sens où d'autres schémas n'en sont pas nécessairement dérivés en grand nombre, et se comportent comme des unités symboliques complexes. La question est de savoir si ces réalisations sous-schématiques nous permettent de conclure quoi que ce soit

9. Cf. (35), Chapitre 2, p. 39)
10. Dans les deux cas, l'association est mesurée à l'aide du rapport de log-vraisemblance.

Tableau 4.11: Extrait d'un tableau de sortie de l'ACC : *A as GN* dans le COCA (DESAGULIER, 2015c, Tableau 7)

ADJ	NP	freq. A	freq NP	obs A_NP in C	exp A_NP in 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 4.12: Extrait d'un tableau de sortie de l'ACC : *A as GN* dans le BNC (DESAGULIER, à paraître, adapté du Tableau 9.18)

ADJ	NP	freq. A	freq NP	obs A_NP in C	exp A_NP in C	relation	coll.strength
good	*gold*	29	30	29	0.48	attraction	288.8138
quick	*flash*	27	20	20	0.3	attraction	189.2885
right	*rain*	20	20	18	0.22	attraction	175.9832
large	*life*	18	21	17	0.21	attraction	164.5468
safe	*houses*	17	14	14	0.13	attraction	148.3236
sure	*hell*	50	98	33	2.69	attraction	141.9763
old	*hills*	31	15	15	0.26	attraction	130.8729
pretty	*picture*	16	12	12	0.11	attraction	126.4332
bold	*brass*	12	10	10	0.07	attraction	113.2006
solid	*rock*	22	26	15	0.31	attraction	110.9541
...	...	...	...	...	...	...	...
sharp	*hell*	36	98	1	1.94	repulsion	0.5895
hard	*hell*	23	98	1	1.24	repulsion	0.0527

quant à la (non) productivité de *A as GN*. Dans DESAGULIER (2015a), j'explore l'hypothèse selon laquelle il existe un gradient de productivité inversement proportionnel au degré d'attraction entre l'adjectif et le GN (voir Section 4.7.2 plus bas).

L'apport de l'analyse collostructionnelle

En plus de proposer un cadre pour quantifier la coalescence des constituants constructionnels, la valeur ajoutée de l'analyse collostructionnelle est de proposer des mesures qui vont au-delà de ce qu'apportent les fréquences brutes. Pour chacune des méthodes, la corrélation est forte mais pas exacte entre les associations mesurées sur la base de fréquences brutes et les associations, comme illustré en Figure 4.3 [11].

Le graphe met en avant certains profils, notamment *good as gold* et *sure as hell*. Les profils de *cold as ice*, *white as snow*, *solid as a rock*, et *white as a sheet* se distinguent également de par leurs déviations respectives vis-à-vis de la droite de régression (en rouge). La fréquence brute ne permet de mesurer l'association entre l'adjectif et le GN qu'imparfaitement.

11. Voir également DESAGULIER (2015c) et DESAGULIER (à paraître, Figures 9.1 et 9.2).

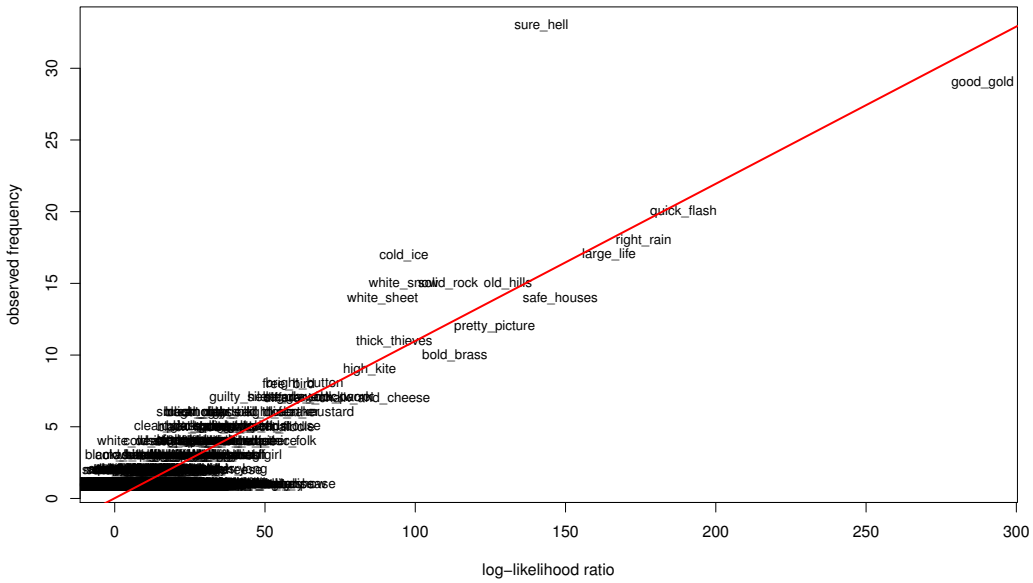


Figure 4.3: Comparaison de la fréquence observée et du score du rapport de log-vraisemblance en ACC pour *A as GN* dans le BNC (DESAGULIER, à paraître, Figure 9.3)

Les limites de l'analyse collostructionnelle

La diffusion relativement importante de l'analyse collostructionnelle en linguistique cognitive (et plus particulièrement en grammaires de constructions) ne doit pas en cacher les limites (DESAGULIER, 2015c, Section 2.3). La première concerne plus généralement l'analyse collocationnelle. Elle est formulée par FIRTH (1957, p. 181) lui-même : par leur recours à la distribution lexicale, les collocations ne suffisent pas à circonscrire exhaustivement le sens d'un mot.

Concernant plus spécifiquement l'analyse collostructionnelle, deux critiques peuvent être formulées [12]. Tout d'abord, les tableaux de sortie de chacune des méthodes présentées ci-dessus peuvent contenir jusqu'à plusieurs milliers de lignes (en fonction de l'étude de cas et de la taille du corpus). Leur synthèse à l'œil nu est impossible, ce qui crée un décalage avec les exigences d'une grammaire de l'usage censée rendre compte de l'expérience « statistique » des locuteurs, ces derniers parvenant sans mal à extraire des régularités dans la masse des données propre à la pratique langagière.

Par ailleurs, tout comme les mesures d'associations sur lesquelles elle s'appuie, l'analyse collostructionnelle ne rend pas compte des associations asymétriques (GRIES, 2013a). Or, l'attraction mutuelle entre lexèmes ou entre lexèmes et constructions est loin d'être systématiquement symétrique. Je développe ce point ci-dessous.

12. En plus de celles de SCHMID et KÜCHENHOFF (2013) auxquelles GRIES (2015) répond.

4.5.3 Associations asymétriques

Qu'il me soit permis ici un rapprochement entre les appariements lexicaux, constructionnels ou lexico-constructionnels et un couple dysfonctionnel. Dans un couple mal assorti, l'attraction d'un conjoint vis-à-vis de son/sa partenaire est par exemple plus forte que l'attraction ressentie par le/la partenaire vis-à-vis de lui [13]. Dans un appariement linguistique impliquant des lexèmes et des constructions, l'attraction n'est que rarement symétrique. À titre d'illustration, on devine que dans *ad hominem*, *hominem* attire plus *ad* que vice versa parce que *ad* intervient dans plus de types d'appariements différents que *hominem*. J'ai consacré deux articles à l'exploration de l'association asymétrique et à ses enjeux en matière de mesure de la productivité constructionnelle (DESAGULIER, 2015c ; DESAGULIER, 2015a).

L'idée selon laquelle les collocations sont directionnelles est mise en avant par GRIES (2013a). Elle a été inspirée par les travaux de ELLIS (2006) et ELLIS et FERREIRA-JUNIOR (2009) qui rejettent la théorie classique du conditionnement au profit d'une théorie renouvelée de l'apprentissage associatif centré sur l'idée de contingence. Je développe ce point plus en détail dans DESAGULIER (2015a, p. 18).

L'ACC réalisée dans l'étude de *A as GN* confond deux types de probabilités :

— la probabilité d'obtenir un adjectif donné sachant que l'on a un GN donné ;

— la probabilité d'obtenir un GN donné sachant que l'on a un adjectif donné.

Une mesure d'association inspirée des travaux de L. G. ALLAN (1980) permet de distinguer ces deux probabilités : ΔP (« delta P »). Décrite par ELLIS (2006) en acquisition du langage, ΔP s'appuie sur un tableau de contingence tel que le Tableau 4.13.

Tableau 4.13: Tableau d'entrée générique pour un événement impliquant un résultat (*O* : *outcome*) et un indice (*C* : *cue*)

	O	¬O
C	a	b
¬C	c	d

Le calcul de la dépendance directionnelle entre C et O se fait de la manière suivante :

$$\begin{aligned} \Delta P &= P(O|C) - P(O|\neg C) \\ &= \frac{a}{a+b} - \frac{c}{c+d} \end{aligned} \tag{4.1}$$

Plus ΔP est proche de 1, plus C augmente la probabilité de O. Inversement, plus ΔP est proche de -1, moins C augmente la probabilité de O. Si $\Delta P = 0$, il n'y a pas de covariation entre C et O.

GRIES (2013a) a montré que ΔP était applicable aux collocations. J'ai pour ma part transposé la mesure à l'étude des collostructions. Le tableau de contingence ci-dessous (Tableau 4.14).

13. Lors de mes communications, j'ai souvent recours à ce rapprochement. Ce que je nomme « the ill-matched-couple metaphor » me permet de faire comprendre le sens de l'opposition entre associations symétriques et asymétriques à un public pas nécessairement formé à l'analyse quantitative des collocations.

Tableau 4.14: Tableau de contingence pour le calcul de l'association directionnelle à l'œuvre dans *A as GN*

	GN : présent	GN : absent
A : présent	a	b
A : absent	c	d

Deux valeurs de ΔP doivent être calculées de manière à rendre compte à la fois de la probabilité d'obtenir un GN donné sachant que l'on a un adjectif donné (4.2) et de la la probabilité d'obtenir un adjectif donné sachant que l'on a un GN donné (4.3) :

$$\begin{aligned} \Delta P_{(GN|A)} &= p(GN|A) - p(GN|\neg A) \\ &= \frac{a}{a+b} - \frac{c}{c+d} \end{aligned} \tag{4.2}$$

$$\begin{aligned} \Delta P_{(A|GN)} &= p(A|GN) - p(A|\neg GN) \\ &= \frac{a}{a+c} - \frac{c}{b+d} \end{aligned} \tag{4.3}$$

Prenons un exemple : *mad as a March hare* dans le BNC. Les fréquences sont indiquées dans le Tableau 4.15.

Tableau 4.15: Tableau de contingence : fréquences de *mad* et *March hare* dans *mad as a March hare* (BNC)

	March hare : présent	*March hare* : absent
mad : présent	2	7
mad : absent	0	98363774

On calcule alors les deux valeurs de ΔP :

$$\begin{aligned} \Delta P_{(March\ hare|mad)} &= p(March\ hare|mad) - p(March\ hare|\neg mad) \\ &= \frac{2}{2+7} - \frac{0}{0+98363774} \\ &\approx 0.22 \end{aligned} \tag{4.4}$$

$$\begin{aligned} \Delta P_{(mad|March\ hare)} &= p(mad|March\ hare) - p(mad|\neg March\ hare) \\ &= \frac{2}{2+0} - \frac{0}{7+98363774} \\ &= 1. \end{aligned} \tag{4.5}$$

Si la différence $\Delta P_{(GN|A)} - \Delta P_{(A|GN)}$ est positive, alors l'adjectif est un meilleur prédicteur du GN que réciproquement. Inversement, si la différence est négative, alors le GN est un meilleur prédicteur de l'adjectif que réciproquement. Si la différence est nulle, ni l'adjectif ni le GN ne sont des prédicteurs l'un de l'autre.

$$\Delta P_{(March\ hare|mad)} - \Delta P_{(mad|March\ hare)} \approx -0.78 \tag{4.6}$$

Sans surprise (parce que l'exemple est, à dessein, extrême), *March hare* est un bien meilleur prédicteur de *mad* que réciproquement, la différence étant négative et proche de -1. La

raison en est que *mad* est employé dans d'autres types de *A as GN* (par exemple *mad as hell*, *mad as a hatter*).

Parmi les 1206 paires adjectif–GN de *A as GN* dans le BNC, 153 ne révèlent aucune covariation (DESAGULIER, 2015a, Figure 2). Les autres paires affichent une attraction asymétrique. L'asymétrie est liée au fait qu'un constituant constructionnel est impliqué dans d'autres sous-schémas de *A as GN*. Le Tableau 4.16 compare les dix valeurs de ΔP les plus proches de -1 et de 1.

Tableau 4.16: Comparaison des 20 valeurs asymétriques les plus extrêmes

A	GN	G^2	$\Delta P_{NP\|A} - \Delta P_{A\|NP}$
white	*ash*	6.6782	-0.984614734
white	*candlewax*	6.6782	-0.984614734
white	*desert*	6.6782	-0.984614734
white	*fire*	6.6782	-0.984614734
white	*flour*	6.6782	-0.984614734
white	*jellyfish*	6.6782	-0.984614734
white	*office paper*	6.6782	-0.984614734
white	*quicklime*	6.6782	-0.984614734
white	*snow* + postmod.	6.6782	-0.984614734
white	*towel* + postmod.	6.6782	-0.984614734
...	...	...	...
awkward	*hell*	5.8518	0.989794932
bleary	*hell*	5.8518	0.989794932
conservative	*hell*	5.8518	0.989794932
depressed	*hell*	5.8518	0.989794932
difficult	*hell*	5.8518	0.989794932
frustrated	*hell*	5.8518	0.989794932
gloomy	*hell*	5.8518	0.989794932
lonesome	*hell*	5.8518	0.989794932
nosy	*hell*	5.8518	0.989794932
nutty	*hell*	5.8518	0.989794932

Deux sous-schémas productifs (au sens défini plus haut) s'en dégagent nettement : *white as GN* et *A as hell*[14]. Ces deux constructions ont d'ailleurs un score de G^2 relativement faible, ce qui suggère une conventionnalisation moins grande. Si ΔP ne supplante pas à elle seule les autres mesures de productivité traditionnelles (cf. Section 4.7.2), elles permettent de mettre en avant des îlots de productivité à un niveau sous-schématique d'analyse constructionnelle.

4.6 Les statistiques exploratoires

Je reviens ici sur un problème mentionné plus haut (p. 73), à savoir la grande taille des tableaux de sortie de l'analyse collostructionnelle, qui rend leur synthèse difficile. De fait, la plupart des études sur les collocations et les collostructions se contentent d'analyser les valeurs les plus extrêmes, en général les attractions les plus fortes. Dans un premier temps,

14. Si l'on prend en compte les 300 sous-schémas dont les valeurs sont les plus proches de -1, on trouve les types suivants : *white/black/red/pale/clear/big/stiff/hard/heavy/solid/strong/smooth/soft/sharp/thick/dry/old/quick/happy/sick/dead/hot/cold/sure as NP*. La productivité de ces types est indexée sur l'adjectif. Les sous-schémas dont les valeurs sont les plus proches de 1 sont pour la plupart indexés sur le nom *hell*.

j'ai eu recours aux méthodes statistiques exploratoires dites « de clustering » pour synthétiser des tableaux issus de l'analyse collostructionnelle. Dans un second temps, j'ai utilisé ces méthodes de clustering dans une optique multifactorielle, en intégrant ponctuellement des mesures d'association à d'autres types de mesures.

4.6.1 Définition

Le clustering fait appel à un ensemble de techniques exploratoires non supervisées. Ces techniques ont en commun d'aider le chercheur à trouver une structure cohérente dans des données multivariées.

Sous-jacente au « cognitive commitment » (LAKOFF, 1990) est l'idée que la faculté de langage n'est pas modulaire et que ses manifestations linguistiques sont le produit de l'interaction simultanée de plusieurs facteurs. Pour rendre compte de cette multifactorialité, il n'est rien de plus naturel en linguistique cognitive que de rendre compte de faits linguistiques en caractérisant chaque observation par plusieurs types de variables (DESAGULIER, 2012b). Les observations sont ainsi réunies dans un tableau à plusieurs dimensions dans lequel se jouent des corrélations à plusieurs niveaux. L'objectif des méthodes statistiques exploratoires est de mettre en avant les corrélations les plus significatives.

Pour faire émerger une structure cohérente à partir d'un jeu de données multivarié, les méthodes de clustering font appel à des outils de visualisation dont j'explique intuitivement la logique dans DESAGULIER (à paraître, Section 10.1.2). La linguistique quantitative contemporaine étant confrontée à des jeux de données de plus en plus grands (tant en nombre d'observations qu'en nombre de variables), réfléchir à leur synthèse et à leur visualisation est devenu incontournable.

Je présente quatre méthodes ci-dessous : la classification ascendante hiérarchique, l'analyse factorielle des correspondances, l'analyse des correspondances multiples et l'analyse en composantes principales. Dans cette synthèse, je n'aborde pas en détail des statistiques sur lesquelles ces méthodes s'appuient et renvoie le lecteur à DESAGULIER (2014) et DESAGULIER (2015b) ainsi qu'à DESAGULIER (à paraître, Chapitre 10). Je mets ici l'accent sur l'application de ces méthodes et leur apport à la démarche linguistique, notamment leur faculté à s'approcher d'une appréhension empirique de l'idée de réseau constructionnel.

4.6.2 La classification ascendante hiérarchique

La classification ascendante hiérarchique (CAH) est une méthode multifactorielle employée pour résumer et visualiser des tableaux de contingence (DESAGULIER, à paraître, Section 10.6) [15]. La méthode convertit le tableau de contingences en une matrice de distances, à savoir un tableau symétrique semblable aux distances kilométriques entre les villes dans les atlas routiers [16].

15. Dans la pratique, la CAH fonctionne également avec d'autres types de données.
16. Cf. DESAGULIER (à paraître, Tableau 10.7)

Dans DESAGULIER (2014), je compare 23 intensifieurs extraits sur la base des 35 collexèmes adjectivaux avec lesquels ils sont le plus associés dans le COCA. Le tableau de contingence regroupe les fréquences de 9936 types de constructions <intensifieur + adjectif>. Ce tableau est ensuite converti en matrice de distances à l'aide d'une métrique de distance, en l'occurrence la mesure Canberra puisqu'elle est particulièrement adaptée aux tableaux dont bon nombre de cellules ont des fréquences nulles [17]. La matrice est ensuite amalgamée à l'aide de l'algorithme de Ward (WARD, 1963), qui évalue les distances entre clusters à l'aide d'une analyse de variance. Le résultat graphique est projeté sous la forme du dendrogramme en Figure 4.4.

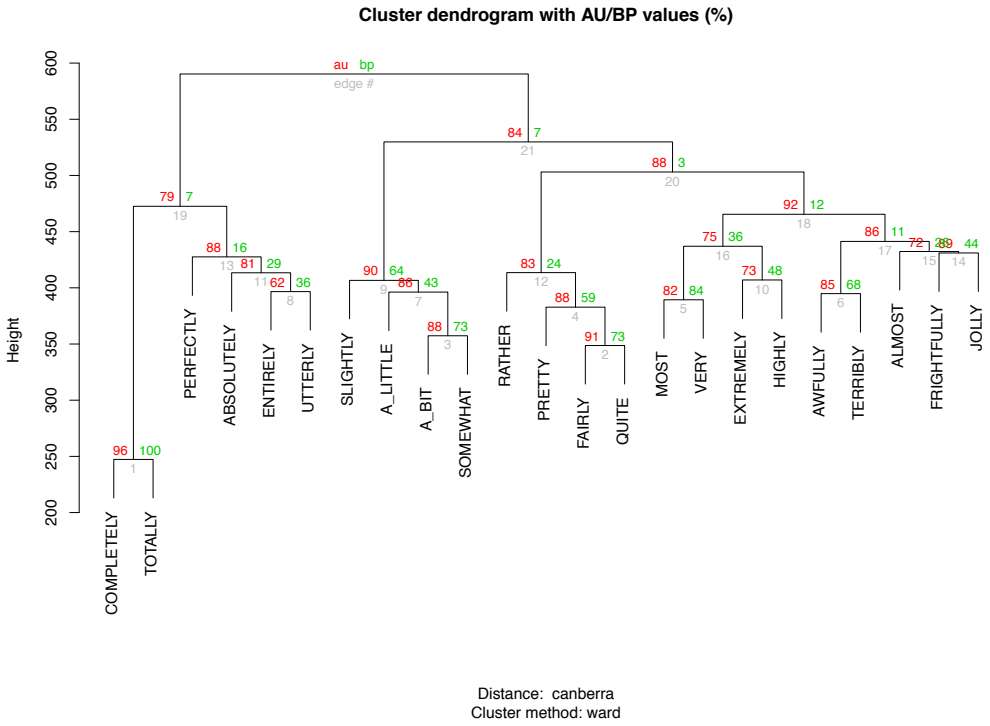


Figure 4.4: Graphe issu de la CAH (DESAGULIER, 2014, Figure 1)

La classification hiérarchique étant ascendante, le dendrogramme se lit du bas vers le haut. Trois nombres sont associés à chaque cluster :

— le nombre sous chaque nœud indique le rang du cluster (de 1 à 21) ;

— les deux nombres au dessus de chaque nœud sont des indications numériques quant à la qualité du cluster :

  — AU : « approximately unbiased p-value » ;

  — BP : « bootstrap probability ».

Le nombre de gauche (AU) est considéré comme plus fiable que le nombre de droite. Dans l'ensemble, le dendrogramme contient des clusters homogènes puisque l'on retrouve, de gauche à droite, les principales catégories reconnues dans la littérature linguistique sur les intensifieurs (PARADIS, 1997) :

— « maximizers » : *completely*, *totally*, *perfectly*, *absolutely*, *entirely* et *utterly* ;

— « diminishers » : *slighlty*, *a little*, *a bit* et *somewhat* ;

17. Cf. EVERITT et al. (2011, Section 4.2).

— « moderators » : *rather*, *pretty*, *fairly* et *quite* ;

— « boosters » : *most*, *very*, *extremely*, *highly*, *awfully*, *terribly*, *frightfully* et *jolly*.

La présence de *almost* est surprenante mais probablement due à son profil particulier (c'est un adverbe de phrase avant d'être un intensifieur d'adjectif). Quoi qu'il en soit, la CAH produit des clusters relativement cohérents en dépit de la simplicité du type de données (des données de co-occurrence).

4.6.3 L'analyse factorielle des correspondances

Tout comme la CAH, l'analyse factorielle des correspondances (AFC) est une méthode multifactorielle employée pour résumer et visualiser des tableaux de contingence (DESAGULIER, à paraître, Section 10.4). À la différence de la CAH, l'AFC permet de visualiser les distances entre les observations (les lignes du tableau) et les variables descriptives (les colonnes).

Dans DESAGULIER (2014), j'ai rassemblé les fréquences des 25 collexèmes les plus distinctifs de *quite*, *rather*, *fairly* et *pretty* dans un tableau de contingence, dont le Tableau 4.17 présente un aperçu (le tableau total contient 400 cellules).

Tableau 4.17: Extrait du tableau d'entrée pour l'AFC (DESAGULIER, 2014, Tableau 8)

adjective (distinctive collexeme)	*fairly*	*pretty*	*quite*	*rather*
able	3	4	67	0
abstract	14	7	5	22
accurate	83	66	109	4
amazing	6	347	97	22
aware	1	11	109	0
awful	0	100	12	6
awkward	0	12	7	27
bad	19	758	32	22
beautiful	1	2	129	21
big	56	591	45	23
...	...	...	...	...

L'AFC se fonde sur ces fréquences pour comparer les profils :

— des lignes entre elles (les adjectifs) ;

— des colonnes entre elles (les adverbes) ;

— des lignes et des colonnes.

L'AFC convertit chaque observation et chaque variable en un point. Chaque point dispose de coordonnées euclidiennes, et peut être projeté sur un plan à deux dimensions. Dans la mesure où il y a plus de deux dimensions, l'AFC les compare deux par deux. Le graphe représenté en Figure 4.5 compare les deux premières dimensions, qui regroupent à elles seules $52.77\% + 28.14\% = 80.91\%$ de la variance du tableau.

Plus deux points sont proches sur le plan euclidien, plus leurs profils respectifs sont similaires [18]. Sur l'axe horizontal (dimension 1), *pretty* (à gauche) et *quite* (à droite) s'opposent.

18. Il faut ajouter un bémol toutefois. Lorsque l'on regarde le ciel par une nuit étoilées, deux étoiles peuvent nous sembler proches alors qu'elles sont situées l'une dernière l'autre à plusieurs années lumières de distance. Toute proportion gardée, la même logique peut être à l'œuvre dans un graphe d'AFC. Pour s'en prémunir, il faut revenir régulièrement aux données brutes.

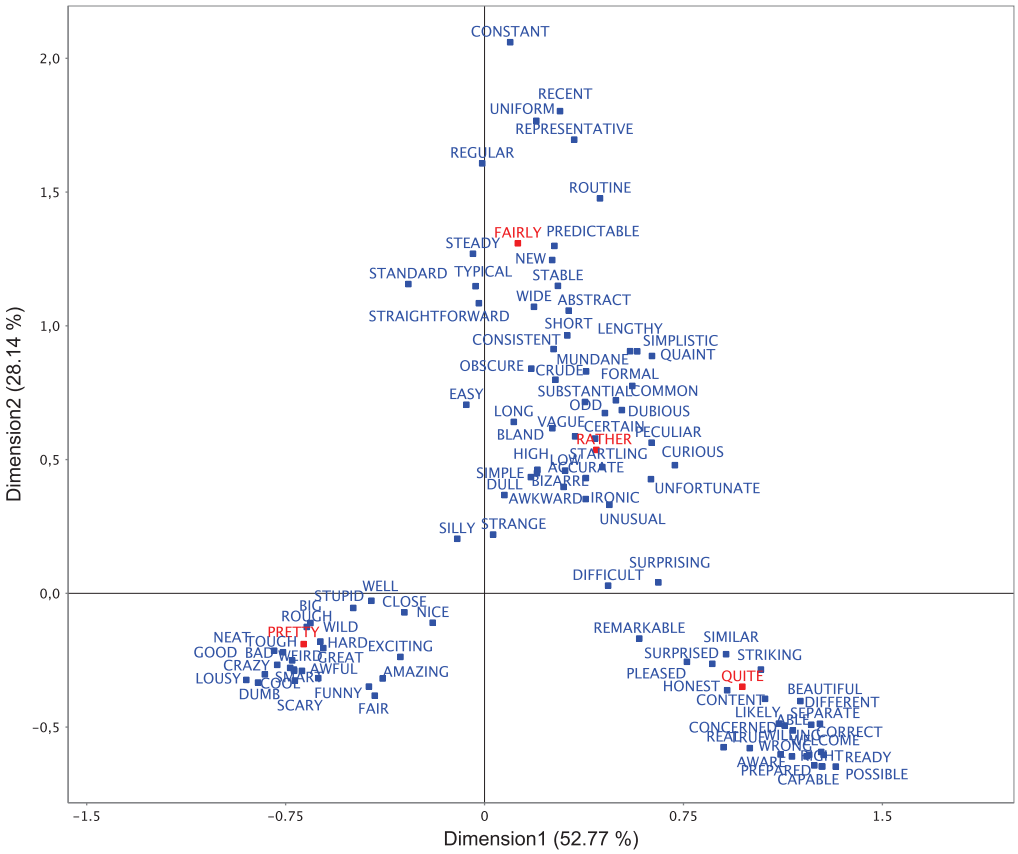


Figure 4.5: Graphe issu de l'AFC (DESAGULIER, 2014, Figure 2)

Les nuages d'adjectifs entourant chacun d'eux ne se recoupent pas. L'axe vertical (dimension 2) oppose *fairly* et *rather* (en haut) à *pretty* et *quite* (en bas). Notons que cette fois les nuages d'adjectifs respectifs de ces deux adverbes forment un continuum. Parce qu'il repose sur des données, ce clustering apporte un regard nouveau sur la classification des intensifieurs, les grammaires traditionnelles regroupant sémantiquement plutôt *pretty* et *fairly* d'un côté et *rather* et *quite* de l'autre (DOWNING et LOCKE, 2006). Je propose une interprétation sémantique détaillée dans l'article (DESAGULIER, 2014, p. 172). Il se dégage de l'analyse que chaque adverbe opère bien à la fois sur des domaines conceptuels spécifiques et que plusieurs adverbes se partagent souvent l'intensification de deux aspects complémentaires d'un même domaine conceptuel. Par exemple, *rather* intervient de préférence sur le domaine conceptuel de l'atypicalité (*odd*, *bizarre*), tandis que *fairly* intervient de préférence sur le domaine conceptuel de la typicalité (*typical*, *common*, *standard*).

Dans ce que je viens de résumer, je fais une utilisation monofactorielle d'une méthode multifactorielle. En effet, les variables du Tableau 4.17 sont d'un seul type (les quatre colonnes correspondent aux quatre intensifieurs étudiés). Mon utilisation de l'AFC ne devient véritablement multifactorielle qu'en intégrant des variables sémantiques dans DESAGULIER (2015b). Le Tableau 4.18 regroupe les fréquences des adjectifs intervenant dans l'étude comparative de *quite* et *rather*. Les observations sont de deux types (les adjectifs et leurs classes sémantiques). Les variables sont également de plusieurs types (syntaxe de l'adverbe à différents niveaux de granularité, syntaxe de l'adjectif, mode). Dans cette étude, les adjectifs ne sont plus sélectionnés sur la base du score censé mesurer leur attraction. Une confiance

est accordée à l'AFC vis-à-vis de son traitement des fréquences brutes puisque le nombre d'occurrences de chaque observation est ramené aux sommes marginales du tableau.

Tableau 4.18: Extrait du tableau d'entrée pour l'AFC (DESAGULIER, 2015b, Tableau 4)

	X_QUITE_ADJ_X	X_QUITE_ADJ_NP	X_RATHER_ADJ_X	X_RATHER_ADJ_NP	DET_QUITE_ADJ_NP	DET_RATHER_ADJ_NP	QUITE_DET_ADJ_NP	RATHER_DET_ADJ_NP	PRE_DETERMINER_POSITION	PRE_ADJECTIVAL_POSITION	WITH_ATTRIBUTIVE_ADJ	WITH_PREDICATIVE_ADJ	SPOKEN	WRITTEN	QUITE_CONSTRUCTIONS	RATHER_CONSTRUCTIONS
different	770	290	213	97	174	160	136	8	144	1704	865	983	116	1732	1370	478
difficult	197	8	92	4	1	13	14	4	18	315	44	289	109	224	220	113
nice	301	13	68	4	3	10	72	7	79	399	109	369	356	122	389	89
good	509	70	64	11	6	17	205	25	230	677	334	573	550	357	790	117
surprised	47	0	45	3	0	0	0	0	0	95	3	92	34	61	47	48
small	137	46	43	11	6	17	33	7	40	260	120	180	39	261	222	78
similar	47	4	40	15	0	17	0	0	0	123	36	87	10	113	51	72
strange	25	0	38	9	1	20	1	2	3	93	33	63	23	73	27	69
vague	4	2	35	4	0	12	0	2	2	57	20	39	5	54	6	53
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
TOTAL (active rows)	*12551*	*1313*	*2953*	*774*	*584*	*1509*	*1639*	*285*	*1924*	*19684*	*6104*	*15504*	*5229*	*16379*	*16087*	*5521*
DIFFERENCE_contrast	1022	367	218	99	198	162	151	8	159	2066	985	1240	128	2097	1738	487
DIFFICULTY_complexity	364	54	160	15	18	55	46	9	55	666	197	524	215	506	482	239
DIMENSION_POSITION	966	291	343	126	55	192	642	79	721	1973	1385	1309	692	2002	1954	740
DISCOMFORT	16	1	31	11	0	34	5	0	5	93	51	47	17	81	22	76
DULLNESS	59	3	115	39	1	64	2	4	6	281	113	174	26	261	65	222
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
TOTAL (sup. rows)	*12551*	*1313*	*2953*	*774*	*584*	*1509*	*1639*	*285*	*1924*	*19684*	*6104*	*15504*	*5229*	*16379*	*16087*	*5521*

Les variables grisées sont considérées par l'AFC comme illustratives ou supplémentaires. Elles n'interviennent pas dans le calcul de la forme du nuage. Les points supplémentaires sont projetés sur le nuage après que les points actifs ont été placés (les variables actives sont sur fond blanc). Le tableau contient :

— 9632 cellules ;

— 543 lignes actives (les adjectifs) ;

— 59 lignes supplémentaires (les classes sémantiques des adjectifs) ;

— 8 colonnes actives (les types de constructions en *quite* et *rather*) ;

— 8 colonnes supplémentaires.

Le graphe en figure 4.6 est la visualisation graphique du tableau par l'AFC. Afin de ne pas encombrer le graphe, seules les observations et les variables supplémentaires sont représentées.

Ce graphe permet de repérer la répartition des tâches conceptuelles entre *quite* et *rather*, résumée dans le Tableau 4.19.

4.6.4 L'analyse des correspondances multiples

Comme son nom l'indique, l'analyse des correspondances multiples (ACM) se situe dans le prolongement de l'AFC. Alors que l'AFC résume tableau de contingence, l'ACM résume un tableau de données où les observations sont décrites par des variables qualitatives. Tout comme l'AFC, l'ACM étudie les ressemblances entre les observations du point de vue de l'ensemble des variables et dégage des profils d'observations (DESAGULIER, à paraître, Section 10.5).

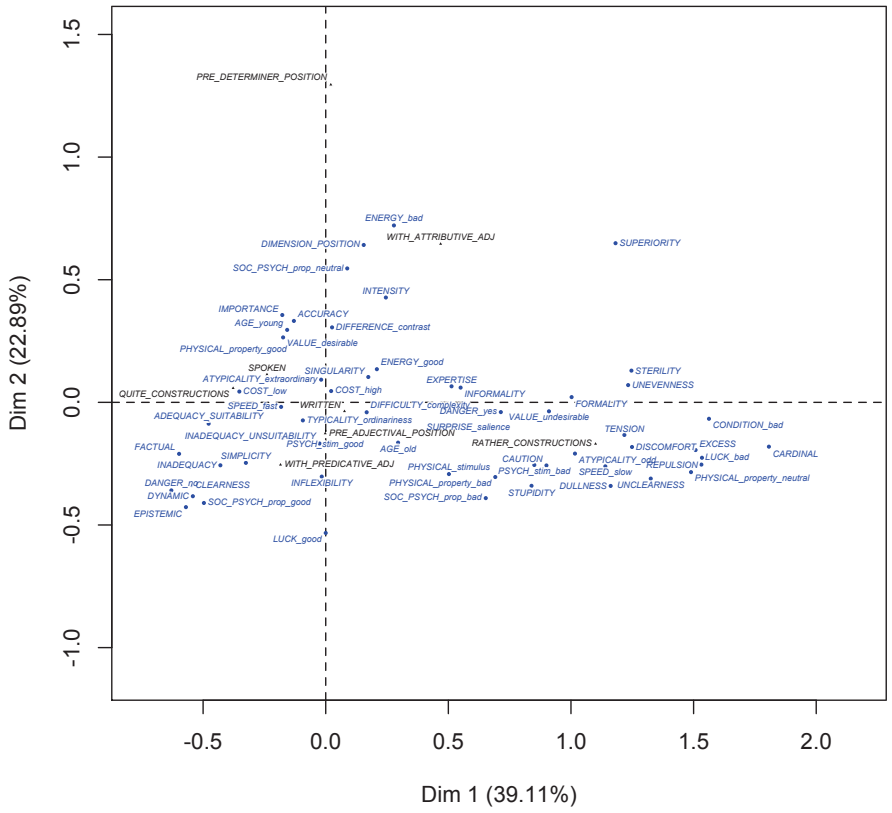


Figure 4.6: Graphe issu de l'AFC (DESAGULIER, 2015b, Figure 2)

Tableau 4.19: Répartition des tâches conceptuelles entre *quite* et *rather* dans le BNC (DESAGULIER, 2015b, Tableau 6)

constructions en *quite*	constructions en *rather*
good luck (*lucky*)	bad luck (*unfortunate*)
atypicality (positive) (*outstanding*, *breathtaking*)	atypicality (negative) (*bizarre*, *unothodox*, *puzzling*)
psychological stimuli (positive) (*exciting*, *interesting*, *moving*)	psychological stimuli (negative) (*disturbing*, *worrying*, *confusing*)
absence of danger (*harmless*, *safe*)	presence of danger (*threatening*, *dangerous*)
high speed (*quick*, *rapid*)	low speed (*slow*, *slower*)
desirable value (*good*, *neat*, *nice*, *perfect*)	undesirable value (*bad*, *poor*, *negative*, *nasty*)

Dans DESAGULIER (2015b), je procède à deux ACM sur la base du Tableau 4.20. Je présente ici la seconde ACM. Toutes les variables sont actives à l'exception de `sem_class`. Le jeu de données comprend 3086 observations, cinq variables contextuelles issues des étiquettes du `BNC` et une variable sémantique issue d'une annotation manuelle. Afin de rester dans le strict cadre de l'alternance entre syntaxe pré-adjectivale et syntaxe pré-déterminantale, seules les occurrences de *quite* et *rather* correspondant aux schémas suivants ont été conservées : *<a(n) quite/rather A NP>* et *<quite/rather a(n) A NP>*. Les adjectifs attributs ont été écartés pour la même raison.

Tableau 4.20: Extrait du tableau d'entrée pour l'ACM (DESAGULIER, 2015b, Tableau 7)

construction	intensifier	text_mode	text_type	text_info	sem_class
PREDETERMINER	QUITE	SPOKEN	OTHERSP	S pub debate	psych_stim_good
PREADJECTIVAL	QUITE	WRITTEN	FICTION	W fict prose	factual
PREADJECTIVAL	RATHER	WRITTEN	FICTION	W fict prose	dullness
PREADJECTIVAL	RATHER	WRITTEN	FICTION	W fict prose	atypicality_odd
PREADJECTIVAL	QUITE	SPOKEN	OTHERSP	S meeting	importance
PREADJECTIVAL	RATHER	WRITTEN	NONAC	W religion	difficulty_complexity
PREADJECTIVAL	RATHER	WRITTEN	NEWS	W newsp other : social	singularity
PREADJECTIVAL	QUITE	WRITTEN	NONAC	W biography	factual
PREDETERMINER	QUITE	SPOKEN	OTHERSP	S lect soc science	age_young
PREADJECTIVAL	RATHER	WRITTEN	NONAC	W nonAc : nat science	simplicity
…	…	…	…	…	…

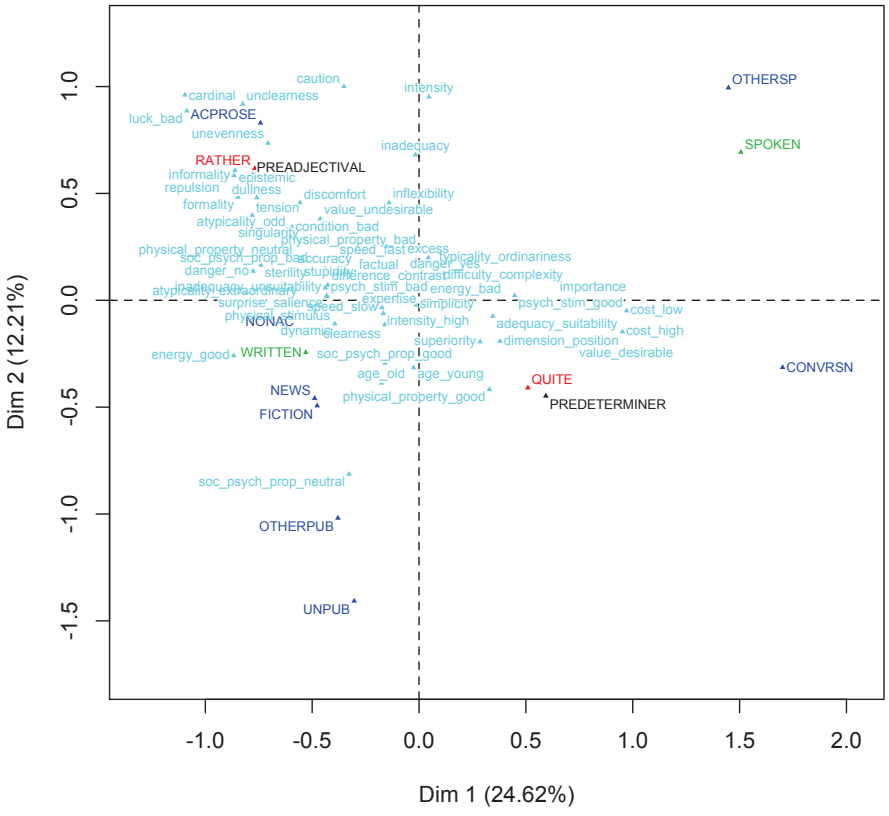


Figure 4.7: Graphe issu de l'ACM (DESAGULIER, 2015b, Figure 4)

Le graphe de la Figure 4.7 est la transposition graphique de l'ACM. Le graphe permet de saisir simultanément les profils de *quite* et *rather* sous le double angle de la forme et du sens. L'axe horizontal (dimension 1) distingue nettement *rather* (à gauche) et *quite* (à droite). Le premier apparaît de préférence en position pré-adjectivale tandis que le second apparaît de préférence en position pré-déterminantale. Toujours sur le même axe, il apparaît que *rather* apparaît de préférence dans les contextes écrits tandis que *quite* apparaît dans les contextes oraux. Cette tendance s'inverse sur l'axe vertical (dimension 2), ce qui amène à tempérer les préférences contextuelles des deux constructions [19]. À quelques exceptions près, on observe une répartition des tâches entre *quite* et *rather*. Par exemple, *quite* en position pré-déterminantale intensifie de préférence les adjectifs à connotations positives (`value_desirable`, `psych_stim_good`, `physical_property_good`). Inversement, *rather* en position pré-adjectivale intensifie de préférence les adjectifs à connotations négatives (`value_undesirable`, `psych_stim_bad`, `physical_property_bad`) [20].

4.6.5 L'analyse en composantes principales

La dernière méthode exploratoire dont je rends compte ici est l'analyse en composantes principales (ACP). Historiquement, cette méthode a vu le jour avant l'AFC et l'ACM et a d'ailleurs inspiré celles-ci à la fois dans leurs logiques et leurs fondements mathématiques (DESAGULIER, à paraître, Section 10.2). L'ACP est particulièrement adaptée à la synthèse de tableaux contenant des données continues et des données catégorielles/nominales [21]. Elle peut également s'accommoder de données discrètes.

Dans DESAGULIER (2015a), le Tableau 4.21 me sert à évaluer la productivité de *A as NP* au niveau sous-schématique en comparant différentes mesures de productivité appliquées aux deux places instanciables de la construction :

— V : la fréquence de type ;

— $V1$: le nombre d'hapax legomena ;

— $\mathscr{P}$: la productivité potentielle ;

— P^* : la productivité globale ;

— la fréquence brute ;

— le score moyen de G^2 ;

— la différence moyenne $\Delta P_{GN|A} - \Delta P_{A|GN}$ (cf. p. 75).

$\mathscr{P}$ et P^* sont deux mesures développées et décrites par BAAYEN (1989), BAAYEN et LIEBER (1991) et BAAYEN (1993). $\mathscr{P}$ mesure la productivité d'une construction (par exemple *A as hell* ou *white as NP*) en divisant le nombre d'hapax legomena générés par ladite construction par la taille du corpus. Il s'agit de la probabilité d'obtenir de nouveaux types. P^* mesure la productivité globale d'une construction en divisant V par $\mathscr{P}$ (DESAGULIER, 2015a, p. 12).

19. Cependant, la deuxième dimension rend compte d'une moindre variance que la première dimension. On peut donc parler de préférences et non de tendances tranchées.

20. Cette tendance est plus évidente dans la première ACM de l'article (DESAGULIER, 2015b, Figure 3).

21. Pour cette raison, c'est une méthode que j'utilise systématiquement pour traiter des données issues de tests psycho-cliniques avec les étudiants du parcours Diapason (Master FLDL) à l'Université Paris Ouest Nanterre. Deux travaux que j'ai encadrés, portant sur le diagnostic de la maladie d'Alzheimer d'un point de vue linguistique, ont exploité l'ACP pour classer des profils de patients en fonction de leurs résultats aux tests. L'un de ces mémoires peut être téléchargé à l'adresse suivante : `https://masterdiapason.wikispaces.com/file/view/Me%CC%81moire+Fre%CC%81de%CC%81rique+Gayet.pdf`.

Tableau 4.21: Extrait du tableau d'entrée pour l'ACP (DESAGULIER, 2015b, Tableau 7)

lexème	categorie	V	$V1$	$\mathscr{P}$	global_prod	const_freq	coll.strength	ΔP_diff
hell	NP	98	30	0.016492578	5942.067	2.227273	9.320452	0.642846875
white	A	65	17	0.009345794	6955.000	2.500000	13.394865	-0.681717831
black	A	50	19	0.010445300	4786.842	1.785714	10.090600	-0.776289082
sure	A	50	4	0.002199010	22737.500	4.545455	22.037282	-0.653766497
cold.adj	A	49	23	0.012644310	3875.261	1.814815	9.483585	-0.659506401
big	A	41	28	0.015393073	2663.536	1.205882	8.172321	-0.807597642
bright	A	39	23	0.012644310	3084.391	1.500000	9.623788	-0.737820145
clear	A	39	13	0.007146784	5457.000	2.052632	11.297047	-0.646846024
sharp	A	36	24	0.013194063	2728.500	1.285714	8.989675	-0.791237137
strong	A	32	19	0.010445300	3063.579	1.391304	9.254965	-0.703121225
thick	A	32	17	0.009345794	3424.000	1.600000	11.986035	-0.792360812
old	A	31	7	0.003848268	8055.571	2.818182	21.950418	-0.662587155
gold	NP	30	1	0.000549753	54570.000	15.000000	145.029850	0.031249924
good	A	29	0	0.000000000	NA	29.000000	288.813800	0.033333323
…	…	…	…	…	…			

Il y a 1278 observations (402 types d'adjectifs et 876 types de GN). Quatre variables sont actives : $\mathscr{P}$, P^*, coll.strength et ΔP_diff. Les trois autres variables sont illustratives. L'hypothèse de recherche est la suivante : *A as GN* est une construction productive au niveau sous-schématique et ne peut pas être reléguée au rang de simple patron innovant, contrairement à ce que postule KAY (2013) (cf. Chapitre 2, p. 39).

L'ACP génère deux graphes : le graphe des variables (Figure 4.8) et le graphe des observations (Figure 4.9). Je propose une analyse détaillée de ces deux graphes dans l'article (DESAGULIER, 2015a, p. 34–38).

Non pas une mais trois strates de productivité se dégagent sur le graphe des observations. La première, parallèle à l'axe vertical, correspond aux sous-schémas potentiellement productifs (ex. *A as hell*). La seconde, parallèle à l'axe horizontal et située au dessus de celui-ci, correspond aux sous-schémas globalement productifs (ex. *A as gold*). La troisième, parallèle à l'axe horizontal et située au dessous de celui-ci, correspond aux sous-schémas peu productifs car conventionnels et caractérisés par une association forte avec l'adjectif ou le GN.

4.7 Les statistiques prédictives

Les clusters qui émergent grâce à une méthode exploratoire sont les points de convergence d'une multitude de corrélations entre variables. Aussi utiles soient-ils sur le plan de la description, ils n'ont pas de valeur prédictive.

4.7.1 Définition

Il existe plusieurs types de méthodes statistiques prédictives. J'aborde tout d'abord la prédiction dans une optique monofactorielle.

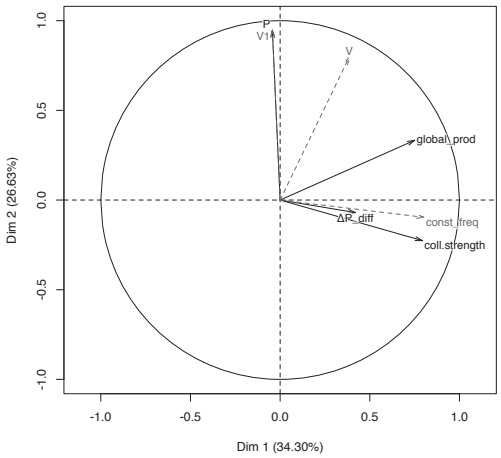


Figure 4.8: ACP – 4 variables actives (flèches pleines) and 3 variables supplementaires (flèches en pointillés) sur les dimensions 1 & 2

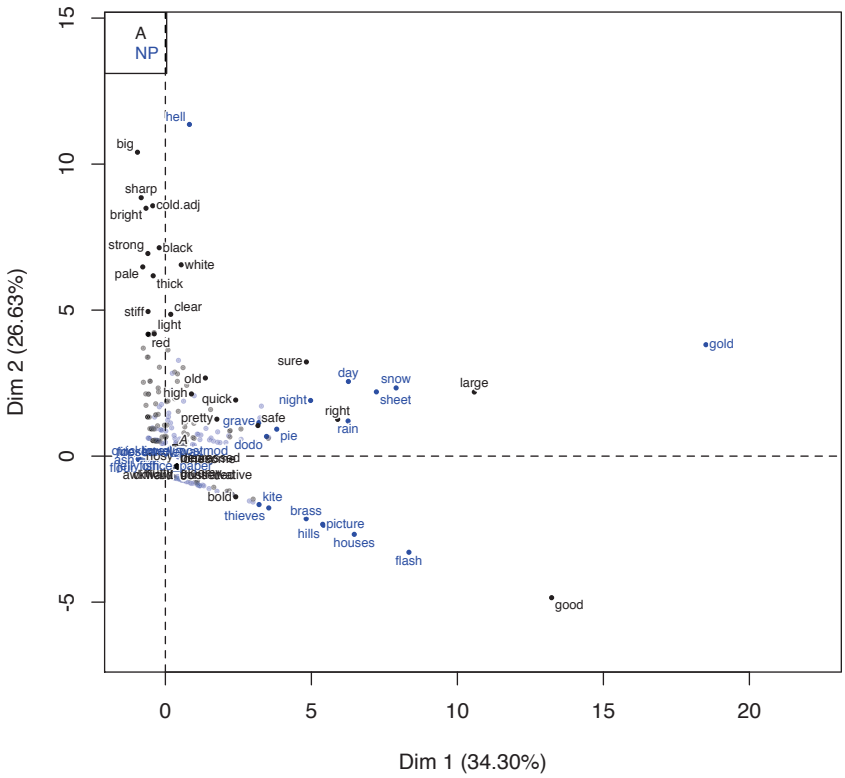


Figure 4.9: ACP – représentation plane des observations sur les dimensions 1 & 2

4.7.2 Les courbes de croissance lexicale et la productivité constructionnelle

Au Chapitre 2 (p. 39) j'ai présenté ma réponse théorique à la définition radicale de la productivité constructionnelle proposée par KAY (2013). Je présente ici un point clé de ma réponse empirique : les courbes de croissance lexicale (DESAGULIER, 2015a) [22].

C'est de la morphologie que sont issues la plupart des mesures de productivité en linguistique. La plus évidente de ces mesures est ce que BAAYEN (2009) nomme « la productivité réalisée », ou $V(C, N)$, à savoir la fréquence de type V des membres d'une catégorie morphologique C dans un corpus de N occurrences. Baayen reproche à cette mesure de ne pas distinguer les formes établies des formes nouvelles. En effet, les types ne sont pas distribués de manière uniforme dans un corpus. Le *type-token ratio*, une autre mesure très populaire, souffre du même défaut (DESAGULIER, à paraître, p. 9.4.2). Comparer les scores de *type-token ratio* dans deux corpus n'a de sens que si les corpus sont de même taille.

Pour qu'une mesure puisse capturer l'évolution de la distribution d'une forme linguistique dans un corpus, il faut qu'elle mesure le rapport des types et des tokens à intervalles réguliers. Le résultat est projeté sur une courbe de croissance lexicale (BAAYEN, 1993). Les tokens sont en abscisse et les types en ordonnées. Au début, la courbe croît rapidement puisque chaque token définit un type nouveau. Plus on avance dans le corpus et plus la courbe s'aplatit car les tokens se répètent et intègrent des types déjà vus. Plus la pente de la courbe est abrupte, plus l'unité linguistique étudiée est productive. La mesure $\mathscr{P}$, mentionnée ci-dessus est la pente de la tangente d'une courbe de croissance lexicale à son extrémité. Elle indique le taux de croissance du vocabulaire à lorsque la courbe s'interrompt.

La courbe en Figure 4.10 croise les tokens et les types de *A as GN* dans le BNC. Le corpus contient 1819 occurrences de la construction et 1316 types. La pente de la courbe tout au long du corpus indique une productivité relativement élevée.

Faisant référence à la loi de Zipf (ZIPF, 1949), Baayen part du principe que la productivité est fonction d'un grand nombre d'événements rares et d'un petit nombre d'événements fréquents (cf. Section 3.5.2). Nous savons que le plus grand corpus du monde ne contient pas tous les types possibles. Dénombrer ces types dans un corpus ne suffit donc pas à estimer la distribution de ces types dans une langue à partir d'un échantillon. Pour répondre à ce problème, BAAYEN (2001) se fonde sur des modèles LNRE (*Large Number of Rare Events*). La méthode qu'il propose comporte plusieurs étapes. Premièrement, il faut réaliser un spectre de fréquences (un tableau de fréquences de fréquences). Ensuite, il faut ajuster un modèle sur le spectre de fréquences. EVERT et BARONI (2006) et EVERT et BARONI (2007) proposent trois modèles :

— Zipf-Mandelbrot (ZM) ;
— finite Zipf-Mandelbrot (fZM) ;
— Generalized Inverse Gauss-Poisson (GIGP).

Lorsqu'un modèle satisfaisant a été ajusté, les valeurs attendues de l'unité étudiée sont calculées à partir des fréquences empiriques. Ces valeurs attendues permettent de lisser

22. Voir également (DESAGULIER, à paraître, Section 9.4.3).

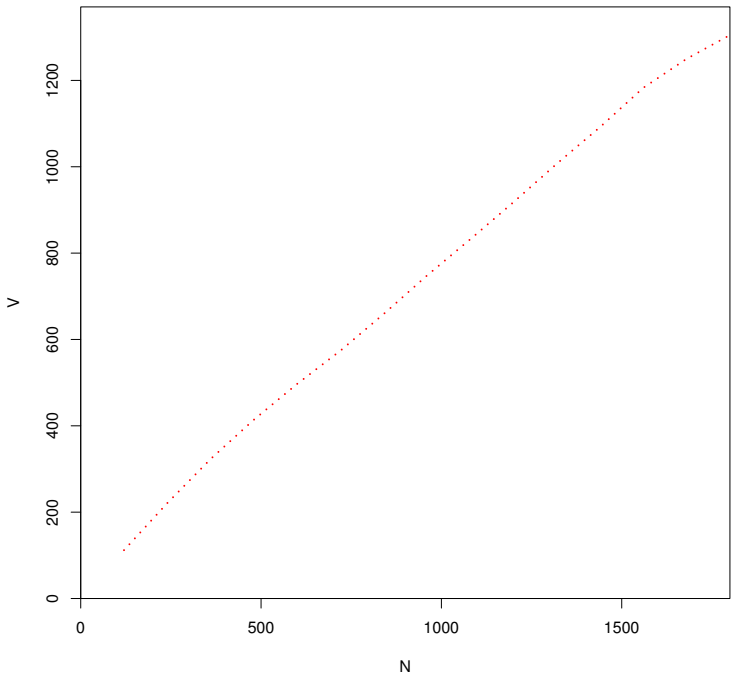


Figure 4.10: Courbe de croissance lexicale : *A as GN* dans le BNC

la courbe empirique sur la longueur du corpus. On obtient alors une courbe théorique interpolée. Les valeurs attendues permettent également de projeter la courbe au-delà de l'échantillon. On obtient alors une courbe théorique extrapolée. Cette dernière permet de prédire l'évolution de la courbe empirique au-delà de l'échantillon.

La Figure 4.11 compare les courbes de croissance lexicale de *A as GN* à quatre niveaux :

— niveau 1 : les occurrences exactes de *A as GN* ;

— niveau 2 : les paires A-GN ;

— niveau 3 : les adjectifs ;

— niveau 4 : les GN.

Pour chaque niveau, on mesure deux fréquences :

— la fréquence de type (V) ;

— la fréquence d'hapax ($V1$).

Pour chaque fréquence, trois courbes sont projetées :

— une courbe empirique ;

— une courbe interpolée ;

— une courbe extrapolée [23].

La construction *A as GN* est productive à tous les niveaux (DESAGULIER, 2015a, p. 23–24). Les courbes continuent de croître au-delà du corpus. Les occurrences exactes de *A as GN* comprennent des éléments de variation structurelle (des déterminants, des modifieurs du

23. Les modèles LNRE choisis sont précisés dans (DESAGULIER, 2015a, Tableau 9).

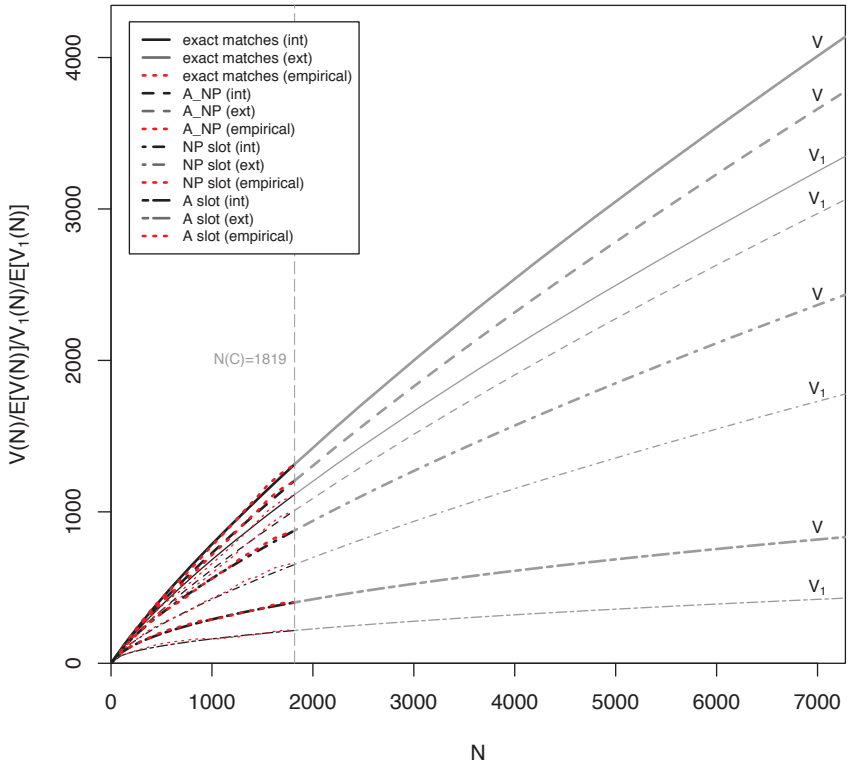


Figure 4.11: Courbes de croissance lexicale pour les occurrences exactes de *A as GN*, les paires A_GN, la place adjectivale et la place nominale avec des interpolations (int) et des extrapolations (ext) jusqu'à quatre fois la taille du corpus (DESAGULIER, 2015a, Figure 1)

GN, etc.) qui accentuent la pente de la courbe. Une fois ces élements retirés, les paires A-GN affichent un degré de productivité moindre. Cette productivité est néanmoins significative de par la richesse des combinaisons entre les adjectifs et les GN. Considérés isolément, les adjectifs et les GN sont moins productifs que leurs appariements. Cela est dû à leur récurrence.

4.7.3 L'apprentissage ciblé

Dans une optique multifactorielle, les méthodes prédictives les plus populaires sont les méthodes de régression (DESAGULIER, à paraître, Chapitre 11). Ces méthodes consistent à expliquer et prédire des variables quantitatives ou qualitatives par des variables quantitatives et/ou qualitatives. Comme les méthodes de clustering, les méthodes de régression sont conçues pour croiser plusieurs variables (et plusieurs types de variables) de manière à dégager des corrélations. En plus de dégager de ce croisement des corrélations significatives, les méthodes de régression vont au-delà puisqu'elles ont une valeur prédictive.

Une corrélation est une association significative entre (au moins) deux variables (DESAGULIER, à paraître, Section 8.11). Nous avons vu plus haut qu'il existait une corrélation entre le choix de l'intensifieur et la syntaxe de l'adverbe : l'emploi de *quite* est corrélé positivement avec le choix de la syntaxe pré-déterminantale tandis que l'emploi de *rather* est corrélé positivement avec le choix de la syntaxe pré-adjectivale. Aussi statistiquement significatives soient-elles pour le statisticien et aussi pertinentes soient-elles pour le linguiste,

ces corrélations n'ont aucune valeur causale : de la corrélation entre le choix de l'intensifieur et le choix de la syntaxe, on ne peut pas conclure que la réalisation d'une variable déclenche la réalisation de l'autre. En théorie, les méthodes prédictives permettent de dégager des relations causales.

La linguistique quantitative a accordé une large place à la régression logistique. Cette méthode fait ses preuves en sociolinguistique et en linguistique historique depuis plusieurs décennies sous le nom de « variable rule analysis » . Elle est à présent appliquée à tous les domaines de l'analyse linguistique (DESAGULIER, à paraître, Section 11.4), en particulier lorsqu'il s'agit d'expliquer et de prédire le phénomène d'alternance (BRESNAN et al., 2007 ; SPEELMAN, 2014).

Tout en étant convaincu par la justesse de la régression logistique, j'ai souhaité tester les prétentions de cette méthode en matière d'explication et de prédiction (CHAMBAZ et DESAGULIER, 2016). J'ai collaboré pour l'occasion avec le mathématicien Antoine Chambaz (Université Paris Ouest Nanterre, Modal'X). Nous avons choisi de travailler sur l'alternance dative, illustrée en (45) :

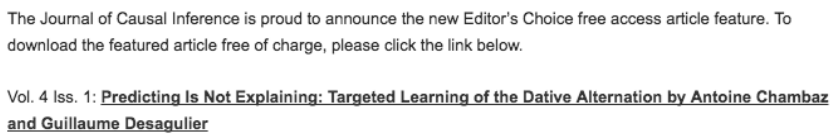
(45) a. Will gave his best student a manuscript. (double objet)
b. Will gave his manuscript to his best student (datif prépositionnel)

En (45a), le bénéficiaire est réalisé en tant que groupe nominal tandis qu'en (45b) il est réalisé en tant que groupe prépositionnel. Dans le premier cas, il s'agit d'un datif à double objet et dans le deuxième d'un datif prépositionnel. Le choix de l'alternance dative est motivé à la fois pour son intérêt linguistique (CHAMBAZ et DESAGULIER, 2016, Section 3.1) et pour le fait que c'est un phénomène sur lequel plusieurs méthodes prédictives ont été testées (CHAMBAZ et DESAGULIER, 2016, Section 3.2).

L'article a été publié dans une revue prestigieuse de mathématiques spécialisée dans l'analyse causale : *Journal of Causal Inference* (`http://www.degruyter.com/view/j/jci`) [24]. Le choix de cette revue a deux motivations. La première est la suivante : notre approche prend en partie le contre-pied de la régression logistique. Cette dernière étant très ancrée en linguistique quantitative (a fortiori en linguistique cognitive quantitative), la présentation de nos résultats a suscité au premier abord un sentiment d'incrédulité auprès de certains linguistes. Il nous a semblé important de valider notre approche causale auprès d'un public expert avant de la présenter aux linguistes [25]. La seconde motivation tient au fait que la revue invite les publications pluridisciplinaires. Antoine Chambaz et moi-même avons proposé une analyse linguistique croisée avec une réflexion mathématique sans avoir fait de compromis sur les exigences de nos deux disciplines respectives. Cet effort a été salué par l'équipe éditoriale de la revue, qui a choisi de mettre l'article en avant et de rendre son accès gratuit (Figure 4.12).

24. L'article en soi est le résultat tangible d'une collaboration entamée en 2012. Cette collaboration a donné lieu à plusieurs projets, dont un un projet PEPS ainsi qu'un séminaire commun entre le laboratoire de mathématiques Modal'X et mon laboratoire de linguistique.

25. La revue a été co-fondée notamment par Judea Pearl (`http://bayes.cs.ucla.edu/jp_home.html`), théoricien de l'informatique moderne, et Mark van der Laan (https ://www.stat.berkeley.edu/ laan/), expert mondial en analyse causale appliquée.


The Journal of Causal Inference is proud to announce the new Editor's Choice free access article feature. To download the featured article free of charge, please click the link below.

Vol. 4 Iss. 1: **Predicting Is Not Explaining: Targeted Learning of the Dative Alternation by Antoine Chambaz and Guillaume Desagulier**

Figure 4.12: Page d'accueil de *Journal of Causal Inference* en septembre 2016

L'article comprend une bonne part de raisonnements mathématiques qui ne sont malheureusement pas à la portée de la plupart des linguistes [26]. Je résume ci-dessous les points-clés de l'article.

L'analyse traditionnelle fait un usage explicatif de méthodes prédictives. Cependant, la prédiction et l'explication ne se recouvrent exactement ni en théorie ni en pratique. En effet, on peut très bien prédire un phénomène (par exemple la trajectoire d'une comète) sans pour autant connaître toutes les lois qui le conditionnent (les lois du cosmos). Les mathématiciens nous alertent sur le fait que les méthodes prédictives ne sont pas toujours adaptées à l'explication. Ceci ne veut pas dire que les linguistes ne doivent pas faire de prédiction mais que lorsque l'on cherche à expliquer un phénomène linguistique, la prédiction n'est qu'un moyen et non une fin en soi. Dans l'article, nous distinguons ce qui relève de la prédiction de l'alternance dative de ce qui relève de son explication (Chambaz et Desagulier, 2016, Section 4.2).

Prédire l'alterance suppose de construire un algorithme qui passe pour un locuteur natif de l'anglais formulant une construction impliquant l'une des deux formes du datif. À ce stade, l'algorithme n'a pas besoin de nous expliquer comment l'alternance fonctionne (Chambaz et Desagulier, 2016, Section 4.2.1). Expliquer l'alternance suppose de découvrir ce qui conditionne le choix d'un alternant plutôt que l'autre. Même si nous n'avons pas accès à la loi (au sens mathématique du terme) de l'alternance dative, nous pouvons l'estimer (Chambaz et Desagulier, 2016, Section 4.2.2).

Notre approche repose sur l'apprentissage ciblé [27]. Tout d'abord, nous opérationnalisons une série de questions concernant l'alternance dative dans un cadre causal. La question principale est la suivante : quel est l'effet du contexte sur la réalisation de la construction dative ? Nous y répondons en plusieurs étapes. Nous utilisons des algorithmes d'apprentissage polyvalents issus des dernières avancées en statistiques semi-paramétriques. Nous dérivons des estimations, des intervalles de confiance et des p-valeurs pour des paramètres bien définis. Ces paramètres sont conçus comme l'influence de chaque variable contextuelle sur la réalisation de l'alternance (double objet vs. datif prépositionnel).

Le Tableau 4.22 présente l'estimation numérique de l'effet des variables contextuelles sur la syntaxe de l'alternance dative dans le jeu de données de Bresnan et al. (2007). Il est différent des tableaux de sortie de la régression logistique (Chambaz et Desagulier, 2016, Section 6.1). Cette dernière propose un modèle prédictif ajusté et signale les variables les plus prédictives (Desagulier, à paraître, Section 11.4).

26. Nous rédigeons actuellement une version plus accessible aux linguistes. Toutefois, la publication de cet article ne sera pas effective au moment de la soutenance.
27. Cette méthode est employée à l'origine en biostatistiques.

Tableau 4.22: Estimation de l'effet des variables contextuelles catégorielles sur la syntaxe de l'alternance dative (CHAMBAZ et DESAGULIER, 2016, Tableau 1)

variable	*vs.*	estimate	CI	p-value
`Modality`	`written%spoken`	0.0277	[-0.0031,0.0585]	0.0776
`AnimacyOfRec`	`inanimate%animate`	0.0938	[0.0549,0.1327]	0.0000
`DefinOfRec`	`indefinite%definite`	0.0395	[0.0102,0.0688]	0.0083
`PronomOfRec`	`pronominal%nonpronominal`	-0.1398	[-0.2171,-0.0624]	0.0004
`AnimacyOfTheme`	`inanimate%animate`	0.0843	[0.0337,0.1348]	0.0011
`DefinOfTheme`	`indefinite%definite`	-0.0568	[-0.0865,-0.0272]	0.0002
`PronomOfTheme`	`pronominal%nonpronominal`	-0.1168	[-0.1377,-0.0959]	0.0000
`AccessOfRec`	`new%accessible`	-0.3824	[-0.5458,-0.2189]	0.0000
	`given%accessible`	0.0411	[-0.0149,0.0971]	0.1506
`AccessOfTheme`	`new%accessible`	-0.0782	[-0.1100,-0.0463]	0.0000
	`given%accessible`	-0.0415	[-0.0673,-0.0157]	0.0016
`SemanticClass`	`t%a`	0.1152	[0.0548,0.1755]	0.0002
	`p%a`	-0.0928	[-0.1532,-0.0324]	0.0026
	`f%a`	-0.1471	[-0.1946,-0.0997]	0.0000
	`c%a`	0.1657	[0.1238,0.2077]	0.0000

Ce tableau (dont je ne reproduis pas ici l'analyse détaillée) nous indique la probabilité d'obtenir un datif prépositionnel lorsque l'on change la modalité d'une variable. Par exemple, la deuxième ligne du tableau indique que la probabilité d'obtenir un datif prépositionnel décroît de 9,38% lorsque le bénéficiaire passe d'animé à inanimé [28]. L'effet des variables contextuelles numériques est résumé en Figure 4.13.

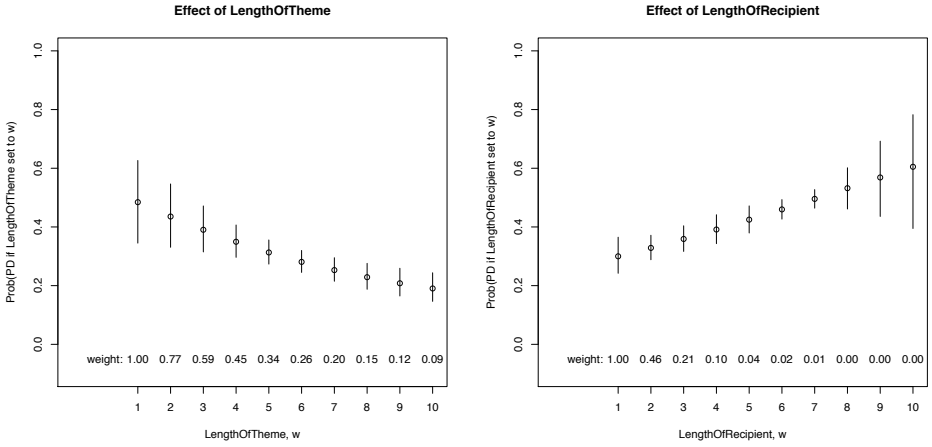
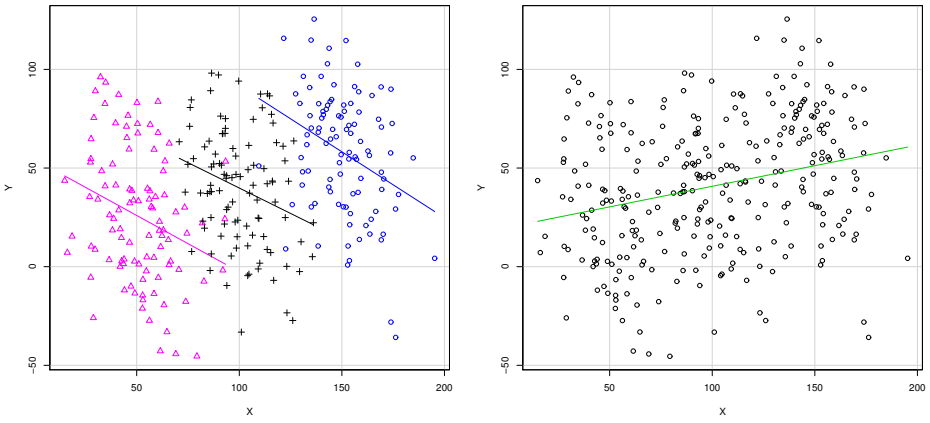


Figure 4.13: Estimation de l'effet des variables contextuelles numériques sur la syntaxe de l'alternance dative (CHAMBAZ et DESAGULIER, 2016, Figure 1)

L'apprentissage ciblé m'a fait prendre conscience des pièges qui jalonnent l'interprétation des données. L'un de ces pièges est le paradoxe de Simpson (CHAMBAZ et DESAGULIER, 2016, Section 6.2). Celui-ci est très simple à comprendre : une tendance observée séparément dans plusieurs groupes de données s'inverse lorsque les groupes sont confondus (Figure 4.14). En Figure 4.14a, une corrélation négative existe entre X et Y pour chacun des trois groupes de données. En Figure 4.14b, on voit que la corrélation est positive pour l'ensemble des données [29].

28. Nous ignorons dans l'article les lignes pour lesquelles la p-valeur est supérieure à 0,05.
29. Je propose une illustration numérique simple dans DESAGULIER (à paraître, Section 11.5.3).



(a) une corrélation négative dans chacun des trois groupes...

(b) ... s'inverse lorsque les trois groupes sont confondus

Figure 4.14: Illustration graphique du paradoxe de Simpson

Tableau 4.23: Tableau de contingence croisant la fréquence des datifs prépositionnels (PD) et des datifs à double objet (DO) en fonction de la définitude du thème (CHAMBAZ et DESAGULIER, 2016, Tableau 2)

	theme : definite	theme : indefinite
PD	63	378
DO	28	858

Bien connu des mathématiciens, ce paradoxe n'est, à ma connaissance, pas pris en compte dans les études linguistiques faisant intervenir la régression logistique. Il est pourtant susceptible de se réaliser dès lors qu'un biais de confusion est possible. Le biais de confusion intervient lorsque, dans l'étude des liens entre deux variables, une autre variable oubliée interfère [30] . Le tableau 4.23 croise la fréquence des datifs prépositionnels et des datifs à double objet en fonction de la définitude du thème.

Si l'on ignore les variables de confusion associées au contexte, la probabilité d'obtenir un datif prépositionnel dont le thème est défini ou indéfini est de 38,65% avec un intervalle de confiance de [28,82%, 48,48%]. Si l'on prend le contexte en compte, comme nous l'avons fait dans l'étude, la probabilité d'obtenir un datif prépositionnel décroît de 5,68% lorsque le thème passe de défini à indéfini. Ce résultat suggère qu'il y a de la confusion et qu'il faut la prendre en compte.

4.8 Quels outils pour les statistiques ?

Les statistiques constituent la seconde partie de mon livre. Je ne les ai pas abordées toutes ici, certains chapitres étant des introductions au raisonnement statistique (en particulier le

30. En dehors de la linguistique, l'exemple classique est une étude affirmant que boire du café augmente les risques de cancer du poumon. La corrélation est avérée, mais il semble bien qu'une variable de confusion intervienne : bon nombre d'amateurs de café sont aussi des fumeurs.

chapitre 8 « Notions of statistical testing »). Tout comme la première partie, la seconde invite le lecteur à prendre en compte la spécificité des statistiques appliquées à la linguistique.

J'ai fait le choix de rester dans l'environnement de R . Il y a trois avantages immédiats. Premièrement, R a été développé par et pour des statisticiens. Des scripts permettant de réaliser les tests statistiques les plus courants sont inclus par défaut. Deuxièmement, le langage de programmation appris pour les techniques de corpus est le même que pour les statistiques. Enfin, R est devenu la plateforme de référence pour une communauté croissante de linguistes. En maîtrisant R , le linguiste peut donc constituer ses corpus, extraire des observations, les tabuler, les quantifier, réaliser des tests de significativité et des graphiques dans le même environnement. Enfin, le développement de R est collaboratif et invite le chercheur à faire partie d'une communauté d'utilisateurs.

5 Premier prolongement : modéliser les réseaux de constructions

> *As a first approximation, then, we can describe a knowledge system as a network where nodes correspond to conceived entities, and arcs to the conceived relationships in which they participate. However, cognitive and linguistic considerations suggest the need for a model that is more elaborate in certain respects than a simple, well-behaved network.*
>
> **— Ronald Langacker**
> (1987, p. 162)

5.1 Introduction

Ce chapitre et le suivant sont une projection des recherches que je souhaite mener dans un futur proche. Elles n'ont pas encore fait l'objet de publications sous forme d'articles [1]. Il s'agit pour moi d'élaborer des méthodes permettant de modéliser des schémas constructionnels sous forme de réseaux dynamiques à partir de grands corpus. Jusqu'ici, je me suis attaché à le faire par l'entremise de méthodes de classification. Ces méthodes sont toutefois limitées parce qu'elles privilégient la visualisation de clusters au détriment de ce qui les relie (Desagulier, 2014).

Au Chapitre 2, j'ai eu l'occasion d'insister sur le fait qu'en linguistique cognitive, la compétence linguistique est conçue comme un réseau complexe et structuré d'unités symboliques. Sur le plan conceptuel, cette idée a été développée à plusieurs reprises (Lakoff, 1987 ; Langacker, 1987 ; Fillmore, Kay et O'Connor, 1988 ; Goldberg, 1995 ; Croft, 2001). La grammaire mentale est un inventaire structuré d'appariements forme/sens interconnectés, non une liste statique de principes modulaires.

Un graphe se compose de nœuds (ou sommets) et d'arêtes (ou arcs) dont les attributs peuvent être naturellement associés à des traits linguistiques. Chaque appariement forme/sens peut être représenté par un nœud et chaque lien d'hérédité par une arête (Figure 5.1).

Dans les faits, les réseaux de constructions sont plus complexes que ne le suggère la Figure 5.1 dans la mesure où interviennent des composantes formelles, sémantiques et pragmatiques.

1. J'ai publié un réseau cumulatif des interactions entre les personnages de *Macbeth* à destination d'étudiants dans le cadre d'un séminaire de Master sur les humanités numériques : `https://gdlinguistics.shinyapps.io/macbeth/`.

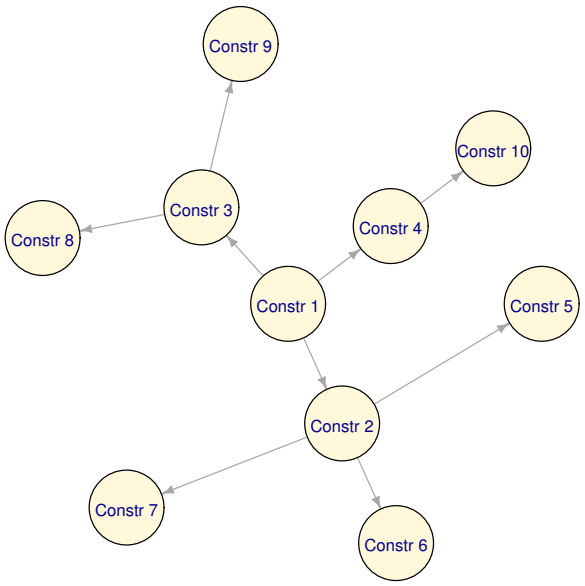


Figure 5.1: Un exemple fictif de réseau de constructions (les cercles sont des nœuds et les flèches des arêtes ; la présence de flèches indique que le graphe est directionnel et caractérise le schéma d'hérédité)

5.2 Les réseaux comme heuristique

En linguistique, le concept de réseau a longtemps été confiné au statut de métaphore dans une optique heuristique, sous l'influence probable des modèles interactionnistes, connexionnistes, et émergentistes de la cognition, qui décrivent le cerveau comme un ensemble de réseaux neuronaux denses. La densité de ces réseaux est liée au grand nombre de connexions entre un très grand nombre de nœuds (ELMAN et MCCLELLAND, 1984 ; ELMAN, BATES et al., 1996).

En Grammaire Cognitive, le réseau en tant qu'outil heuristique intervient à plusieurs niveaux dans l'analyse linguistique. Par exemple, LANGACKER (1987) s'appuie sur la modélisation en réseau pour rendre compte du phénomène d'allophonie, de l'allomorphie du pluriel en anglais (Figure 5.2) et de la catégorisation et des relations sémantiques telles que l'hyponymie, l'hypéronymie ou la métonymie, etc. (LANGACKER, 1987, chapitre 10).

L'idée que les constructions sont ancrées sous forme de réseaux est aussi une idée force des grammaires de constructions.

> It is argued that constructions form a network and are linked by inheritance relations which motivate many of the properties of particular constructions. The inheritance network lets us capture generalizations across constructions while at

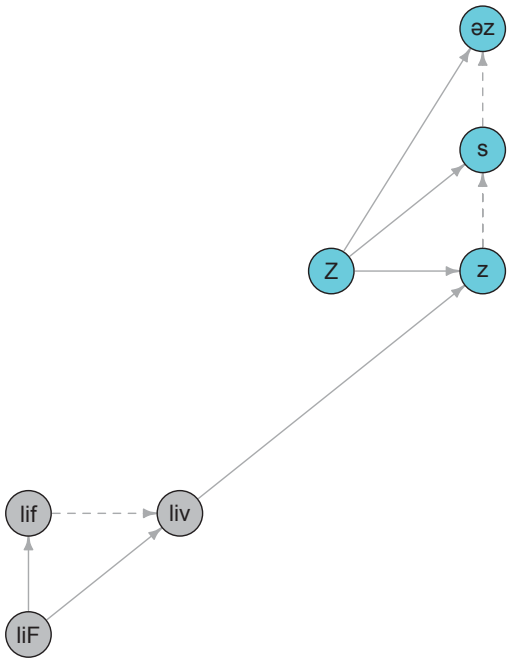


Figure 5.2: Le réseau allomorphique du pluriel en anglais (LANGACKER, 1987, p. 395) : le cas de *leaves* (adapté à l'aide d'`igraph` pour `R` ; en gris les réalisations phonologiques de *leaf* ; en cyan les allomorphes du pluriel {Z})

> the same time allowing for subregularities and exceptions. (GOLDBERG, 1995, p. 67)

À titre d'illustration, la Figure 5.3 présente le réseau d'héritage de la construction *Is to* en anglais, illustrée en (46), selon GOLDBERG et VAN DER AUWERA (2012, p. 121). Le réseau décrit ici un schéma d'héritage par défaut entre des nœuds « mères » et des nœuds « filles ». Les nœuds « filles » héritent des propriétés non-conflictuelles des nœuds « mères ».

(46) a. We are to meet at the station at 10am.

b. You are to brush your teeth before going to bed.

L'idée selon laquelle la grammaire est un réseau complexe et dynamique d'unités symboliques est largement consignée à l'origine au statut d'abstraction théorique en linguistique cognitive de première génération. Elle n'a reçu de fondations empiriques que récemment. Les linguistes prennent peu à peu conscience de l'intérêt de visualiser les unités linguistiques en réseaux. Par exemple, ELLIS, O'DONNELL et al. (2014) et GRIES et ELLIS (2015) parviennent à quantifier les traits sémantiques distinctifs de constructions Verbe-Argument telles que les causatives en *into* de manière à projeter les verbes qui interviennent dans ces constructions sous la forme d'un réseau. En somme, grâce à la théorie des graphes, le réseau de constructions a l'occasion de dépasser le simple statut de métaphore.

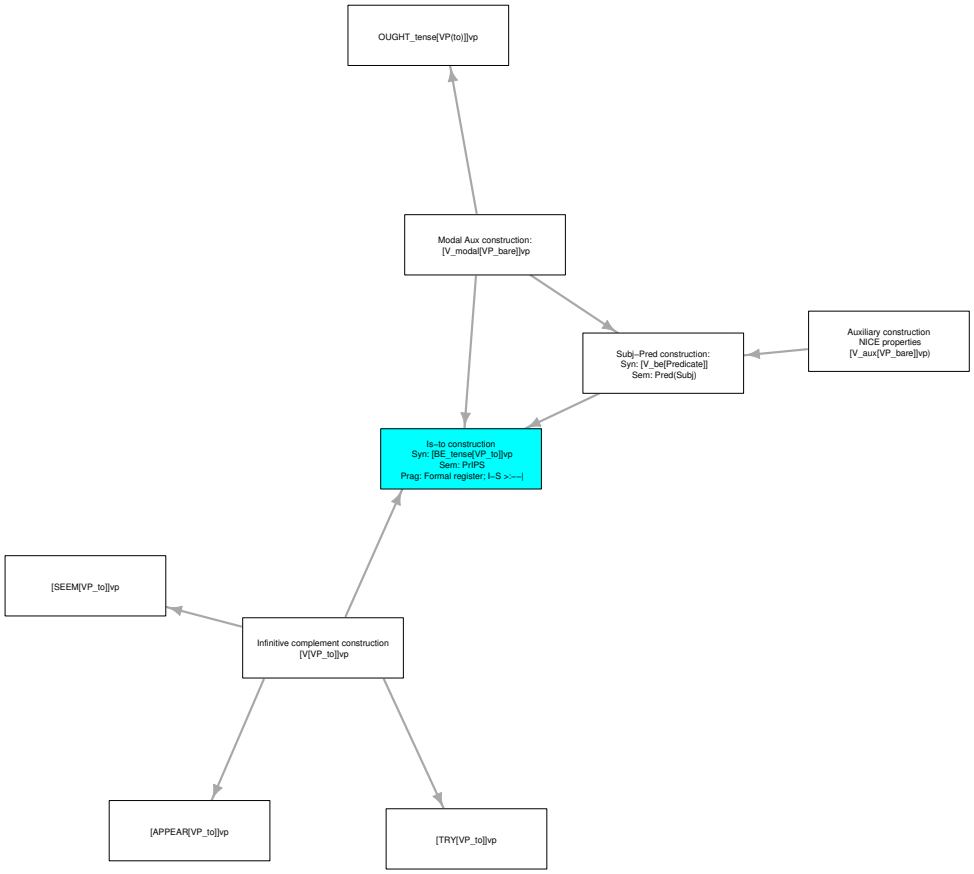
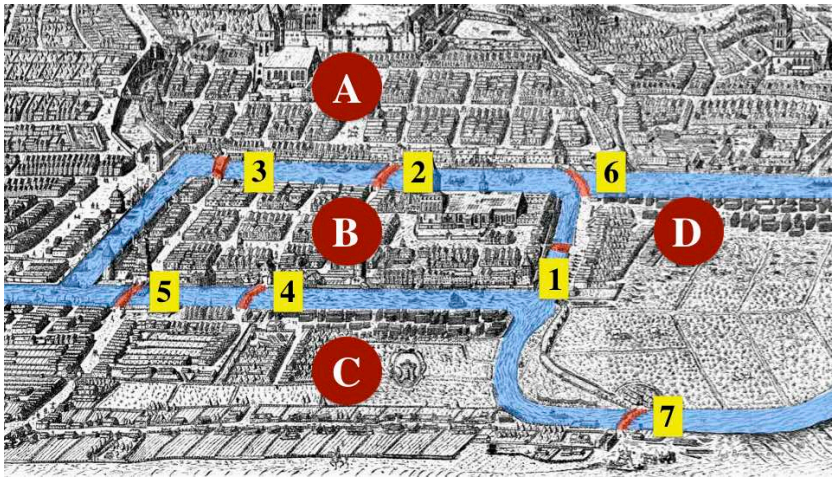


Figure 5.3: La construction *Is to* selon GOLDBERG et VAN DER AUWERA (2012) adaptée à l'aide du package `igraph` pour `R`

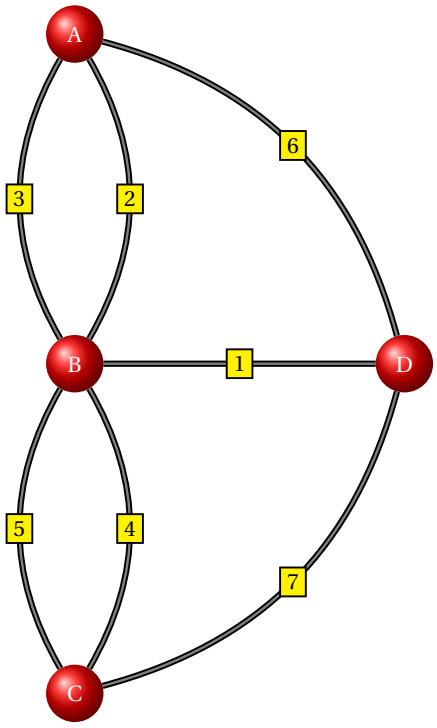
5.3 La théorie des graphes

La science des réseaux trouve son origine dans la théorie des graphes, amorcée par Euler lors de sa résolution du problème des ponts de Königsberg au XVIII^{e} siècle (Figure 5.4). La Figure 5.4a est une carte de la ville de Königsberg en 1736. Elle est divisée en plusieurs parties : l'île Kneiphof (B) et trois quartiers délimités par la rivière Pregel : A, C et D. Les quartiers sont reliés par sept ponts (numérotés de 1 à 7). Le problème est le suivant : il faut trouver un itinéraire tel qu'un piéton n'aurait à emprunter chaque pont qu'une seule fois pour passer par tous les quartiers de la ville. Leonhard Euler remplaça chacun des quatre quartiers par un nœud et chaque pont par une arête, de manière à obtenir un graphe (Figure 5.4b).

À l'aide de ce graphe, il prouva que cet itinéraire est impossible. Les nœuds comportant un nombre impair d'arêtes ne peuvent être que des points de départ ou des points d'arrivée. Un chemin continu traversant tous les ponts ne peut avoir qu'un seul point de départ et un seul point d'arrivée. Par conséquent, ce chemin ne peut pas être tracé sur un graphe



(a) carte de Königsberg en 1736



(b) transposition de la carte sous forme de graphe

Figure 5.4: Le problème des sept ponts de Königsberg

comportant plus de deux nœuds reliés à un nombre impair d'arêtes. Parce que graphe de la Figure 5.4b comporte quatre nœuds reliés à un nombre impair d'arêtes, il s'avère que l'itinéraire recherché est impossible [2].

À l'époque, c'est la preuve mathématique qui retint l'attention et non le moyen de l'obtenir par le biais d'un graphe. Depuis, toutefois, le graphe d'Euler a fait l'objet de plusieurs formalisations (voir par exemple KÖNIG, 1950) et l'on attache désormais plus d'importance à la formalisation du problème qu'à sa solution. C'est ce qui a donné naissance à la théorie des graphes.

5.4 Le « petit monde » des phénomènes langagiers

La théorie des graphes aléatoires (ERDŐS et RÉNYI, 1959) est probablement la déclinaison la plus influente de la théorie des graphes. Selon la théorie des graphes aléatoires, tous les nœuds ont la même probabilité d'être reliés par des arêtes et sont positionnés de manière aléatoire (Figure 5.6a). Lorsque l'on étudie des phénomènes langagiers, cette propriété a de fortes chances de ne pas se réaliser.

En effet, les mots n'interagissent pas de manière aléatoire dans une phrase ou un texte, et encore moins dans une construction (KILGARRIFF, 2005). Comme nous l'avons vu, la distribution des mots dans les langues naturelles est caractérisée par la loi de Zipf (ZIPF, 1949). Ainsi, la fréquence des mots décroît inversement à leurs rangs dans un tableau de fréquence en vertu d'une loi de puissance. Cette distribution se caractérise par un grand nombre d'événements rares coexistant avec un petit nombre d'événements très fréquents. À titre d'exemple, la Figure 5.5 illustre la distribution des mots pour chacun des huit types de textes du BNC. Chacune des courbes est typiquement zipfienne.

CANCHO et SOLÉ (2001) observent que les cooccurrences de mots suivent la structure en réseau du lexique. Cette structure est telle qu'elle peut être décrite en référence à des interactions de type « petit-monde ». Comme l'ont observé les spécialistes de biologie, de physique ou des réseaux sociaux, certains graphes ont les propriétés d'un « petit monde » (D. J. WATTS et STROGATZ, 1998). Les graphes « petit-monde » ont deux traits distinctifs :

— ils sont très denses,

— la distance moyenne entre deux nœuds est courte (Figure 5.6b).

Ce qui précède a été vérifié lorsque du texte linéaire est transformé en un réseau de mots interconnectés de manière à visualiser leurs relations (DRIEGER, 2013). En d'autres termes, les mots interagissent de manière dense avec leur voisinage propre.

Des recherches menées dans le cadre des grammaires de constructions (ELLIS et FERREIRA-JUNIOR, 2009 ; ELLIS, O'DONNELL et al., 2014 ; GRIES et ELLIS, 2015) ont confirmé cette tendance. Elles ont révélé les propriétés suivantes :

2. En 1875, un nouveau pont fut construit entre A et C. Cela eut pour conséquence de réduire le nombre de nœuds reliés à un nombre impair d'arêtes à deux et l'itinéraire devint possible. L'histoire ne dit pas si le pont fut construit pour résoudre le problème des ponts de Königsberg.

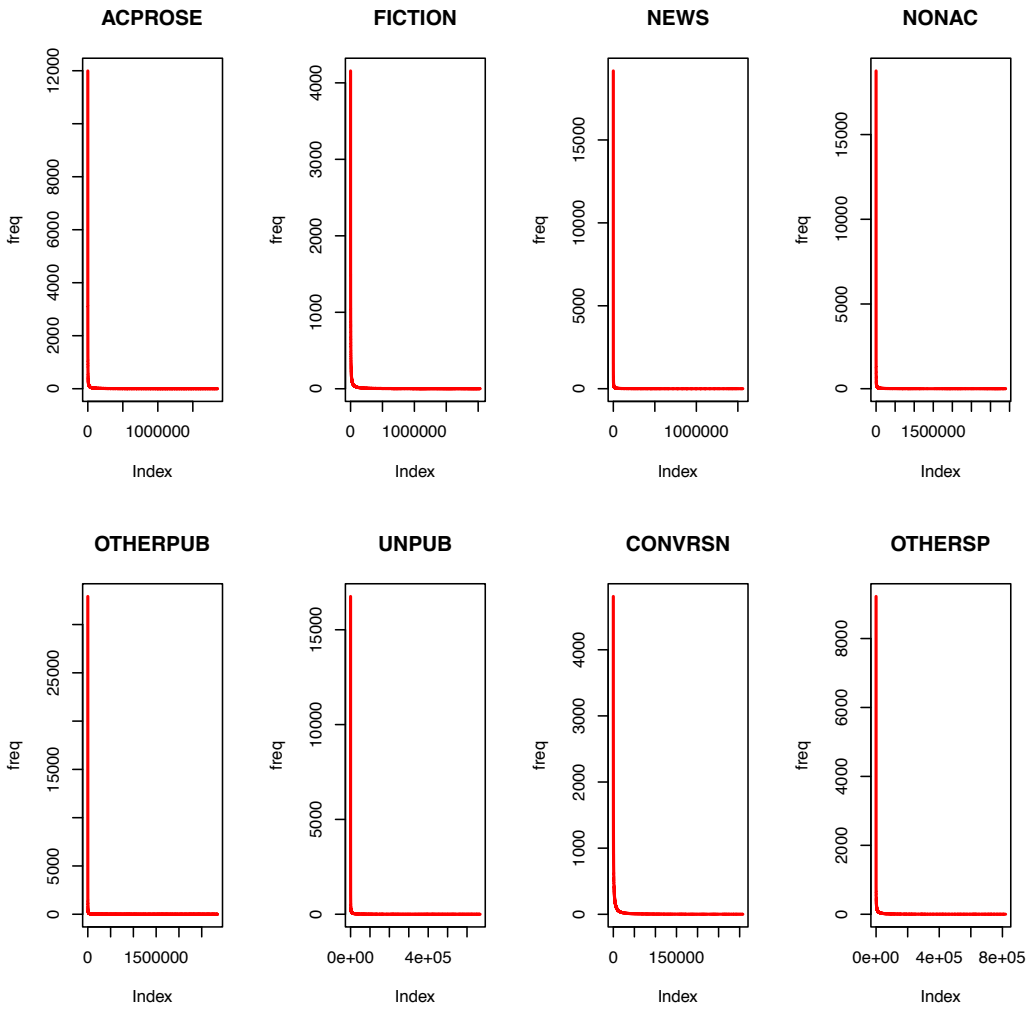
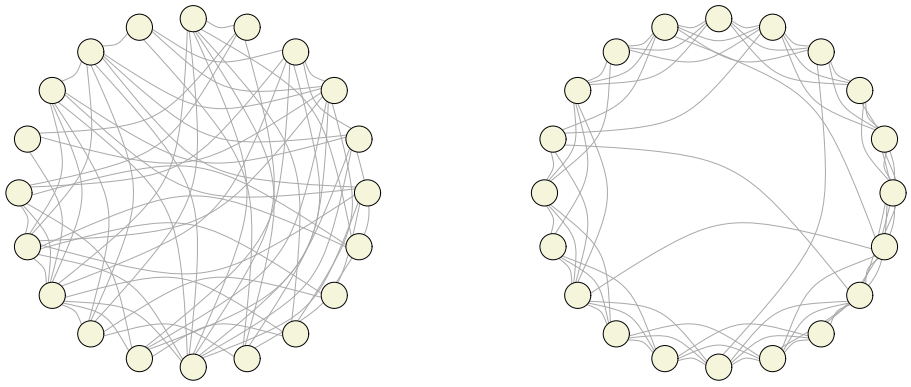


Figure 5.5: Distribution lexicale pour chacun des huit types de textes dans le BNC



(a) graphe aléatoire

(b) graphe « petit monde »

Figure 5.6: Comparaison d'un graphe aléatoire et d'un graphe « petit monde »
(les deux graphes ont le même nombre de nœuds et d'arêtes)

— la réalisation des places instanciables des constructions suit une distribution zipfienne ;

— les constructions sont sélectives quant à la réalisation de leurs places instanciables ;

— les constructions sont homogènes quant aux réseaux qu'elles génèrent.

Jusqu'ici, la théorie des graphes a été appliquée à la sémantique des constructions. J'ai le projet de modéliser les réseaux de constructions en rétablissant l'équilibre entre sens et forme, la deuxième ayant été quelque peu laissée de côté au profit de la première.

5.5 Visualiser les réseaux de constructions

La théorie des graphes est adaptée à l'analyse en réseaux des constructions. Nous avons vu à plusieurs reprises que, dans le cadre de la linguistique de l'usage, la fréquence était un facteur central dans l'ancrage de représentations linguistiques [3]. Plus un locuteur est exposé à un événement linguistique, plus cet événement aura de chances d'être ancré dans la mémoire du locuteur, et plus ce dernier y aura accès rapidement. La fréquence d'une unité linguistique a un corrélat dans la théorie des graphes. Les nœuds les plus fréquents sont plus importants que les nœuds les moins fréquents.

L'exemple qui suit est fondé sur le jeu de données utilisé par BRESNAN et al. (2007) dans leur étude sur l'alternance dative, illustrée en (45), p. 90 [4].

Le graphe présenté en Figure 5.7 synthétise les collocations entre les deux schémas syntaxiques de l'alternance dative (NP : datif à double objet ; PP : datif prépositionnel) et les verbes recrutés par la construction. La sémiotique du graphe est la suivante :

— chaque constituant constructionnel est représenté par un nœud ;

— chaque nœud est représenté par un disque ;

— les nœuds correspondant aux deux schémas syntaxiques sont en bleu ;

— les nœuds correspondant aux verbes sont colorés suivant un dégradé de couleur correspondant à un spectre de chaleur [5] ;

— le diamètre des disques est fonction de la fréquence d'occurrence des constituants dans la construction (plus le diamètre est grand, plus le constituant est fréquent).

Le graphe met en avant plusieurs aspects saillants de l'alternance dative, à savoir :

— le datif à double objet est plus fréquent que le datif prépositionnel dans le corpus ;

— trois groupes de verbes se dégagent : les verbes propres au datif à double objet, les verbes propres au datif prépositionnel et les verbes communs aux deux schémas ;

— les verbes les plus fréquents sont ceux qui apparaissent dans les deux schémas ;

— le verbe *give*, considéré dans la littérature comme étant le plus représentatif du sens de la construction dative, est le plus fréquent et apparaît dans les deux schémas (les verbes les plus fréquents après *give* sont proches du sens prototypique : *send*, *sell*, *pay* et *tell*).

3. Même si nous avons vu également qu'il faut éviter d'établir une équation trop rapide entre fréquence élevée et ancrage cognitif.

4. Ce jeu de données est disponible dans le package `languageR` (BAAYEN, 2013).

5. Le spectre de chaleur est fonction d'une mesure propre à la théorie des graphes, à savoir la mesure de centralité (voir plus bas). La mesure de centralité mesure l'importance relative des nœuds dans le graphe. La mesure de centralité retenue ici est la centralité dite « de vecteur propre ».

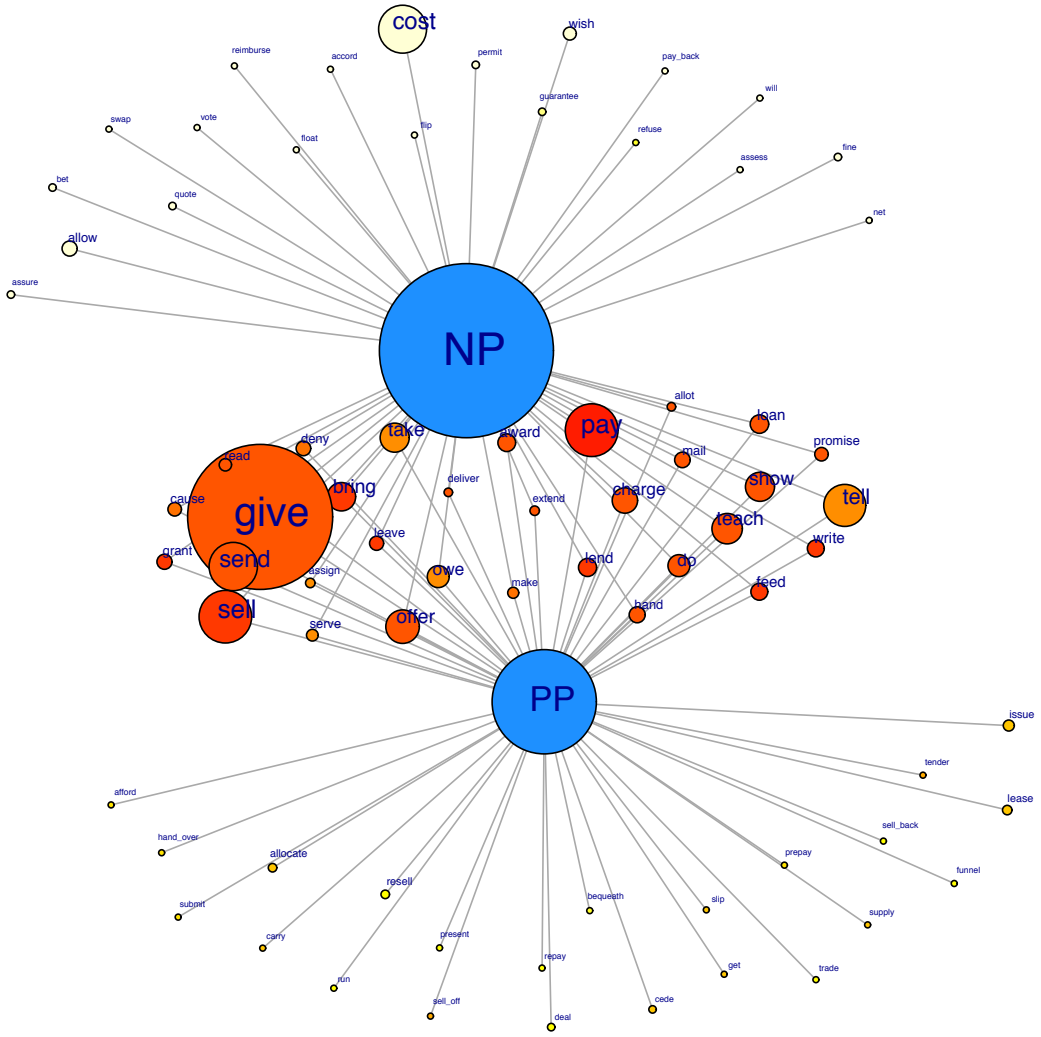


Figure 5.7: Collocations entre les verbes intervenant dans la construction dative en anglais et deux schémas syntaxiques : le schéma à double objet (NP) et le schéma prépositionnel (PP) (algorithme de Fruchterman Reingold)

Par rapport à la Figure 5.7, la Figure 5.8 ajoute un niveau de description supplémentaire, à savoir la catégorie sémantique des verbes (en vert). BRESNAN et al. (2007) répartissent les verbes en cinq catégories :

— *t* : transfert de possession (par ex. *send*, *I sent her flowers*) ;

— *a* : transfert abstrait (par ex. *give*, *give it some thought*) ;

— *f* : futur transfert de possession (par ex. *owe*, *I owe you $25*) ;

— *p* : transfert de possession bloqué (par ex. *deny*, *He was denied a raise*) ;

— *c* : communication (e.g. *tell*, par ex. *He told me the truth*).

Il apparaît que les trois classes sémantiques les plus fréquentes sont le transfert de possession, le transfert abstrait et la communication. Les verbes les plus fréquents interviennent dans plusieurs classes : a, t, c et f. Ces verbes, principalement en rouge, sont regroupés le long d'un arc autour du centre du réseau. Les verbes prototypiquement datifs (*give*, *send*, *pay*, *sell*) sont au centre du réseau. Ils sont compatibles avec les quatre classes sémantiques que nous venons de mentionner. Seule la classe p regroupe des verbes que l'on ne retrouve dans aucune autre classe. Cela est lié au sémantisme particulièrement contraint de la classe.

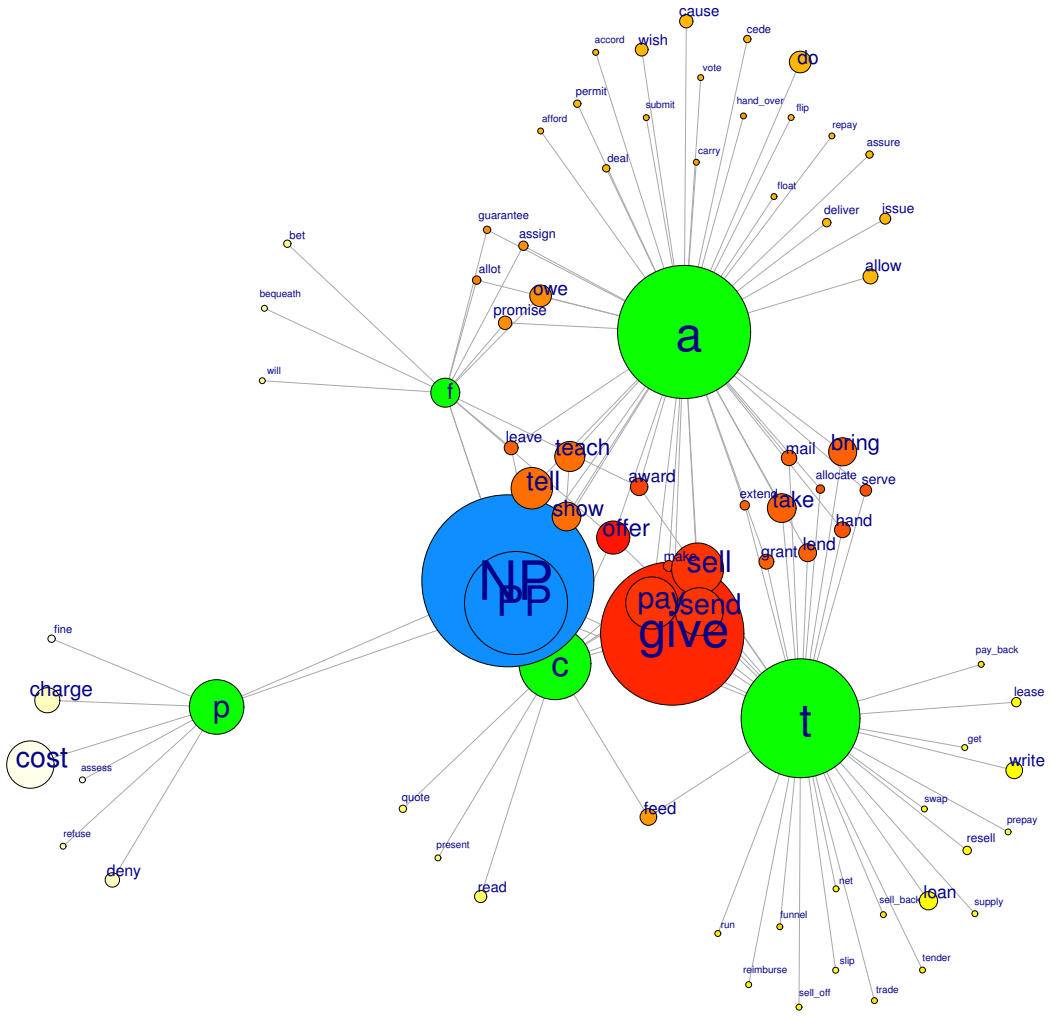


Figure 5.8: Collocations entre les verbes intervenant dans la construction dative en anglais, les classes de verbes et deux schémas syntaxiques (algorithme de Fruchterman Reingold)

La modélisation des collocations purement lexicales sous forme de réseaux n'est pas nouvelle (WILLIAMS, 1998 ; WILLIAMS, 2001 ; WILLIAMS, 2002). WILLIAMS (1998, p. 156) définit un réseau collocationnel de la manière suivante : « (. . .) a web of interlocking conceptual clusters realised in the form of words linked through the process of collocation ». La modélisation en réseau se prête également bien à l'étude des collocations directionnelles, telles que je les ai présentées dans l'étude de la construction *A as NP* présentée au Chapitre 4 (DESAGULIER, 2015a). La Figure 5.9 présente un graphe directionnel fondé sur le jeu de données de l'étude en question. Le but principal de cette visualisation est de mieux cerner le recrutement de constituants par la construction. La sémiotique du graphe reprend plusieurs mesures :

— la couleur et la forme du nœud dénotent la catégorie grammaticale du composant constructionnel : disque rouge pour les adjectifs et carré bleu pour les GN ;

— la taille du nœud est fonction de la fréquence d'occurrence de l'adjectif ou du GN dans la construction *A as NP* ;

— l'épaisseur de l'arête dénote le degré d'attraction symétrique entre l'adjectif et le GN mesurée à l'aide du rapport de log-vraisemblance ;

— la direction de l'arête dénote l'attraction asymétrique entre l'adjectif et le GN mesurée à l'aide de ΔP.

Il apparaît que certains types fonctionnent de manière isolée (*large as life*, *honest as the day is long*, *gauche as a schoolgirl*, *thin as a rake*, *taut as a bowstring*, etc.). D'autres s'insèrent au contraire dans une constellation de réseaux plus complexes (au centre du graphe). Ces réseaux complexes s'articulent sur des pôles attracteurs, dont le repérage est facilité par la visualisation. C'est le cas par exemple des adjectifs *white*, *cold*, *clear* ou *smooth* et du GN *hell*. Certains de ces pôles attracteurs sont eux-mêmes attirés par d'autres pôles. Par exemple, l'adjectif *sure* attire les GN *death* et *night follows day* tout en étant attiré par le GN *hell*.

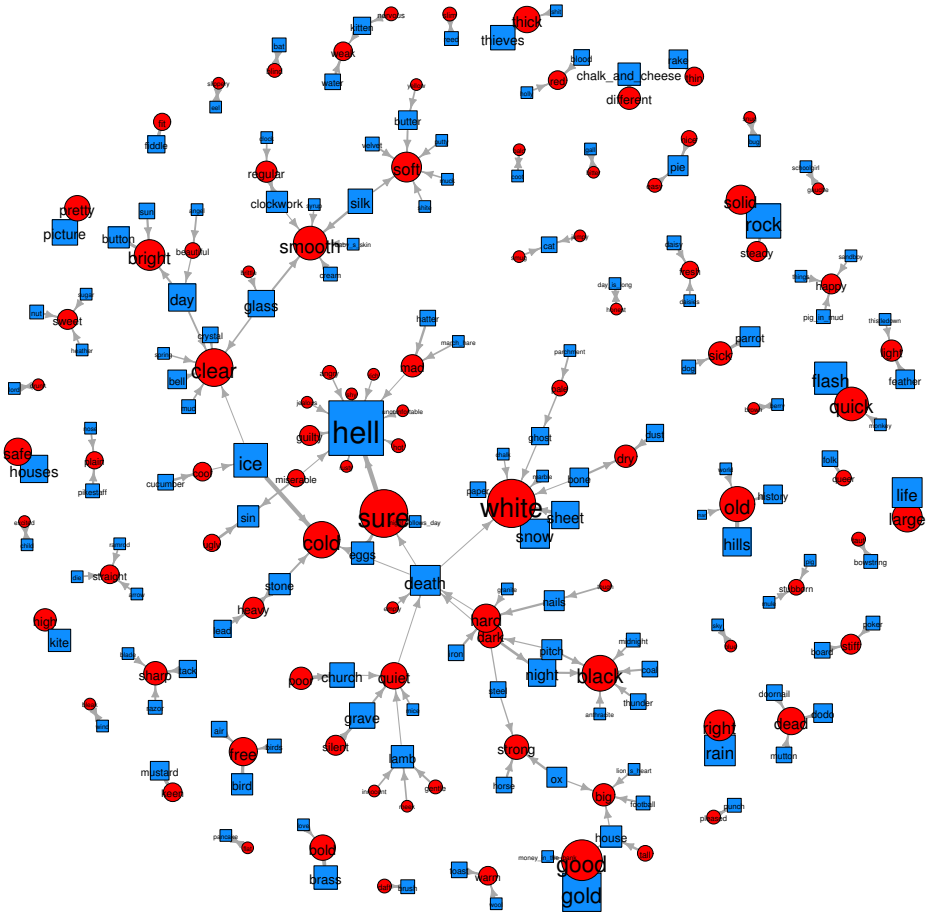


Figure 5.9: Graphe des collocations symétriques et asymétriques de *A as NP* (algorithme de Fruchterman Reingold)

On est assez proche ici de l'idée de « chaîne collocationnelle » développée par WILLIAMS (1998, p. 156–157) :

> (…) networks are formed by following through the 'collocational chains' which consists of initial look-up forms, referred to here as nodes in that they act as points of intersection in the graphical representation, and their collocates.

En Figure 5.9, toutefois, il s'agirait plutôt d'une « chaîne collostructionnelle » dont la nouveauté réside à la fois dans la directionnalité des attractions et dans le fait que nous ne partons pas d'entrées pré-définies (ce que Williams appelle « initial look-up forms »). Chaque constituant est à la fois un point d'entrée et un point d'arrivée. Ces chaînes collostruction-

nelles constituent un réseau dense dont les nœuds sont reliés par des arêtes courtes. Une fois combinés, ces nœuds forment un graphe « petit monde ».

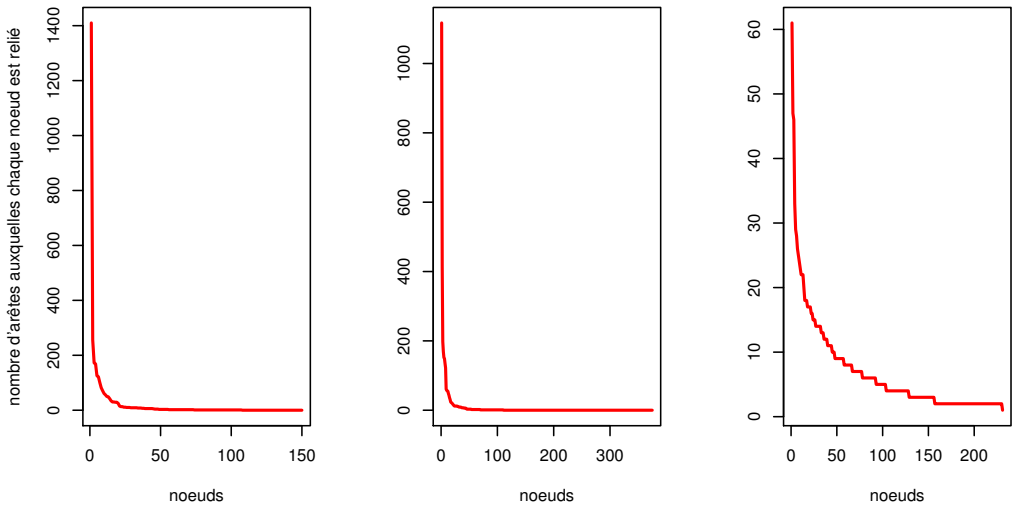


Figure 5.10: Distributions zifpiennes des nœuds en fonction des arêtes auxquelles ils sont reliés (à gauche : graphe de la Figure 5.7 ; au centre : graphe de la Figure 5.8 ; à droite : graphe de la Figure 5.9)

Notons que la courbe de Zipf n'est pas aussi marquée pour le graphe le plus à droite. Cela est lié au fait que ce graphe est n'est pas autant dominé par un petit nombre de « hubs » que les autres graphes [6]. Cela est lié au choix des variables qui interviennent dans la constitution du graphe. Les constituants constructionnels de *A as NP* sont globalement plus fréquents (car plus récurrents) que les variables de l'alternance dative (graphes de gauche et du centre).

5.6 Détecter les réseaux de constructions

Visualiser les réseaux de constructions sur la base de tableaux de données à l'aide de la théorie des graphes est une procédure relativement aisée. Les constructions sont pré-identifiées et les données pré-classées. Le graphe synthétise les données. Cette méthodologie suit, peu ou prou, une démarche exploratoire, à la manière de ce que proposent les méthodes de clustering. Plus ambitieuse et autrement plus difficile est la détection de réseaux de constructions à partir de grands corpus. J'expose ici les objectifs, les méthodes et les défis posés par cet axe de recherche, sur lequel mes recherches les plus récentes s'articulent.

5.6.1 Objectifs

Il est entendu en linguistique cognitive que les réseaux sont par essence dynamique :

> Our characterization of schematic networks has emphasized their "static" properties, but it is important to regard them as dynamic, continually evolving

6. Les hubs sont des nœuds au croisement d'un grand nombre d'arêtes, à l'image des échangeurs autoroutiers.

> structures. A schematic network is shaped, maintained, and modified by the pressures of language use. (LANGACKER, 1987, p. 381–382)

Générer un graphe dynamique permet de simuler la genèse des constructions. L'hypothèse sous-jacente est que des constituants formels s'agrègent en vertu d'une logique sémantico-pragmatique de manière à former des assemblages conventionnels.

5.6.2 Méthodes

La typologie des constructions est surtout une affaire de classification manuelle ou semi-automatique sur la base de corpus permettant ce type d'analyse. Mon projet consiste en la projection sous forme de réseaux sur la base de très grands corpus, tels que ceux compilés à partir de la Toile par BARONI et al. (2009). L'hypothèse de travail est que des propriétés macroscopiques du langage émergeront si le corpus est suffisamment grand. Dans la mesure où je souhaite répliquer les méthodes sur d'autres grands corpus, il est nécessaire d'élaborer une méthode permettant un passage à l'échelle.

La méthode repose sur un apprentissage semi-supervisé. L'apprentissage est effectué sur la base d'un répertoire de schémas pré-identifiés tels que ceux proposés par la Pattern Grammar (G. FRANCIS et al., 1996 ; HUNSTON et G. FRANCIS, 2000). Ces schémas sont vectorisés à l'aide d'un algorithme de vecteurs lexicaux (`GloVe`). Un premier réseau est construit sur la base de proximités vectorielles. Une fois l'apprentissage effectué, le modèle est appliqué à de nouveaux corpus. Chaque mot est vectorisé en fonction de son contexte d'occurrence. Les combinaisons de vecteurs les plus proches des combinaisons identifiées dans l'apprentissage initial sont alors intégrées au réseau originel.

Extraction des données

Pour extraire des données, deux méthodes seront comparées. La première méthode consiste à détecter la présence de schémas grammaticaux en corpus à partir de répertoires tel que ceux proposés par la Pattern Grammar (G. FRANCIS et al., 1996 ; HUNSTON et G. FRANCIS, 2000). Chaque constituant du schéma grammatical est alors traité comme un nœud du réseau. Chaque nœud est relié ou non aux autres nœuds du réseau. La seconde méthode consiste à utiliser le répertoire de constructions pré-identifiées pour entraîner un algorithme à détecter les schémas absents du répertoire via l'identification d'associations signifiantes.

La principale difficulté dans l'identification d'une construction est sa délimitation formelle. La méthode dite des n-grammes, qui consiste à calculer l'association entre mots voisins (qu'il y ait un lien sémantique entre eux ou non) a montré ses limites. On sait par exemple qu'il y a une association significative entre eux même lorsqu'ils n'appartiennent pas au même syntagme (GRIES, 2013a). Par ailleurs, certaines constructions illustrent le phénomène de dépendance à longue distance. Détecter automatiquement ce type de distance est une procédure ambitieuse. Heureusement, elle est assez bien traitée par certains outils de traitement automatique du langage tels que `CoreNLP` (MANNING et al., 2014). La Figure 5.11 est un exemple de dépendance à longue distance traitée par cet algorithme.

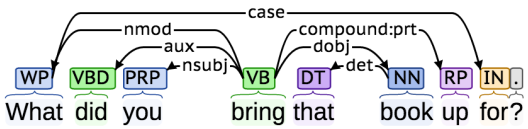


Figure 5.11: Deux dépendances à longue distance prises en charge par l'algorithme `CoreNLP` (Université de Stanford)

Prototypicalité, cohésion, communauté

Le formalisme des réseaux dans la théorie des graphes est à même de modéliser les aspects fondamentaux des constructions complexes. Ces aspects sont : la prototypicalité, la cohésion et les communautés. J'explique chacun de ces concepts ci-dessous et l'associe à une mesure propre à la théorie des graphes. Mon hypothèse est que ces trois paramètres pris en compte simultanément permettent d'identifier des familles de constructions.

Un prototype constructionnel est une caractérisation idéalisée du meilleur exemple d'une catégorie de constructions. La prototypicalité peut être capturée par le concept de centralité. La centralité mesure de l'importance d'un nœud dans le contexte d'un graphe. On peut envisager de l'appliquer pour détecter les constructions les plus prototypiques dans un réseau de constructions (ou les constituants constructionnels les plus prototypiques dans un sous-schéma). J'envisage de comparer trois mesures de centralité :

— la centralité de degré,

— la centralité de vecteur propre,

— la centralité d'intermédiarité.

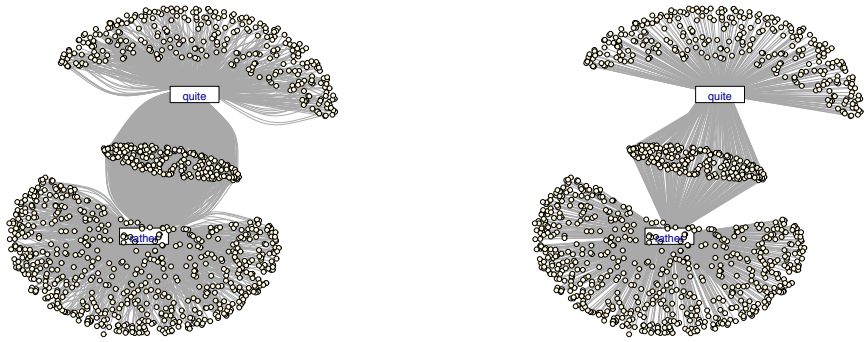
La centralité de degré est la plus simple des trois. Elle classe les nœuds en fonction du nombre d'arêtes auxquelles ils sont connectés. La centralité de vecteur propre (ou centralité spectrale) attribue à chaque nœud un score relatif. Plus ce nœud est relié à des nœuds dont le score est également élevé, plus le score du nœud en question est élevé. Enfin, la centralité d'intermédiarité classe les nœuds en fonction du nombre de fois où un nœud agit comme un point de passage le long du plus court chemin entre deux autres nœuds. L'un des problèmes liés à la visualisation des constructions en réseaux est que le nombre d'unités à représenter sur le graphe est élevé. Si la famille de constructions étudiée compte un grand nombre de types, le graphe peut vite devenir illisible. L'intérêt d'indexer la prototypicalité constructionnelle sur une mesure de centralité est de pouvoir bénéficier d'un seuil en deça duquel une construction considérée comme pas ou peu prototypique n'est pas projetée sur le graphe. Au-dessus dudit seuil, il est possible de colorer la construction en fonction de sa centralité.

Les constructions dont les réalisations sont proches d'un prototype donné sont sémantiquement cohérentes. L'accès cognitif y est supposé plus rapide que pour les constructions hétérogènes, c'est-à-dire dont les réalisations impliquent une variabilité et une divergence plus grande vis-à-vis du prototype. Une fois encore, j'ai l'intime conviction que les graphes offrent une gamme de mesures permettant de capturer la connectivité entre les réalisations d'une construction et son prototype. L'une de ces mesures est la densité, qui est le rapport entre le nombre d'arêtes entre des nœuds et le nombre total d'arêtes possibles dans le graphe.

La cohésion sémantique est un pré-requis pour cerner les communautés. Une communauté est un groupe de nœuds reliés par des connexions plus denses. La densité d'un graphe est le ratio du nombre d'arêtes attestées par rapport au nombre d'arêtes possibles. Jusqu'ici, l'accent a été mis sur des communautés de sens. J'ai pour projet de modéliser des communautés de constructions sur la base du sens, certes, mais aussi de la forme.

Les défis de la visualisation

La visualisation offerte par la théorie des graphes s'appuie sur une sémiotique riche et fine. Cette sémiotique permet de cerner assez aisément la dimension collocationnelle de la plupart des constructions. La difficulté principale vient du grand nombre de nœuds et de liens impliqués dans la projection de réseaux constructionnels impliquant des individus ou des constituants à la fois variés et très représentés en corpus. Dans ce cas, le graphe devient vite chargé et il faut avoir recours à des classes intermédiaires pour limiter le nombre de nœuds. À titre d'illustration, la Figure 5.12 est la transposition sous forme de réseau d'un jeu de données utilisé dans DESAGULIER (2015b). Chaque point représente un adjectif modifié soit par *quite*, soit par *rather*, soit par les deux intensifieurs.



(a) liens multiples **(b)** liens uniques

Figure 5.12: Deux visualisations sous forme de réseau impliquant deux constructions synonymes (*quite* et *rather* dans le BNC ; 1118 nœuds)

On distingue une structure similaire à ce que nous avions en Figure 5.7, à savoir trois groupes d'adjectifs : ceux intensifiés exclusivement par *quite*, ceux intensifiés exclusivement par *rather* et ceux intensifiés par l'un ou l'autre adverbe. Il semble que cette configuration soit celle que l'on obtient lorsque l'on projette des données de collocations faisant intervenir des constructions quasi-synonymes [7]. Sur le graphe de gauche (Figure 5.12a), chaque arête correspond à une collocation entre l'une des deux constructions et un adjectif. Les nœuds sont si nombreux qu'il est impossible de projeter les étiquettes des adjectifs. Les arêtes sont encore plus nombreuses (3995) à tel point qu'il est difficile de distinguer le détail des collocations. Sur le graphe de droite (Figure 5.12b), chaque arête correspond à un type de

7. Notons que l'on a de fortes chances de retrouver cette configuration si l'on étudie deux constructions antonymes.

collocation et non plus à une occurrence. Le graphe est moins chargé mais là encore, du fait du grand nombre de combinaisons attestée, l'interprétation reste difficile.

Je compte exploiter les avancées en matière de visualisation. Celles-ci permettent de projeter des graphes en trois dimensions. C'est le cas du package `igraph` pour R. C'est aussi le cas d'un logiciel de visualisation externe à R : *Gephi*.

6 Deuxième prolongement : les vecteurs lexicaux

> *A common myth is that corpus linguists do everything automatically, which would make corpus linguistic techniques unsuited for the study of meaning.*
>
> — **Dagmar Divjak et Nick Fieller**
> (2014, p. 408)

6.1 Introduction

Je découvre les vecteurs lexicaux à l'occasion d'une séance de travail avec Antoine Chambaz à la Bibliothèque Mathématiques Informatique Recherche de L'Université Pierre et Marie Curie en décembre 2015. Nous cherchons alors un moyen d'annoter sémantiquement un grand nombre d'adjectifs dans un grand jeu de données dans le cadre d'une analyse causale en préparation. Nous découvrons que les dernières avancées en matière de réseaux neuronaux (une branche de l'intelligence artificielle) ouvrent cette possibilité, en plus de déclencher un fort enthousiasme.

J'expose ici mes travaux exploratoires sur les vecteurs lexicaux et les enjeux qui en découlent. Si d'un côté les vecteurs lexicaux incitent à l'optimisme quant à la capture du sens en contexte, ils incitent également à la prudence. Parce que les vecteurs lexicaux permettent de traiter plus de données que le cerveau humain ne peut en traiter, la linguistique de l'usage, qui vise une compréhension à l'échelle humaine de la compétence linguistique, et la linguistique de corpus, qui se distingue du traitement automatique des langues en se déclarant la seule à exiger une connaissance fine de ses échantillons, n'ont-elles pas intérêt à s'en distancier ?

6.2 Les enjeux de l'annotation sémantique à grande échelle

Traditionnellement, la première solution consiste à annoter manuellement son jeu de données à l'aide de données secondaires. L'annotation manuelle est considérée comme étant au cœur de la linguistique de corpus puisqu'elle suppose de bien connaître ses données (Divjak et Fieller, 2014, p. 408) :

> (...) at the heart of the corpus-based study of linguistic phenomena is the manual annotation of examples. This requires the analyst to read and analyze the examples one by one, just as s/he would do with a list of examples collected in a notebook or written out on cards. One of the main differences between a corpus linguist and other linguists is that the corpus linguist selects a random sample from a representative and balanced collection of texts that represent one or more varieties of the language s/he is studying.

Pour rendre l'annotation manuelle possible, l'échantillon doit être de taille modeste. Nous avons vu au Chapitre 3 que la taille d'un échantillon n'est pas considérée comme un problème en linguistique de corpus dans la mesure où elle est contrebalancée par son échantillonnage, sa représentativité et son équilibre.

La Tableau 6.1 en un exemple d'annotation sémantique manuelle d'adjectifs à partir d'un jeu de données issu de DESAGULIER (2015b). Les classes sémantiques ont été empruntées à DIXON et AIKHENVALD (2004) et ont été adaptées aux spécificités du jeu de données. L'annotation manuelle a l'avantage de résoudre les cas d'ambiguïté grâce au contexte. Par exemple, on assigne sans problème un sens différent à *cold* en fonction du NP qu'il qualifie (*a rather cold morning* vs. *a rather cold person*).

Tableau 6.1: Un exemple d'annotation sémantique manuelle d'adjectifs

fichier de corpus	construction	genre texte	expression	adjectif	classe sém. adj.
K1J.xml	pré-déterminant	W news script	*quite a hot shot*	*hot*	value_positive
G2W.xml	pré-adjectival	W pop lore	*a rather hot seller*	*hot*	value_positive
KRT.xml	pré-adjectival	S brdcast news	*a quite clear position*	*clear*	clearness
J0V.xml	pré-déterminant	W ac :humanities arts	*quite a clear understanding*	*clear*	clearness
CHE.xml	pré-déterminant	W biography	*quite a clear view*	*clear*	clearness
FEV.xml	pré-déterminant	W nonAc : nat science	*quite a clear picture*	*clear*	clearness
EWR.xml	pré-adjectival	W nonAc : polit law edu	*a quite clear line*	*clear*	clearness
CRK.xml	pré-déterminant	W religion	*quite a clear stand*	*clear*	clearness
HA7.xml	pré-adjectival	W fict prose	*a rather clouded issue*	*clouded*	unclearness
KPV.xml	pré-déterminant	S conv	*quite a cold day*	*cold*	temperature_cold
G3B.xml	pré-adjectival	W biography	*a rather cold morning*	*cold*	temperature_cold
AB5.xml	pré-adjectival	W biography	*a rather cold person*	*cold*	value_undesirable
CDB.xml	pré-déterminant	W fict prose	*rather a cold note*	*cold*	psych_stim_bad
K23.xml	pré-adjectival	W news script	*a rather colder winter*	*colder*	temperature_cold

Il est commun de faire appel à plusieurs annotateurs pour le même jeu de données. Les étiquettes sémantiques sont déterminées a priori et le codage est vérifié a posteriori à l'aide du κ (kappa) de Cohen (COHEN, 1960 ; COHEN, 1968), qui mesure l'accord entre les différents annotateurs [1]. Cette procédure est clairement exposée par GLYNN (2010, p. 250), qui annote les emplois du verbe *bother* à l'aide de quatre catégories subjectives (affect, thème, agent, cause).

Un problème lié à cette méthode est le manque de flexibilité du jeu d'étiquettes qui, une fois déterminé, ne peut plus être modifié pour mieux coller aux données. Il faut donc veiller à bien avoir ses données en tête au moment de mettre en place le protocole d'annotation. Cela suppose, une fois encore, de travailler à partir d'un jeu de données à taille humaine. Toutefois, même lorsque l'échantillon considéré est de petite taille, l'annotation sémantique manuelle est une méthode extrêmement gourmande en temps et en énergie y compris lorsque l'on choisit de répartir le travail entre plusieurs annotateurs.

1. Le score obtenu pour chacune des classes retenues est compris entre 0 et 1. Un score inférieur à 0.4 est considéré comme moyen ou mauvais, et un score supérieur à 0.6 comme bon ou excellent

Ce problème se pose avec d'autant plus d'acuité que les très grands jeux de données se multiplient de manière exponentielle à l'ère du « big data ». Objet de méfiance pour certains linguistes, qui arguent à juste titre que le travail d'un linguiste à une échelle réduite de données est de bien meilleure qualité, le « big data » nécessite des techniques élaborées. Ces techniques sont heureusement de plus à plus à la portée des linguistes à la faveur de deux facteurs :

— le travail d'excellentes équipes de traitement automatique des langues, notamment The Stanford Natural Language Processing Group [2] ;

— l'effacement progressif de la frontière entre la linguistique de corpus et le traitement automatique des langues [3].

D'un côté, ma position vis-à-vis du « big data » rejoint en partie le sentiment de méfiance partagé par la communauté linguistique. Tout jeu de données contient ou engendre du « bruit », c'est-à-dire des données non pertinentes. On cerne ce phénomène sous l'angle de la dyade « précision »/« rappel ». Dans l'interrogation d'une base de données (comme une requête en corpus), la précision est le rapport du nombre de formes pertinentes incluses dans les résultats par rapport au nombre de résultats obtenus. Le rappel est le rapport entre le nombre de résultats pertinents trouvés par rapport au total de résultats pertinents dans la base de données. En linguistique de corpus, on a tendance à optimiser le rappel de manière à minimiser les risques de « silence » (lorsque la requête ne donne rien). Cela a pour effet de diminuer la précision. Une précision diminuée suppose de nettoyer les données à la main (SMITH et al., 2008). En augmentant la taille des données, on augmente par la même occasion la quantité de nettoyage à effectuer. Avec le « big data », les résultats sont tellement nombreux que le nettoyage manuel est pratiquement impossible.

D'un autre côté, il est pour moi difficile de rester insensible à l'accumulation de jeux de données de grande taille. Après tout, à la sortie du BNC au milieu des années 90, l'un des principaux mérites formulés à son égard par ses créateurs était sa taille, à savoir près de 100 millions de mots (BURNARD, 2000). Depuis les années 50, les pionniers en matière de constitution de corpus ont fait en sorte de proposer des collections de textes de plus en plus grandes. Mark Davies procède de la sorte à chaque ajout d'un corpus à ses bases de données. Il est difficile d'imaginer pourquoi le saut quantitatif proposé par le « big data » serait moins bon que cette tendance à l'œuvre en linguistique de corpus depuis plusieurs décennies. Certes, les linguistes affirment, en grande partie à raison, qu'il existe une différence qualitative entre les corpus pour linguistes et les bases de données compilées de manière algorithmique à partir de la Toile. Les échantillons dont sont constitués les corpus sont sélectionnés pour l'analyse linguistique en vertu de leur équilibre entre les genres et de leur représentativité. Toutefois, d'autres linguistes ont montré que l'extraction à partir de la Toile pouvait être contrôlée et affinée de manière à pouvoir générer des corpus tout à fait exploitables pour une analyse linguistique fine [4].

L'annotation sémantique (semi-)automatique d'un jeu de données de plusieurs milliers d'occurrences n'est pas moins difficile et problématique que l'annotation manuelle. Plusieurs

2. `http://nlp.stanford.edu/`

3. Le laboratoire MoDyCo en est le parfait exemple puisque des linguistes collaborent au quotidient avec des spécialistes du TAL.

4. Voir le corpus constitué par Dirk Speelman (Université Catholique de Leuven) à partir du serveur de blogs *LiveJournal* (GLYNN, 2009, p. 84). Ce corpus a le mérite d'identifier la provenance des auteurs, leur âge, leur langue maternelle, d'estimer l'origine sociale et le thème du billet. Voir également les grands corpus extraits de la Toile et la méthodologie associée (BARONI et al., 2009).

obstacles se dressent sur le chemin du linguiste. Le premier est commun à l'annotation manuelle. Il concerne l'emploi de classes déterminées a priori. Le Tableau 6.2 est identique au Tableau 6.1 à l'exception du schéma d'annotation sémantique. Chaque adjectif a été annoté automatiquement par le schéma USAS (UCREL Semantic Analysis System) développé par PIAO et al. (2015).

Tableau 6.2: Un extrait de tableau de données ; les adjectifs sont annotés sémantiquement avec le schéma USAS

fichier de corpus	construction	genre texte	expression	adjectif	étiquette sém. adj.	classe sém. adj.
K1J.xml	pré-déterminant	W news script	*quite a hot shot*	*hot*	O4.6+	Temperature_Hot_on_fire
G2W.xml	pré-adjectival	W pop lore	*a rather hot seller*	*hot*	O4.6+	Temperature_Hot_on_fire
KRT.xml	pré-adjectival	S brdcast news	*a quite clear position*	*clear*	A7+	Likely
J0V.xml	pré-déterminant	W ac :humanities arts	*quite a clear understanding*	*clear*	A7+	Likely
CHE.xml	pré-déterminant	W biography	*quite a clear view*	*clear*	A7+	Likely
FEV.xml	pré-déterminant	W nonAc : nat science	*quite a clear picture*	*clear*	A7+	Likely
EWR.xml	pré-adjectival	W nonAc : polit law edu	*a quite clear line*	*clear*	A7+	Likely
CRK.xml	pré-déterminant	W religion	*quite a clear stand*	*clear*	A7+	Likely
HA7.xml	pré-adjectival	W fict prose	*a rather clouded issue*	*clouded*	O4.3	Colour_and_colour_patterns
KPV.xml	pré-déterminant	S conv	*quite a cold day*	*cold*	O4.6-	Temperature_Cold
G3B.xml	pré-adjectival	W biography	*a rather cold morning*	*cold*	B2-	Disease
AB5.xml	pré-adjectival	W biography	*a rather cold person*	*cold*	O4.6-	Temperature_Cold
CDB.xml	pré-déterminant	W fict prose	*rather a cold note*	*cold*	O4.6-	Temperature_Cold
K23.xml	pré-adjectival	W news script	*a rather colder winter*	*colder*	O4.6–	Temperature_Cold

On constate que le sens figuré des adjectifs sélectionnés est maladroitement annoté. Par exemple, lorsqu'il modifie *seller*, l'adjectif *hot* n'a que peu à voir avec une haute température. Ce n'est pas faute d'avoir appliqué un inventaire d'étiquettes sémantiques très détaillées (`http://ucrel.lancs.ac.uk/usas/`). L'algorithme ne fait qu'imposer un jeu d'étiquettes établi a priori et pré-entraîné sur une grande base lexicale. Il manque à la procédure un facteur clé : la prise en compte du contexte. C'est sur ce terrain que les vecteurs lexicaux ont une carte à jouer.

6.3 Les vecteurs lexicaux : principes

La conversion des mots en vecteurs numériques sur la base de leurs distributions n'est pas une idée nouvelle. Elle est déjà appliquée depuis un certain temps en analyse sémantique latente (ASL), utilisée pour relier des documents en vertu des termes qu'ils ont en commun (LANDAUER, 2007). L'ASL établit des relations entre des documents et les termes qu'ils contiennent à l'aide d'une matrice d'occurrences dont les lignes correspondent aux termes et les colonnes aux documents. La matrice décrit le nombre de termes dans chaque document. Ce nombre est normalisé à l'aide de la pondération TF-IDF (*term-frequency - inverse document frequency*). La matrice est ensuite transformée en une relation entre les termes et les concepts, puis entre ces concepts et les documents. Les documents sont reliés en vertu des propriétés statistiques des termes qu'ils contiennent.

Plus récemment, des méthodes plus fines ont vu le jour. Elles se fondent sur l'apprentissage profond (*deep learning*) et les réseaux neuronaux (*neural networks*). Elles consistent à apprendre des représentations lexicales à valeur prédictive sur la base du contexte local. Le processus de conversion d'une série de chaînes de caractères en vecteurs lexicaux se fait en trois étapes :

1. le corpus est divisé en mots ;
2. un apprentissage des représentations de vectorielles des mots est lancé ;

3. une matrice de vecteurs est produite.

La Figure 6.3 est extraite d'un exemple de matrice vectorielle réalisée avec *Glove* (jeu de données Common Crawl, 42 milliards de mots). La matrice, que j'ai restreinte aux adjectifs, comporte un mot par ligne. Chaque adjectif est caractérisé par un vecteur de 300 nombres (le vecteur a donc 300 dimensions). Ce vecteur est en quelque sorte la carte d'identité du mot dans le corpus. Plus deux mots sont proches, plus leurs vecteurs respectifs sont similaires. Plus le nombre de dimensions est grand, plus la représentation vectorielle des mots est fine (mais plus il est difficile de trouver des similitudes entre les mots).

Tableau 6.3: Un extrait de matrice vectorielle (10 dimensions sur 300 sont représentées)

adjectif	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	…
amoral	-0.29251	-0.43645	-0.18519	0.27531	-0.1985	0.27466	-0.75204	0.46133	0.0047043	-0.85978	…
moral	-0.48678	-0.0050737	0.58324	0.031326	-0.39364	-0.12834	-2.9809	0.85481	-0.34808	-0.57806	…
eminent	-0.36935	0.24015	0.30613	-0.13224	-0.27834	0.55996	-1.0977	0.059873	0.26424	0.26603	…
prominent	0.2528	0.40627	0.16504	-0.092403	-0.46024	-0.058152	-1.7385	-0.17295	0.076905	0.46242	…
young	-0.43338	0.30648	-0.11446	0.90696	0.20203	-0.42553	-2.9645	-0.42409	-0.0093237	-0.44171	…
younger	-0.2997	0.3935	-0.11797	0.067139	0.31185	-0.086947	-2.7571	-0.19034	-0.40016	-0.17616	…
careful	-0.18278	-0.0033721	0.0041928	-0.41133	-0.13307	0.19522	-3.22	0.39502	-0.065774	-0.25931	…
cautious	-0.17579	-0.070458	0.19399	-0.61177	-0.20105	-0.06016	-2.4399	0.39984	-0.32287	-0.20551	…
fragile	0.54108	-0.32512	0.0098112	0.3353	-0.26892	-0.15349	-2.1593	0.37341	-0.10556	-0.47822	…
delicate	0.043434	-0.24999	-0.3001	-0.34816	0.49445	-0.070195	-2.1159	-0.34641	-0.050869	-0.74683	…
effective	0.1042	0.75314	-0.24779	-0.28348	0.41798	0.053419	-3.5979	1.1205	-0.18439	-0.023927	…
ineffective	-0.26563	0.025115	-0.26815	0.019727	-0.15291	0.23698	-2.4483	0.52012	-0.27205	-0.059292	…
…	…	…	…	…	…	…	…	…	…	…	…

Deux principaux programmes permettent de décomposer des corpus en vecteurs lexicaux : *word2vec* (MIKOLOV, CHEN et al., 2013 ; MIKOLOV, SUTSKEVER et al., 2013 ; MIKOLOV, YIH et al., 2013) et *GloVe* (PENNINGTON et al., 2014) [5]. *word2vec* repose sur deux algorithmes complémentaires (RONG, 2014) :

— *CBOW* ;

— *Skip-gram*.

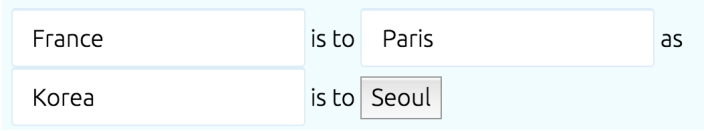
CBOW (*continuous bag of words*) permet de prédire un mot en fonction de son contexte tandis que *continuous skip-gram* permet de faire l'inverse, à savoir prédire le contexte d'un mot en fonction du mot lui-même. Les concepteurs de *word2vec* ont entraîné le programme avec le corpus Google News, qui comporte 100 milliards de mots. Radim Řehůřek a réalisé une application permettant d'évaluer manuellement la qualité de l'apprentissage (http://rare-technologies.com/word2vec-tutorial/#app). L'application propose plusieurs tests d'analogie. Deux d'entre eux sont représentés en Figure 6.2. Un troisième test n'est pas sans rappeler les tests de sémantique structurale. Il permet de déduire $queen$ de l'équation $king - man + woman$.

Les concepteurs de *GloVe* proposent une amélioration de *word2vec* :

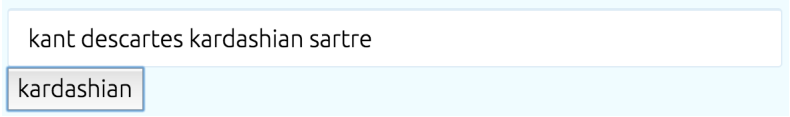
> "Methods like skip-gram may do better on the analogy task, but they poorly utilize the statistics of the corpus since they train on separate local context windows instead of on global co-occurrence counts." (PENNINGTON et al., 2014)

GloVe repose sur deux composantes. La première (des modèles de régression globale log-bilinéaire) est difficile à cerner pour qui n'est pas mathématicien. Je ne la détaille pas ici. La seconde est un modèle entraîné sur une pondération des moindres carrés. Ce modèle

5. Ces programmes sont décrits sur leurs pages respectives : https://code.google.com/p/word2vec et http://nlp.stanford.edu/projects/glove. *word2vec* est disponible pour Java (http://deeplearning4j.org/) et Python (http://rare-technologies.com/deep-learning-with-word2vec-and-gensim/). *GloVe* fonctionne sous Python.



(a) déduction sur la base d'une analogie linéaire



(b) repérage d'une incongruité

Figure 6.1: Tests d'analogie permettant d'évaluer la qualité de l'apprentissage produisant les vecteurs lexicaux

a fait l'objet d'un apprentissage sur la base de fréquences de co-occurrences de mots. Plus simplement, l'avantage de *GloVe* sur *word2vec* serait la prise en compte de cooccurrences plus globales.

Dans tous les cas, la qualité des vecteurs lexicaux dépend de la qualité de l'apprentissage. La qualité de l'apprentissage dépend de plusieurs facteurs, à savoir :

— la taille du corpus de départ ;

— le nombre de dimensions ;

— l'algorithme.

Le Tableau 6.4 compare la performance des deux programmes dans des tâches d'analogie (PENNINGTON et al., 2014). *GloVe* affiche une précision supérieure (75%) requérant près de 3,5 fois moins de dimensions, mais s'appuyant sur un corpus 7 fois plus grand.

Tableau 6.4: Comparaison des performances de *CBOW*, *Skip-gram*, et *GloVe* sur des tâches d'analogie

model	dimensions	size in words	accuracy
CBOW	1000	6B	63.7
Skip-gram	1000	6B	65.6
GloVe	300	42B	75.0

Sur la base de cette comparaison, j'ai utilisé *GloVe* pour la suite de mes exploration sur les vecteurs lexicaux.

6.4 Applications exploratoires

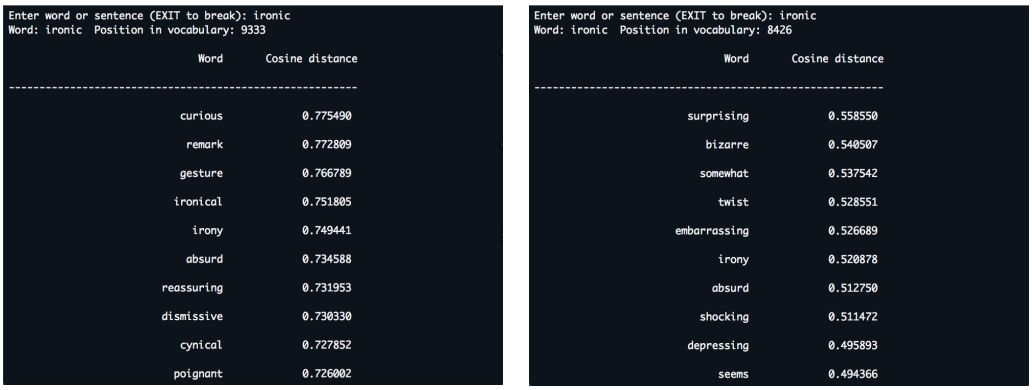
À la faveur d'une journée d'étude organisée dans le cadre du partenariat Hubert Curien « Fouille de textes : l'expression des sentiments en français et en coréen » le 13 janvier 2016 à l'Université Ajou à Suwon, j'ai réalisé des apprentissages sur deux corpus : le BNC et GloWbE. À chaque fois, mon but était d'évaluer le profil vectoriel de mots de sentiments en anglais en inspectant les similitudes lexicales sur la base du sens et du contexte syntaxique. Dans le premier cas, j'ai généré via *GloVe* une matrice de 50 dimensions. Le corpus étant de taille modeste par rapport aux bases de données traditionnellement utilisées en *deep learning*, la précision obtenue fut loin d'être excellente (Tableau 6.5, première ligne). Dans le second cas, j'ai généré une matrice de 300 dimensions. La granularité plus fine associée à un

Tableau 6.5: Comparaison des performances de *GloVe* sur le BNC et GloWbE

	semantic accuracy	syntactic accuracy	**total accuracy**
BNC	18.37%	34.40%	**27.76%**
GloWbE	52.83%	43.08%	**47.47%**

corpus d'apprentissage beaucoup plus gros ont donné de bien meilleurs résultats (Tableau 6.5, première ligne). Pour que l'apprentissage puisse se faire au niveau sémantique, il faut pouvoir disposer d'une base de données suffisamment grande et d'une granularité fine.

En dépit de performances en deçà de ce que nous obtenons à partir des bases de données aspirées sur la Toile, l'apprentissage réalisé à partir des corpus donne des résultats encourageants. La Figure 6.2 montre les termes associés à *ironic* dans chacun des corpus. La mesure cosinus utilisée (« cosine distance ») est un indice de la similitude sémantique et syntaxique entre deux vecteurs. En théorie, plus le score est élevé, plus la similitude sémantique et syntaxique est grande.



(a) BNC **(b)** GloWbE

Figure 6.2: Similitudes entre le vecteur de *ironic* et les vecteurs voisins dans le BNC et GloWbE

Dans les deux tableaux, les mots les plus similaires sont à la fois des adjectifs, des adverbes, des verbes et des noms. Ceci est dû à la nature distributionnelle des vecteurs. Deux mots employés en distribution complémentaire (par exemple *ironic* et *curious*, *ironical*, *surprising*, *bizarre*, *cynical*, etc.) ont de grandes chances d'avoir un profil vectoriel très similaire. Deux mots employés en distribution parallèle (ex. *ironic* et *remark*, *gesture*, *twist*) ont également des chances d'avoir un profil vectoriel similaire, à plus forte raison si le corpus est limité en taille (dans ce cas, un effet de biais est inévitable).

Dans le prolongement de cette étude exploratoire, j'ai cherché à examiner les distances vectorielles entre les adjectifs du jeu de données de DESAGULIER (2015b) (cf. Tableau 6.1). Afin d'éviter le biais sémantique lié à l'utilisation d'un corpus trop petit, et ne disposant pas à l'époque d'un serveur suffisamment puissant pour réaliser un apprentissage à partir d'un très grand corpus, j'ai extrait les vecteurs correspondant aux adjectifs de mon étude de la base de données `Common Crawl`. J'ai ensuite intégré ces vecteurs à mes données (Tableau 6.6). Le tableau ainsi obtenu comporte 303 colonnes.

Tableau 6.6: Extrait du jeu de données augmenté de vecteurs lexicaux pour les adjectifs

construction	intensifieur	adjectif	V1	V2	V3	V4	V5	V6	…
pré-déterminant	*rather*	*gloomy*	-0.36405	-0.44487	-0.33327	-0.16695	-0.52404	0.31066	…
pré-déterminant	*quite*	*sacred*	0.60337	-0.20526	-0.042822	-0.33008	-0.68957	0.26654	…
pré-adjectival	*rather*	*jaundiced*	-0.32168	-0.58319	-0.34614	-0.12474	0.10368	0.1733	…
pré-déterminant	*quite*	*loud*	0.24615	-0.24904	-0.18212	-0.14834	-0.06532	-0.3393	…
pré-déterminant	*quite*	*memorable*	0.30206	0.20307	0.062304	0.66816	0.048326	0.034361	…
pré-adjectival	*rather*	*justified*	-0.080959	-0.23694	-0.43372	-0.31442	-0.31528	0.0057226	…
pré-adjectival	*rather*	*scant*	-0.14467	-0.29329	0.10832	-0.11123	-0.57925	-0.27022	…
pré-déterminant	*rather*	*continuous*	-0.15253	-0.082764	-0.40871	-0.53719	0.0822	-0.31482	…
pré-déterminant	*quite*	*imposing*	0.32043	0.155	-0.10547	-0.23157	-0.35657	-0.097553	…
pré-adjectival	*rather*	*weighty*	0.085281	0.015087	0.58454	0.0094917	-0.082617	0.36811	…
…	…	…	…	…	…	…	…	…	

Le tableau obtenu étant de grande dimension (1108 lignes et 303 colonnes), je l'ai synthétisé graphiquement à l'aide de la méthode `t-SNE` (*t*-distributed Stochastic Neighbor Embedding) (VAN DER MAATEN et HINTON, 2008). Cette méthode visualise des données de grandes dimensions en attribuant à chaque point une coordonnée sur un espace à deux dimensions, à l'image de ce que proposent l'ACP, l'AFC et l'AFM (cf. Chapitre 4). La Figure 6.3 présente le résultat graphique de cette projection. La couleur des adjectifs est fonction de la construction dans laquelle ils interviennent.

Le profil vectoriel regroupe des synonymes ou quasi-synonymes (par exemple *indispensable–essential*, *tasty–delicious*, *unlikely–improbable*, *astute–shrewd*, *weak–feeble*), des paires adjectif-comparatif (par exemple *young–younger*, *short–shorter*, *heavy–heavier*) mais aussi des antonymes (par exemple *fast–slow*, *adequate–inadequate*, *faster–slower*, *lighter–heavier*, *safe–dangerous*), voire des métonymes (*frail–elderly*). En somme, plusieurs types de relations sémantiques sont ici représentées.

6.5 Discussion

Dans la mesure où je n'ai découvert les vecteurs lexicaux que récemment, je me garde de toute conclusion à leur égard. Il peut sembler futile de parler de sémantique lorsqu'il s'agit d'apprentissage. L'une des questions que l'on peut légitimement se poser est la suivante : la machine connaît-elle le sens des mots vectorisés ? La réponse est bien entendu négative. Le sens dont il est question ici n'est qu'affaire de distribution lexicale. Mais si l'on en croit la position d'une grande partie des linguistes de corpus, synthétisée dans la citation en exergue du Chapitre 3, la linguistique est une science distributionnelle. Les régularités sémantiques mises en avant par la vectorisation des mots est d'autant plus intéressante qu'elle fait au moins partiellement écho à la classification humaine et qu'elle est l'œuvre d'une machine qui ne peut pas avoir accès au sens.

Une autre question légitime est la suivante : les vecteurs lexicaux sont-ils arbitraires ? S'ils le sont, ils le sont moins que les schémas d'annotation décontextualisés dont il a été question plus haut. De plus, parce qu'ils sont en prise avec les collocations et plus largement le contexte, ils ne peuvent que trouver leur place dans la linguistique de l'usage. Enfin, la langue étant par essence arbitraire (KILGARRIFF, 2005), la question n'a pas vraiment lieu de se poser.

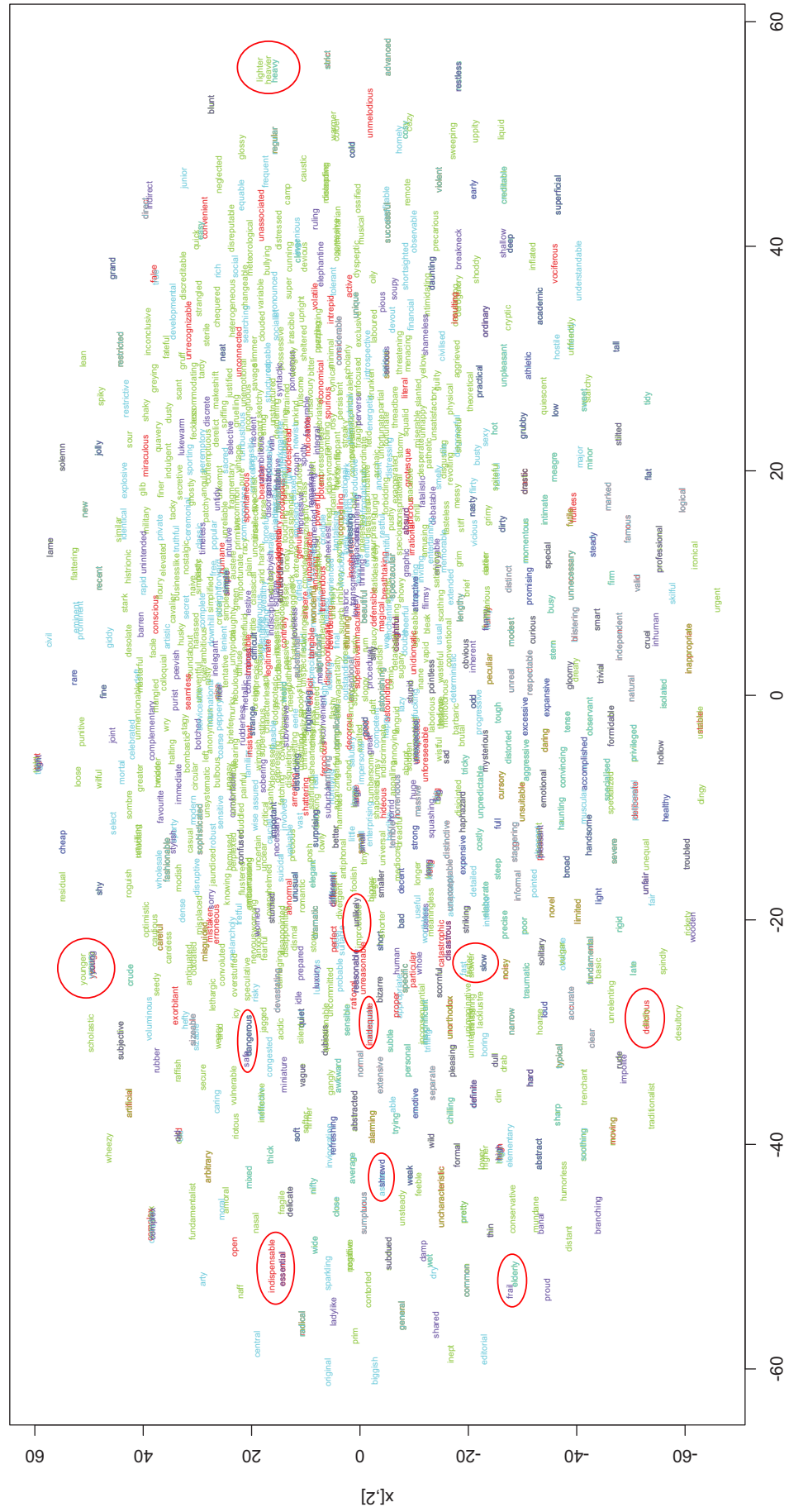


Figure 6.3: Projection des adjectifs sur la base de leurs profils vectoriels (*rather* pré-adjectival, *rather* pré-déterminant, *quite* pré-adjectival, *quite* pré-déterminant)

Il est évident que les réseaux neuronaux et les vecteurs lexicaux ont encore besoin de quelques années pour s'affiner. Quoi qu'il en soit, même s'ils ne sont qu'un outil de plus dans la panoplie du linguiste de corpus, ils ouvrent des pistes de recherche qu'il me semble intéressant d'explorer.

7 Conclusion

À une époque où la rédaction de candidatures à des appels à projets empiète de plus en plus sur la recherche pure, je considère la synthèse comme étant un exercice salutaire. Elle dégage, tout autant qu'elle fonde, une cohérence dans mon parcours de chercheur. Privilège rare à ce moment de ma carrière, elle me permet de partager mes recherches avec un panel d'experts reconnus dans mon domaine.

Une synthèse est à la fois rétrospective et prospective. Après avoir brièvement récapitulé les idées forces présentées dans les chapitres précédents, j'aborde des projets qui m'attendent, la plupart étant déjà amorcés.

7.1 Bilan

Avant d'opter pour le présent chapitrage, j'ai envisagé une autre organisation. Celle-ci aurait consisté en des entrées thématique centrées sur les études de cas. J'aurai pu ainsi distinguer nettement l'étude de la socio-pragmatique (via les constructions directives), l'étude de la sémantique constructionnelle (via la conversion du comptable au massif, l'intensification), et l'étude de phénomènes plus syntaxiques (via le phénomène d'alternance). Au lieu de cela, je suis resté fidèle à la linéarité de mes publications afin de souligner que chaque réorientation était le produit d'une réflexion influencée par les débats qui ont animé la linguistique cognitive et les grammaires de constructions depuis ma soutenance de thèse jusqu'à aujourd'hui.

Formé à la linguistique cognitive de première génération, je me suis tout d'abord prononcé sans complexes sur des questions touchant à la représentation psychologique de phénomènes langagiers. Convaincu par l'idée selon laquelle la grammaire est un réseau d'unités symboliques, ces unités étant des constructions, je me suis tourné vers les méthodes empiriques de la linguistique cognitive de deuxième génértion pour contribuer au débat de manière tangible. Doutant l'espace de quelques mois, à la fin des années 2000, de la légitimité de la linguistique cognitive au sein des sciences cognitives [1], c'est paradoxalement le détour par la linguistique de corpus et les statistiques qui m'a convaincu du bien-fondé du « cognitive commitment ». Linguiste cognitiviste, je n'oublie pas pour autant qu'un cadre théorique n'est pas un vase clos et qu'il faut savoir dialoguer avec des approches complémentaires et des disciplines parfois étrangères pour répondre aux questions touchant à l'usage et à la compétence.

1. Un comble pour moi qui suis l'actuel président de l'Association Française de Linguistique Cognitive.

7.2 Perspectives

Le moment rétrospectif qu'offre la synthèse ne dure qu'un temps et je suis déjà investi dans d'autres projets liés aux thèmes des Chapitres 5 et 6. Le premier de ces projets est une co-direction de thèse dans le cadre d'une cotutelle entre l'Université Paris Ouest Nanterre et l'Université Ajou à Suwon (Corée du Sud). Suite à l'établissement d'une convention entre les deux universités, Monsieur Seongmin Mun est aujourd'hui officiellement inscrit en thèse à Nanterre. Il travaille sur la quantification et la visualisation de données linguistiques issues d'une étude de cas portant sur les médias et les nouvelles technologies. La thèse est dirigée conjointement par M. Kyungwon Lee (Université Ajou, Département des Médias Numériques, `http://madang.ajou.ac.kr/~kwlee/`), Mme Anne Lacheret (Université Paris Ouest Nanterre) et moi-même. L'obtention de l'Habilitation à Diriger les Recherches me permettra d'assurer, seul, la co-direction en France. Dans les faits, Seongmin Mun travaille actuellement au laboratoire MoDyCo sous ma direction. Il souhaite exploiter les données qu'il a collectées au cours de fouilles d'opinions sur la Toile. Ces données sont les avis laissés par les spectateurs après le visionnage d'un film. Je lui ai demandé de se former aux techniques de sémantique distributionnelle (cf. Chapitre 6). L'objectif est de voir si les vecteurs lexicaux permettent une capture plus fine de l'opinion des consommateurs. Au premier semestre de cette année, Seongmin Mun va se concentrer sur les lexèmes. Au second semestre, il abordera les constructions complexes.

Le second projet est la rédaction d'une monographie sur les grammaires de constructions : *Distributional Construction Grammar*. Cet ouvrage, dont j'ai prévu de débuter la rédaction au printemps 2017 a pour but de faire le point sur l'apport des techniques de corpus et des méthodes quantitatives en grammaires de constructions. Il propose un bilan critique de la sémantique distributionnelle. Celle-ci offre autant d'espoirs qu'elle suscite de méfiance. À l'heure où j'écris ces lignes, Adele Goldberg me fait part d'une journée d'étude consacrée à l'apprentissage profond et aux réseaux neuronaux artificiels à laquelle elle a été conviée le 13 août 2016 à Bolzano (Italie). Cette journée, organisée par Marco Baroni, a donné lieu à plusieurs interventions sur les vecteurs lexicaux. Le sentiment général est le suivant : le rôle de la sémantique distributionnelle dans le cadre de la sémantique de l'usage et des grammaires de constructions est indéniable. Toutefois, en l'état actuel des connaissances, les techniques de vectorisation sont en retard par rapport à l'agenda suggéré par la théorie de l'usage. C'est précisément ce décalage qui alimente mon attrait pour ces techniques. Dans ma monographie, je propose d'émettre des desiderata pour que l'intelligence artificielle soit plus en prise avec la spécificité des phénomènes langagiers.

Quelle que soit la direction que prendra ma recherche future, celle-ci s'appuiera sur des disciplines connexes à la linguistique au sein des humanités numériques. De par l'importance qu'elle accorde à l'étude du sens en contexte, à l'interaction et à la variation, la linguistique a une valeur ajoutée que n'ont pas nécessairement les techniques de fouille de textes ou de traitement automatique des langues. À condition de ne pas perdre de vue son objectif principal (décrire et rendre compte des structures de la langue et du langage), la linguistique n'a pas à craindre de se diluer à l'heure où le traitement des données de masse a le vent en poupe.

Bibliographie

ALLAN, Keith (1980). « Nouns and Countability ». In : *Language* 56.3, p. 541–567 (cf. p. 33).

ALLAN, Lorraine G. (1980). « A note on measurement of contingency between two binary variables in judgment tasks ». In : *Bulletin of the Psychonomic Society* 15.3, p. 147–149 (cf. p. 74).

ANDOR, József (2004). « The master and his performance : An interview with Noam Chomsky ». In : *Intercultural Pragmatics* 1.1, p. 93–111 (cf. p. 54).

BAAYEN, Rolf Harald (1989). *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation*. Amsterdam : Centrum Wiskunde en Informatica (cf. p. 84).

– (1993). « On frequency, transparency and productivity ». In : *Yearbook of Morphology 1992*. Sous la dir. de Geert BOOIJ et Jaap van MARLE. Dordrecht ; London : Kluwer, p. 181–208 (cf. p. 84, 87).

– (2001). *Word Frequency Distributions*. Dordrecht : Kluwer Academic Publishers (cf. p. 87).

– (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge : Cambridge University Press (cf. p. 5).

– (2009). « Corpus linguistics in morphology : Morphological productivity ». In : *Corpus Linguistics. An International Handbook*. Sous la dir. d'Anke LÜDELING et Merja KYTÖ. Berlin : Mouton de Gruyter, p. 899–919 (cf. p. 87).

– (2013). *languageR : Data sets and functions with "Analyzing Linguistic Data : A practical introduction to statistics"*. R package version 1.4.1 (cf. p. 102).

BAAYEN, Rolf Harald et Rochelle LIEBER (1991). « Productivity and English derivation : A corpus-based study ». In : *Linguistics* 29, p. 801–843 (cf. p. 84).

BALLY, Charles (1921). *Traité de stylistique française*. 2nde édition. Indogermanische Bibliotek Abteilung 2 Sprachwissenschaftliche Gymnasial-bibliothek. Heidelberg : C. Winter (cf. p. 16).

BARÐDAL, Jóhanna (2008). *Productivity : Evidence from Case and Argument Structure in Icelandic*. Amsterdam : John Benjamins (cf. p. 39, 40).

BARLOW, Michael et Suzanne KEMMER (2000). *Usage-Based Models of Language*. Stanford : CSLI Publications (cf. p. 48).

BARONI, Marco, Silvia BERNARDINI, Adriano. FERRARESI et Eros ZANCHETTA (2009). « The WaCky Wide Web : A Collection of Very Large Linguistically Processed Web-Crawled Corpora ». In : *Language Resources and Evaluation* 43.3, p. 209–226 (cf. p. 42, 107, 113).

BERGEN, Benjamin K. et Nancy CHANG (2005). « Embodied Construction Grammar in simulation-based language understanding ». In : *Construction Grammars : Cognitive Grounding and Theoretical Extensions*. Sous la dir. de Jan-Ola ÖSTMAN et Mirjam FRIED. Amsterdam ; Philadelphia : John Benjamins, p. 147–190 (cf. p. 22).

BERMÚDEZ-OTERO, Ricardo, David DENISON, Richard M. HOGG et C.B. MCCULLY (2000). *Generative Theory and Corpus Studies*. Berlin (cf. p. 36).

« Beyond Linguistic Wars. An Interview with Noam Chomsky » (2011). In : *Intellectica* 56.2, p. 21–27 (cf. p. 15).

BIBER, Douglas (1993). « Representativeness in Corpus Design ». In : *Literary and Linguistic Computing* 8.4, p. 241–257 (cf. p. 42, 43).

BOLINGER, Dwight (1981). *Aspects of Language*. New York : Harcourt Brace Jovanovich (cf. p. 25).

BRATMAN, Michael E. (1992). « Shared cooperative activity ». In : *The Philosophical Review* 101.2, p. 327–341 (cf. p. 38).

– (1993). « Shared intention ». In : *Ethics* 104.1, p. 97–113 (cf. p. 38).

– (1997). « I intend that we ». In : sous la dir. de Ghita HOLMSTRÖM-HINTIKKA et Raimo TUOMELA. T. 2. Dordrecht : Kluwer, p. 49–63 (cf. p. 38).

BRESNAN, Joan (1982). *The Mental Representation of Grammatical Relations*. Cambridge, Mass. : MIT Press (cf. p. 2).

BRESNAN, Joan, Anna CUENI, Tatiana NIKITINA et R. Harald BAAYEN (2007). « Predicting the dative alternation ». In : *Cognitive Foundations of Interpretation*, p. 69–94 (cf. p. 90, 91, 102, 103).

BROWN, Penelope et Stephen C. LEVINSON (1987). *Politeness : Some Universals in Language Usage*. T. 4. Studies in interactional sociolinguistics. Cambridge ; New York : Cambridge University Press (cf. p. 25, 27).

BUNT, Harry C. (1985). *Mass Terms and Model-Theoretic Semantics*. T. 42. Cambridge Studies in Linguistics. Cambridge ; New York : Cambridge University Press (cf. p. 32).

BURNARD, Lou (2000). *Reference Guide for the British National Corpus (World Edition)*. Web Page (cf. p. 113).

BYBEE, Joan L. (1985). *Morphology : A Study of the Relation between Meaning and Form*. Amsterdam : John Benjamins (cf. p. 7, 9, 40, 48).

– (2001). *Phonology and Language Use*. Cambridge : Cambridge University Press (cf. p. 40).

– (2006). « From Usage to Grammar : The Mind's Response to Repetition ». In : *Language* 82.4, p. 711–733 (cf. p. 48).

– (2007). *Frequency of Use and the Organization of Language*. Oxford ; New York : Oxford University Press (cf. p. 48).

– (2010). *Language, Usage, and Cognition*. Cambridge : Cambridge University Press (cf. p. 40, 48).

CAFFI, Claudia (1999). « On mitigation ». In : *Journal of Pragmatics* 31.7, p. 881–909 (cf. p. 26).

– (2007). *Mitigation*. Amsterdam : Elsevier (cf. p. 25, 26).

CANCHO, Ramon Ferrer i et Ricard V. SOLÉ (2001). « The Small World of Human Language ». In : *Proceedings of The Royal Society of London. Series B, Biological Sciences* 268, p. 2261–2265 (cf. p. 100).

CHAMBAZ, Antoine et Guillaume DESAGULIER (2016). « Predicting Is Not Explaining : Targeted Learning of the Dative Alternation ». In : *Journal of Causal Inference* 4.1, p. 1–30 (cf. p. 5, 52, 90–93).

CHENG, Chung-Ying (1973). « Response to Moravcsik ». In : *Approaches to Natural Language*. Sous la dir. de Jaakko HINTIKKIA, Julius Matthew Emil MORAVCSIK et Patrick SUPPES. Dordrecht : Reidel, p. 286–288 (cf. p. 32).

CHOMSKY, Noam (1957). *Syntactic Structures*. The Hague : Mouton (cf. p. 51).

– (1995). *The Minimalist Program*. Cambridge, MA : MIT Press (cf. p. 2, 17).

CHURCH, Kenneth, William A. GALE, Patrick HANKS et Donald HINDLE (1991). « Using statistics in lexical analysis ». In : *Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon*. Sous la dir. d'Uri ZERNIK. Hillsdale : Lawrence Erlbaum, p. 115–164 (cf. p. 68).

COHEN, Jacob (1960). « A coefficient of agreement for nominal scales ». In : *Educational and Psychological Measurement* 20, p. 37–46 (cf. p. 112).

– (1968). « Weighted kappa : Nominal scale agreement with provision for scaled disagreement or partial credit ». In : *Psychological Bulletin* 70.4, p. 213–220 (cf. p. 112).

CROFT, William (1995). « Some issues in frame (domain) representation : The case of *eat* and *feed* ». Unpublished paper presented to the Department of Linguistics. The University of Colorado, Boulder (cf. p. 9).

– (1998). « Linguistic evidence and mental representations ». In : *Cognitive Linguistics* 9.2, p. 151–174 (cf. p. 6, 8, 9).

– (2000). *Explaining Language Change : An Evolutionary Approach*. Harlow : Longman, xv, 287 p. (Cf. p. 6).

– (2001). *Radical construction grammar : syntactic theory in typological perspective*. Oxford : Oxford University Press (cf. p. 6, 22, 95).

– (2009). « Toward a social cognitive linguistics ». In : *New Directions in Cognitive Linguistics*. Sous la dir. de Vyvyan EVANS et Stéphanie POURCEL. Amsterdam : John Benjamins, p. 395–420 (cf. p. 38).

CROFT, William et Timothy C. CLAUSNER (1997). « Productivity and schematicity in metaphors ». In : *Cognitive Science* 21.3, p. 247–282 (cf. p. 40).

CROFT, William et D. A. CRUSE (2004). *Cognitive Linguistics*. Cambridge ; New York : Cambridge University Press (cf. p. 18, 22).

CRYSTAL, David et Donald A. SEARS (2008). *A Dictionary of Linguistics and Phonetics*. 6e édition. Malden, MA ; Oxford : Blackwell (cf. p. 25).

CULICOVER, Peter W. et Ray JACKENDOFF (1999). « The View from the Periphery : The English Comparative Correlative ». In : *Linguistic Inquiry* 30.4, p. 543–571 (cf. p. 18).

CULICOVER, Peter William et Ray JACKENDOFF (2005). *Simpler Syntax*. Oxford : Oxford University Press (cf. p. 2).

CZERWIONKA, Lori (2012). « Mitigation : The combined effects of imposition and certitude ». In : *Journal of Pragmatics* 44.10, p. 1163–1182 (cf. p. 25).

DAMOURETTE, Jacques et Édouard PICHON (1930). *Des mots à la pensée ; essai de grammaire de la langue française*. Paris : Collection des linguistes contemporains (cf. p. 32).

DAVIES, Mark (2013). *Corpus of Global Web-Based English : 1.9 billion words from speakers in 20 countries* (cf. p. 52).

– (2015). *The Wikipedia Corpus : 4.6 million articles, 1.9 billion words*. Adapted from Wikipedia (cf. p. 52).

– (2016a). *Hansard Corpus (British Parliament) : 1803–2005, 7.6 million speeches, 1.6 billion words* (cf. p. 52).

– (2016b). *NOW corpus* (cf. p. 53).

DE SWART, Henriëtte (1998). « Aspect shift and coercion ». In : *Natural Language & Linguistic Theory* 16.2, p. 347–385 (cf. p. 33).

DESAGULIER, Guillaume (2003). « *Want to/wanna* : verbal polysemy versus constructional compositionality ». In : *Berkeley Linguistics Society* 29, p. 91–102 (cf. p. 1).

– (2005). « Grammatical blending and the conceptualization of complex cases of interpretational overlap : The case of *want to/wanna* ». In : *Annual Review of Cognitive Linguistics* 3, p. 22–40 (cf. p. 1, 24).

– (2007a). « A cognitive perspective on the indirect framing of directive constructions ». Second international AFLiCo conference. Université Charles de Gaulle–Lille 3 (cf. p. 24).

DESAGULIER, Guillaume (2007b). « Figures et forces en linguistique cognitive : pour une redéfinition du concept de représentation dans une Grammaire de Constructions Floue ». In : *Théorie, Littérature, Enseignement* 24, p. 95–113 (cf. p. 3, 18, 23).

– (2007c). « Quelques remarques sur l'origine des Grammaires de Constructions et le statut des représentations en linguistique ». Base de données du PRI Anthropologie et Linguistique (cf. p. 2).

– (2008a). « Cognitive Arguments for a Fuzzy Construction Grammar ». In : *Du fait grammatical au fait cognitif/From Gram to Mind : Grammar as Cognition*. Sous la dir. de Guillaume DESAGULIER, Jean-Baptiste GUIGNARD et Jean-Rémi LAPAIRE. Bordeaux : Presses Universitaires de Bordeaux, p. 125–150 (cf. p. 3, 23).

– (2008b). « Constructions, mental representations, and sociopragmatics : The case of pseudo-directive constructions ». Language, Communication and Cognition. Université de Brighton (cf. p. 24).

– (2010). « Intégration conceptuelle, représentation et performativité ». AFLiCo JET 2010, What Type of Cognition for Cognitive Linguistics ? Université Bordeaux Montaigne (cf. p. 37).

– (2011a). « Le programme sociopragmatique des grammaires de constructions : bilan et perspectives ». In : *Intellectica* 56, p. 99–123 (cf. p. 3, 19, 37, 38).

– (2011b). « Quelles alternatives linguistiques à la théorie des faces ? Le cas du japonais ». In : *Japon pluriel 8 : La modernité japonaise en perspective*. Sous la dir. de Noriko BERLINGUEZ KÔNO et Thomann BERNARD. Paris : Philippe Picquier, p. 33–42 (cf. p. 3).

– (2012a). « *C'est de la bombe !* Qualitative count-to-mass conversion in French copular subject-predicate constructions ». In : *Constructions in French*. Sous la dir. de Myriam BOUVERET et Dominique LEGALLOIS. Amsterdam : John Benjamins, p. 201–232 (cf. p. 5, 32, 46).

– (2012b). « The vagaries of frequency (keynote) ». Sorbonne Nouvelle University Graduate Linguistics Symposium. Université Sorbonne Nouvelle–Paris3 (cf. p. 48, 60, 77).

– (2014). « Visualizing distances in a set of near synonyms : *rather*, *quite*, *fairly*, and *pretty* ». In : *Corpus Methods for Semantics : Quantitative Studies in Polysemy and Synonymy*. Sous la dir. de Dylan GLYNN et Justyna ROBINSON. Amsterdam : John Benjamins, p. 145–178 (cf. p. 5, 49, 67–70, 77–80, 95).

– (2015a). « A lesson from associative learning : asymmetry and productivity in multiple-slot constructions. » In : *Corpus Linguistics and Linguistic Theory* (cf. p. 5, 19, 39, 40, 67, 68, 71, 72, 74, 76, 84, 85, 87–89, 104).

– (2015b). « Forms and meanings of intensification : a multifactorial comparison of *quite* and *rather* ». In : *Anglophonia* 20.2 (cf. p. 5, 50, 51, 67–70, 77, 80–85, 109, 112, 117).

– (2015c). « Le statut de la fréquence dans les grammaires de constructions : *simple comme bonjour* ? » In : *Langages* 197.1, p. 99–128 (cf. p. 5, 60, 61, 67, 68, 71–74).

– (à paraître). *Corpus Linguistics and Statistics with R*. New York : Springer (cf. p. 5, 12, 42–45, 53, 57, 61–64, 67–73, 77, 79, 81, 84, 87, 89–92).

DIRVEN, René, Louis GOOSENS, Yvan PUTSEYS et Emma VORLAT (1982). *The Scene of Linguistic Action and its Perspectivization by Speak, Talk, Say and Tell*. Amsterdam : John Benjamins (cf. p. 59).

DIRVEN, René et John R. TAYLOR (1988). « The conceptualisation of vertical space in English : the case of tall ». In : *Topics in Cognitive Linguistics*. Sous la dir. de Brygida RUDZKA-OSTYN. Amsterdam ; Philadelphia : John Benjamins (cf. p. 59).

DIVJAK, Dagmar et Nick FIELLER (2014). « Cluster analysis : Finding structure in linguistic data ». In : *Corpus Methods for Semantics : Quantitative Studies in Polysemy and Synonymy*. Sous la dir. de Dylan GLYNN et Justyna ROBINSON. Amsterdam : John Benjamins, p. 405–441 (cf. p. 111).

DIXON, Robert M. W. et Alexandra Y. AIKHENVALD (2004). *Adjective Classes : A Cross-Linguistic Typology*. Oxford : Oxford University Press (cf. p. 112).

DOI, Takeo (1973). *The Anatomy of Dependence*. Tokyo, New York : Kodansha International (cf. p. 3).

– (1986). *The Anatomy of Self : The Individual versus Society*. Tokyo, New York : Kodansha International (cf. p. 3).

DOWNING, Angela et Philip LOCKE (2006). *English grammar : A University Course*. 2nde édition. London : Routledge (cf. p. 80).

DRIEGER, Philipp (2013). « Semantic network analysis as a method for visual text analytics ». In : *Procedia – Social and Behavioral Sciences* 79, p. 4–17 (cf. p. 100).

DUNNING, Ted (1993). « Accurate methods for the statistics of surprise and coincidence ». In : *Computational Linguistics* 19.1, p. 61–74 (cf. p. 69).

ELLIS, Nick C. (2006). « Language acquisition as rational contingency learning ». In : *Applied Linguistics* 27.1, p. 1–24 (cf. p. 74).

ELLIS, Nick C. et Fernando FERREIRA-JUNIOR (2009). « Constructions and their acquisition : Islands and the distinctiveness of their occupancy ». In : *Annual Review of Cognitive Linguistics* 7, p. 187–220 (cf. p. 74, 100).

ELLIS, Nick C., Matthew Brook O'DONNELL et Ute RÖMER (2014). « The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality ». In : *Cognitive Linguistics* 25.1, p. 55–98 (cf. p. 97, 100).

ELMAN, Jeffrey, Elizabeth A. BATES, Mark H. JOHNSON et al. (1996). *Rethinking Innateness : A Connectionist Perspective on Development*. Cambridge, MA : MIT press (cf. p. 96).

ELMAN, Jeffrey et James L. MCCLELLAND (1984). « Speech perception as a cognitive process : the interactive activation model ». In : *Speech and Language*. Sous la dir. de Norman LASS. T. 10. New York : Academic Press, p. 337–374 (cf. p. 96).

ERDŐS, Paul et Alfréd RÉNYI (1959). « On random graphs ». In : *Publicationes Mathematicae* 6, p. 290–297 (cf. p. 100).

EVERITT, Brian S., Sabine LANDAU, Morven LEESE et Daniel STAHL (2011). *Cluster Analysis*. T. 5. Wiley series in probability and statistics. Oxford : Wiley Blackwell (cf. p. 78).

EVERT, Stefan (2005). « The statistics of word cooccurrences : word pairs and collocations ». Unpublished Work. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart (cf. p. 68).

– (2009). « Corpora and collocations ». In : *Corpus Linguistics : An International Handbook*. Sous la dir. d'Anke LÃŒDELING et Merja KYTÃ¶. T. 2. Berlin, New York : Mouton de Gruyter, p. 1212–1248 (cf. p. 68).

EVERT, Stefan et Marco BARONI (2006). « The *zipfR* library : Words and other rare events in R ». Presentation at useR ! 2006 : The Second R User Conference, Vienna, Austria (cf. p. 87).

– (2007). « *zipfR* : Word frequency distributions in R ». In : *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Posters and Demonstration Sessions*. (R package version 0.6-6 of 2012-04-03). Prague, Czech Republic, p. 29–32 (cf. p. 87).

FAUCONNIER, Gilles (1985). *Mental Spaces : Aspects of Meaning Construction in Natural Language*. Cambridge, Mass. : MIT Press (cf. p. 1, 10).

– (1997). *Mappings in Thought and Language*. Cambridge, New York : Cambridge University Press (cf. p. 1, 10).

FAUCONNIER, Gilles et Mark TURNER (2002). *The Way We Think : Conceptual Blending and the Mind's Hidden Complexities*. New York : Basic Books (cf. p. 1).

FERRARESI, Adriano, Eros ZANCHETTA, Marco BARONI et Silvia BERNARDINI (2008). « Introducing and evaluating ukWaC, a very large web-derived corpus of English ». In : *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, p. 47–54 (cf. p. 52).

FILLMORE, Charles (1997). *Construction Grammar lecture notes*. Manuscript (cf. p. 19).

– (2002). « "Idiomaticity" » (cf. p. 39).

FILLMORE, Charles et Paul KAY (1995). *Construction Grammar*. Manuscript. Department of Linguistics, University of California, Berkeley (cf. p. 19, 33).

FILLMORE, Charles, Paul KAY et Catherine O'CONNOR (1988). « Regularity and Idiomaticity in Grammatical Constructions : The Case of *let alone* ». In : *Language* 64.3, p. 501–538 (cf. p. 16–19, 95).

FIRTH, J. R. (1957). « A synopsis of linguistic theory 1930-55. » In : *Studies in Linguistic Analysis (special volume of the Philological Society)*. T. 1952-59. Oxford : The Philological Society, p. 1–32 (cf. p. 47, 73).

FITCH, W. Tecumseh, Marc D. HAUSER et Noam CHOMSKY (2005). « The evolution of the language faculty : clarifications and implications ». In : *Cognition* 97, p. 179–210 (cf. p. 2).

FLORES-FERRÁN, Nydia et Kelly LOVEJOY (2015). « An examination of mitigating devices in the argument interactions of L2 Spanish learners ». In : *Journal of Pragmatics* 76, p. 67–86 (cf. p. 25).

FRANCIS, Gill, Susan HUNSTON et Elizabeth MANNING (1996). *Grammar Patterns. 1, Verbs*. London : Harper Collins (cf. p. 107).

FRANCIS, W. Nelson et Henry KUČERA (1979). *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Department of Linguistics. Brown University (cf. p. 43).

FRANÇOIS, Jacques (2008). « Les grammaires de constructions : un bâtiment ouvert aux quatre vents ». Cahier du CRISCO n^{o}26 (cf. p. 15).

FRASER, Bruce (1975). « Hedged performatives ». In : *Syntax and Semantics*. Sous la dir. de Peter COLE et Jerry L. MORGAN. T. Vol. 3 : Speech Acts. New York : Harcourt Brace et Jovanovitch, p. 187–210 (cf. p. 25).

– (1980). « Conversational mitigation ». In : *Journal of Pragmatics* 4.4, p. 341–350 (cf. p. 25).

– (1990). « Perspectives on politeness ». In : *Journal of Pragmatics* 14.2, p. 219–236 (cf. p. 25).

– (2010). « Pragmatic competence : The case of hedging ». In : *New Approaches to Hedging*. Sous la dir. de Gunther KALTENBÖCK, Wiltrud MIHATSCH et Stefan SCHNEIDER. Studies in Pragmatics. Bingley : Emerald, p. 15–34 (cf. p. 25).

FRIED, Mirjam et Jan-Ola ÖSTMAN (2004). *Construction Grammar in a Cross-Language Perspective*. Amsterdam ; Philadelphia : John Benjamins (cf. p. 37).

GALMICHE, Michel (1989). « Massif/comptable : de l'un à l'autre et inversement ». In : *Termes massifs et termes comptables*. Sous la dir. de Jean DAVID et Georges KLEIBER. Recherches Linguistiques. Paris : Klincksieck, p. 63–77 (cf. p. 33).

GEERAERTS, Dirk (2000). « Salience phenomena in the lexicon : A typology ». In : *Meaning and Cognition*, p. 79–101 (cf. p. 61).

– (2010a). « The doctor and the semantician ». In : *Quantitative Methods in Cognitive Semantics : Corpus-Driven Approaches*. Sous la dir. de Dylan GLYNN et Kerstin FISCHER. Berlin, New York : Mouton de Gruyter, p. 61–78 (cf. p. 12).

– (2010b). *Theories of Lexical Semantics*. Oxford ; New York : Oxford University Press (cf. p. 60).

GIBBS, Raymond W. (1983). « Do people always process the literal meanings of indirect requests ? » In : *Journal of Experimental Psychology : Learning, Memory, and Cognition* 9.524–533 (cf. p. 27).

– (1994). *The Poetics of Mind : Figurative Thought, Language, and Understanding*. Cambridge, New York : Cambridge University Press (cf. p. 9).

– (2007). « Why cognitive linguists should care more about empirical methods ». In : *Methods in Cognitive Linguistics*. Sous la dir. de Monica GONZALEZ-MARQUEZ, Irene MITTELBERG, Seana COULSON et Michael J. SPIVEY. Amsterdam : John Benjamins, p. 2–18 (cf. p. 1, 10, 11).

GILQUIN, Gaëtanelle (2013). « Making sense of collostructional analysis : On the interplay between verb senses and constructions ». In : *Constructions and Frames* 5.2, p. 119–142 (cf. p. 68).

GINZBURG, Jonathan et Ivan A. SAG (2000). *Interrogative investigations : the Form, Meaning, and Use of English Interrogatives*. Stanford, Calif. : CSLI Publications (cf. p. 2).

GLYNN, Dylan (2009). « New Directions in Cognitive Lnguistics ». In : sous la dir. de Vyvyan EVANS et Stéphanie POURCEL. Amsterdam ; Philadelphia : John Benjamins, p. 77–104 (cf. p. 113).

– (2010). « Testing the hypothesis : Objectivity and verification in usage-based cognitive semantics ». In : *Corpus-Driven Cognitive Semantics. Quantitative Approaches*. Sous la dir. de Dylan GLYNN et Kerstin FISCHER. Berlin : Mouton de Gruyter, p. 239–270 (cf. p. 112).

– (2014). « Polysemy and synonymy. Cognitive theory and corpus method ». In : *Corpus Methods for Semantics : Quantitative Studies in Polysemy and Synonymyemantics. Quantitative studies in polysemy and synonymy*. Sous la dir. de Dylan GLYNN et Justyna ROBINSON. Amsterdam ; Philadelphia : John Benjamins, p. 7–38 (cf. p. 50).

GLYNN, Dylan et Kerstin FISCHER (2010). « Corpus-driven Cognitive Semantics. Introduction to the field ». In : *Quantitative Methods in Cognitive Semantics : Corpus-Driven Approaches*. Berlin : Mouton de Gruyter, p. 1–42 (cf. p. 52).

GOLDBERG, Adele E. (1995). *Constructions : a Construction Grammar Approach to Argument Structure*. Chicago : University of Chicago Press (cf. p. 2, 10, 15, 17–20, 95, 97).

– (2003). « Constructions : a new theoretical approach to language ». In : *Trends in Cognitive Sciences* 7.5, p. 219–224 (cf. p. 20).

– (2006). *Constructions at Work : the Nature of Generalization in Language*. Oxford ; New York : Oxford University Press (cf. p. 2, 18–20, 37).

– (2009). « Constructions work ». In : *Cognitive Linguistics* 20.1, p. 201–224 (cf. p. 20).

– (2016). « Partial productivity of linguistic constructions : Dynamic categorization and statistical preemption ». In : *Language and Cognition* 8 (Special Issue 3), p. 369–390 (cf. p. 40).

GOLDBERG, Adele E. et Ray JACKENDOFF (2004). « The English Resultative as a Family of Constructions ». In : *Language* 80.3, p. 532–568 (cf. p. 17).

GOLDBERG, Adele E. et Johan VAN DER AUWERA (2012). « This is to count as a construction ». In : *Folia Linguistica* 46.1, p. 109–132 (cf. p. 97, 98).

GOODWIN, Charles (2003). *Conversation and Brain Damage*. Oxford ; New York : Oxford University Press (cf. p. 2).

GOODWIN, Marjorie Harness (2006). *The Hidden Life of Girls : Games of Stance, Status, and Exclusion*. Oxford : Blackwell (cf. p. 2).

GRIES, Stefan Thomas (1999). « Particle movement : A cognitive and functional approach ». In : *Cognitive Linguistics* 10.1, p. 105–145 (cf. p. 10).

– (2003a). *Multifactorial Analysis in Corpus Linguistics : A Study of Particle Placement*. Open linguistics series. New York : Continuum (cf. p. 11).

– (2003b). « Towards a corpus-based identification of prototypical instances of constructions ». In : *Annual Review of Cognitive Linguistics* 1, p. 1–27 (cf. p. 11).

– (2009). *Quantitative Corpus Linguistics with R : A Practical Introduction*. New York : Routledge (cf. p. 5, 41, 56).

GRIES, Stefan Thomas (2013a). « 50-something years of work on collocations : what is or should be next... » In : *International Journal of Corpus Linguistics* 18.1 (cf. p. 5, 19, 73, 74, 107).

– (2013b). « Data in Construction Grammar ». In : *The Oxford Handbook of Construction Grammar*. Sous la dir. de Thomas HOFFMANN et Graeme TROUSDALE. Oxford ; New York : Oxford University Press, p. 93–108 (cf. p. 10, 18).

– (2013c). *Statistics for Linguistics with R : A Practical Introduction*. Mouton de Gruyter (cf. p. 5).

– (2014). « Frequency tables : Tests, effect sizes, and explorations ». In : *Corpus Methods for Semantics : Quantitative Studies in Polysemy and Synonymy*. Sous la dir. de Dylan GLYNN et Justyna ROBINSON. Amsterdam : John Benjamins, p. 365–389 (cf. p. 41, 46).

– (2015). « More (old and new) misunderstandings of collostructional analysis : On Schmid and Küchenhoff (2013) ». In : *Cognitive Linguistics* 26.3, p. 505–536 (cf. p. 73).

GRIES, Stefan Thomas et Dagmar DIVJAK (2010). « Quantitative approaches in usage-based Cognitive Semantics : Myths, erroneous assumptions, and a proposal ». In : *Corpus-Driven Cognitive Semantics. Quantitative Approaches*. Sous la dir. de Dylan GLYNN et Kerstin FISCHER. Berlin : Mouton de Gruyter, p. 333–353 (cf. p. 54).

GRIES, Stefan Thomas et Nick C. ELLIS (2015). « Statistical measures for usage-based linguistics ». In : *Language Learning* 65.S1, p. 228–255 (cf. p. 62, 97, 100).

GRIES, Stefan Thomas, Beate HAMPE et Doris SCHÖNEFELD (2005). « Converging evidence : Bringing together experimental and corpus data on the association of verbs and constructions ». In : *Cognitive Linguistics* 16.4, p. 635–676 (cf. p. 11).

GRIES, Stefan Thomas et Anatol STEFANOWITSCH (2004a). « Co-varying collexemes in the into-causative ». In : *Language, Culture, and Mind*. Sous la dir. de Michel ACHARD et Suzanne KEMMER. Stanford, Calif. : CSLI, p. 225–236 (cf. p. 68).

– (2004b). « Extending collostructional analysis : A corpus-based perspective on 'alternations' ». In : *International Journal of Corpus Linguistics* 9.1, p. 97–129 (cf. p. 68).

– (2006). *Corpora in Cognitive Linguistics : Corpus-based Approaches to Syntax and Lexis*. Berlin : Mouton de Gruyter (cf. p. 11).

HAUSER, Marc D., Noam CHOMSKY et W. Tecumseh FITCH (2002). « The faculty of language : What is it, who has it, and how did it evolve ? » In : *Science* 298, p. 1569–1579 (cf. p. 2).

HENGEVELD, Kees et Evelien KEIZER (2011). « Non-straightforward communication ». In : *Journal of Pragmatics* 43.7, p. 1962–1976 (cf. p. 25).

HILPERT, Martin (2006). « Distinctive collexeme analysis and diachrony ». In : *Corpus Linguistics and Linguistic Theory* 2.2, p. 243–256 (cf. p. 68).

– (2008). *Germanic Future Constructions : A Usage-Based Approach to Language Change*. Constructional approaches to language. Amsterdam ; Philadelphia : John Benjamins (cf. p. 37).

– (2014). *Construction Grammar and its Application to English*. Edinburgh textbooks on the English language Advanced. Edinburgh : Edinburgh University Press (cf. p. 15).

HOEY, Michael (2005). *Lexical Priming : A New Theory of Words and Language*. New York : Routledge (cf. p. 49).

HOFFMANN, Thomas et Graeme TROUSDALE (2013). *The Oxford handbook of construction grammar*. Oxford ; New York : Oxford University Press (cf. p. 15).

HOLLMANN, Willem B. et Anna SIEWIERSKA (2007). « A construction grammar account of possessive constructions in Lancashire dialect : some advantages and challenges. » In : *English Language and Linguistics* 11.2, p. 407–424 (cf. p. 54).

HOLMES, Janet (1984). « Modifying illocutionary force ». In : *Journal of Pragmatics* 8.3, p. 345–365 (cf. p. 25).

– (1995). *Women, Men, and Politeness*. Real Language Series. London ; New York : Longman (cf. p. 25).

HOUSE, Juliane et Gabriele CASPER (1981). « Interpersonal markers in English and German ». In : *Conversational Routines*. Sous la dir. de Florian COULMAS. The Hague : Mouton de Gruyter, p. 157–185 (cf. p. 25).

HUNSTON, Susan et Gill FRANCIS (2000). *Pattern Grammar : A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam : John Benjamins (cf. p. 107).

ITAKURA, Hiroko (2013). « Hedging praise in English and Japanese book reviews ». In : *Journal of Pragmatics* 45.1, p. 131–148 (cf. p. 25).

JACKENDOFF, Ray (1997). *The Architecture of the Language Faculty*. Cambridge, Mass. ; London : MIT Press (cf. p. 36).

JACKENDOFF, Ray et Steven PINKER (2005a). « The faculty of language : what's special about it ? » In : *Cognition* 95, p. 201–236 (cf. p. 2).

– (2005b). « The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky) ». In : *Cognition* 97, p. 211–225 (cf. p. 2).

JOHANSSON, Stig, Geoffrey LEECH et Helen GOODLUCK (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Department of English. University of Oslo (cf. p. 43).

KAY, Paul (1995). « Construction Grammar ». In : *Handbook of Pragmatics (Manual)*. Sous la dir. de Jef VERSCHUEREN, Jan-Ola ÖSTMAN, Jan BLOMMAERT et Chris BULCAEN. Amsterdam : John Benjamins, p. 171–177 (cf. p. 19, 37).

– (2013). « The Limits of (Construction) Grammar ». In : *The Oxford Handbook of Construction Grammar*. Sous la dir. de Thomas HOFFMANN et Graeme TROUSDALE. Oxford : Oxford University Press (cf. p. 19, 36, 38–40, 85, 87).

KAY, Paul et Charles FILLMORE (1999). « Grammatical constructions and linguistic generalizations : the *What's X doing Y ?* construction ». In : *Language* 75, p. 1–33 (cf. p. 19, 39).

KERBRAT-ORECCHIONI, Catherine (2001). *Les actes de langage dans le discours. Théorie et fonctionnement*. Paris : Nathan (cf. p. 25).

KILGARRIFF, Adam (2005). « Language is never, ever, ever, random ». In : *Corpus Linguistics and Linguistic Theory* 1.2, p. 263–276 (cf. p. 59, 63, 100, 118).

KLEIBER, Georges (1997). « Massif/comptable et partie/tout ». In : *Verbum* 3, p. 321–337 (cf. p. 33).

– (2014). « Massif/comptable : d'une problématique à l'autre ». In : *Langue Française* 183.3, p. 3–24 (cf. p. 34).

KÖNIG, Dénes (1950). *Theorie der Endlichen und Unendlichen Graphen*. New York : American Mathematical Society (cf. p. 100).

LAFON, Pierre (1980). « Sur la variabilité de la fréquence des formes dans un corpus ». In : *Mots* 1.1, p. 127–165 (cf. p. 67).

– (1981). « Analyse lexicométrique et recherche des cooccurrences ». In : *Mots* 3.1, p. 95–148 (cf. p. 67).

– (1984). *Dépouillements et statistiques en lexicométrie*. Travaux de linguistique quantitative. Genève, Paris : Slatkine, Champion, p. XII–217 (cf. p. 67).

LAKOFF, George (1973). « Hedges : A study in meaning criteria and the logic of fuzzy concepts ». In : *Journal of Philosophical Logic* 2.4, p. 458–508 (cf. p. 25).

– (1987). *Women, Fire, and Dangerous Things*. University Of Chicago Press (cf. p. 10, 18, 48, 95).

– (1990). « The Invariance Hypothesis : is abstract reason based on image-schemas ? » In : *Cognitive Linguistics* 1.1, p. 39–74 (cf. p. 18, 77).

LANDAUER, Thomas K. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, N.J. ; London : Lawrence Erlbaum Associates (cf. p. 114).

LANGACKER, Ronald W. (1987). *Foundations of Cognitive Grammar. Theoretical prerequisites*. T. 1. Stanford, CA : Stanford University Press (cf. p. 7, 9, 10, 16, 20, 21, 40, 47, 48, 60, 95–97, 107).

– (1988). « A usage-based model ». In : *Topics in Cognitive Linguistics*. Sous la dir. de Brygida RUDZKA-OSTYN. Amsterdam ; Philadelphia : John Benjamins, p. 127–161 (cf. p. 16, 48).

– (1990). *Concept, Image, and Symbol : The Cognitive Basis of Grammar*. Berlin ; New York : Mouton de Gruyter (cf. p. 10).

– (1998). « Conceptualization, symbolization, and grammar ». In : *The New Psychology of Language*. Sous la dir. de Michael TOMASELLO. Hillsdale, NJ : Erlbaum, p. 1–39 (cf. p. 2).

– (1999). *Grammar and Conceptualization*. Cognitive linguistics research. Berlin : Mouton de Gruyter (cf. p. 16, 29).

– (2005). « Construction Grammars : Cognitive, Radical and Less So ». In : *Cognitive Linguistics : Internal Dynamics and Interdisciplinary Interaction*. Sous la dir. de Francisco J. RUIZ DE MENDOZA IBÁÑEZ et M. Sandra PEÑA CERVEL. Berlin, New York : Mouton de Gruyter, p. 101–159 (cf. p. 21).

– (2008). *Cognitive Grammar : A Basic Introduction*. Oxford : Oxford University Press (cf. p. 21, 60).

– (2009a). « Cognitive (Construction) Grammar ». In : *Cognitive Linguistics* 20.1, p. 167–176 (cf. p. 19, 21).

– (2009b). « Constructions and Constructional Meaning ». In : *New Directions in Cognitive Linguistics*. Sous la dir. de Vyvyan EVANS et Stéphanie POURCEL. Amsterdam : John Benjamins, p. 225–267 (cf. p. 21).

LEBART, Ludovic et André SALEM (1994). *Statistique textuelle*. Paris : Dunod (cf. p. 67).

LEECH, Geoffrey N. (1983). *Principles of Pragmatics*. London : Longman (cf. p. 25, 27, 29).

LEGALLOIS, Dominique et Jacques FRANÇOIS (2006). « Autour des grammaires de constructions ». Cahier du CRISCO n^{O}21 (cf. p. 15).

LEVINSON, Stephen C. (1983). *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge ; New York : Cambridge University Press, p. 420 (cf. p. 26).

MANNING, Christopher D., Mihai SURDEANU, John BAUER et al. (2014). « The Stanford CoreNLP Natural Language Processing Toolkit ». In : *Association for Computational Linguistics (ACL) System Demonstrations*, p. 55–60 (cf. p. 107).

MICHAELIS, Laura A. (2004). « Type shifting in construction grammar : An integrated approach to aspectual coercion ». In : *Cognitive Linguistics* 15.1, p. 1–67 (cf. p. 33).

MIKOLOV, Tomas, Kai CHEN, Greg CORRADO et Jeffrey DEAN (2013). « Efficient Estimation of Word Representations in Vector Space ». In : *CoRR* abs/1301.3781 (cf. p. 115).

MIKOLOV, Tomas, Ilya SUTSKEVER, Kai CHEN, Greg CORRADO et Jeffrey DEAN (2013). « Distributed Representations of Words and Phrases and their Compositionality ». In : *CoRR* abs/1310.4546 (cf. p. 115).

MIKOLOV, Tomas, Wen-tau YIH et Geoffrey ZWEIG (2013). « Linguistic regularities in continuous space word representations ». In : *Proceedings of NAACL-HLT*, p. 746–751 (cf. p. 115).

MULLER, Charles (1964). *Essai de statistique lexicale. L'Illusion Comique de Pierre Corneille*. Paris : Klincksieck (cf. p. 67).

– (1973). *Initiation aux méthodes de la statistique linguistique*. Paris : Champion (cf. p. 67).

– (1977). *Principes et méthodes de statistique lexicale*. Paris : Hachette (cf. p. 67).

NICOLAS, David (2002). *La Distinction entre noms massifs et noms comptables : aspects linguistiques et conceptuels*. Louvain ; Dudley : Peeters (cf. p. 33).

NUNBERG, Geoffrey, Ivan A. SAG et Thomas WASOW (1994). « Idioms ». In : *Language* 70.3, p. 491–538 (cf. p. 18).

PALANCAR, Enrique (1999). « What do we give in Spanish when we hit ? A constructionist account of hitting expressions ». In : *Cognitive Linguistics* 10.1, p. 57–91 (cf. p. 10).

PAPAFRAGOU, Anna (2000). « Early communication : Beyond speech-act theory ». In : *Proceedings of the 24th Annual Boston University Conference on Language Development*. T. 2, p. 571–582 (cf. p. 27).

PARADIS, Carita (1997). *Degree Modifiers of Adjectives in Spoken British English*. Lund : Lund University Press (cf. p. 69, 78).

PECINA, Pavel (2010). « Lexical association measures and collocation extraction ». In : *Language Resources and Evaluation* 44.1, p. 137–158 (cf. p. 68).

PELLETIER, F. Jeffry (1975). « Non-singular reference : some preliminaries ». In : *Philosophia* 5.4, p. 451–465 (cf. p. 33).

PENNINGTON, Jeffrey, Richard SOCHER et Christopher D. MANNING (2014). « GloVe : Global Vectors for Word Representation ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, p. 1532–1543 (cf. p. 115, 116).

PIAO, Scott, Francesca BIANCHI, Carmen DAYRELL, Angela D'EGIDIO et Paul RAYSON (2015). « Development of the multilingual semantic annotation system ». In : *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Association for Computational Linguistics, p. 1268–1274 (cf. p. 114).

PINKER, Steven (1989). *Learnability and Cognition : the Acquisition of Argument Structure*. Cambridge, Mass. ; London : MIT Press (cf. p. 36).

POLLARD, Carl Jesse et Ivan A. SAG (1994). *Head-driven Phrase Structure Grammar*. Stanford, Chicago : Center for the Study of Language et Information ; University of Chicago Press (cf. p. 2, 19).

QUINE, Willard van Orman (1960). *Word and Object*. Studies in communication. Cambridge : Technology Press of the Massachusetts Institute of Technology, p. 294 (cf. p. 32).

R CORE TEAM (2016). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (cf. p. 4, 46, 56).

RENOUF, Antoinette (2015). « Rules of lexical creativity ». English Linguistics and Corpora (Engcorpora2015). Université Paris Est Créteil (cf. p. 45).

RONG, Xin (2014). « `word2vec` parameter learning explained ». In : *CoRR* abs/1411.2738 (cf. p. 115).

RYDER, Mary Ellen (1994). *Ordered Chaos : The Interpretation of English Noun-Noun Compounds*. Berkeley ; London : University of California Press (cf. p. 9).

SADOCK, Jerrold M. et Arnold M. ZWICKY (1985). « Speech Acts Distinctions in Syntax ». In : *Language Typology and Syntactic Description*. Sous la dir. de Timothy SHOPEN. Cambridge : Cambridge University Press, p. 155–196 (cf. p. 26).

SAG, Ivan A. (2008). « Sign-Based Construction Grammar : An Informal Synopsis ». In : *Sign-Based Construction Grammar*. Sous la dir. de C. BOAS Hans et Ivan A. SAG. Stanford, CA : CSLI Publications, p. 39–170 (cf. p. 22).

SANDRA, Dominiek (1998). « What linguists can and can't tell you about the human mind : A reply to Croft ». In : *Cognitive Linguistics* 9.4, p. 361–378 (cf. p. 8–10).

SANDRA, Dominiek et Sally RICE (1995). « Network analyses of prepositional meaning : Mirroring whose mind ?the linguist's or the language user's ? » In : *Cognitive Linguistics* 6, p. 89–130 (cf. p. 9).

SAUSSURE, Ferdinand de (1916). *Cours de linguistique générale*. Lausanne : Payot (cf. p. 16).

SBISÀ, Marina (2001). « Illocutionary force and degrees of strength in language use ». In : *Journal of Pragmatics* 33.12, p. 1791–1814 (cf. p. 25).

SCHMID, Hans-Jörg (2007). « Entrenchment, salience, and basic levels ». In : *The Oxford Handbook of Cognitive Linguistics*. Sous la dir. de Dirk GEERAERTS et Hubert CUYCKENS. Oxford : Oxford University Press, p. 117–138 (cf. p. 61).

– (2010). « Does frequency in text really instantiate entrenchment in the cognitive system ? » In : *Quantitative Methods in Cognitive Semantics : Corpus-Driven Approaches*. Sous la dir. de Dylan GLYNN et Kerstin FISCHER. Berlin, New York : Mouton de Gruyter, p. 101–133 (cf. p. 61).

SCHMID, Hans-Jörg et Helmut KÜCHENHOFF (2013). « Collostructional analysis and other ways of measuring lexicogrammatical attraction : Theoretical premises, practical problems and cognitive underpinnings ». In : *Cognitive Linguistics* 24.3, p. 531–577 (cf. p. 73).

SCHNEIDER, Stefan (2010). « Mitigation ». In : *Interpersonal Pragmatics*. Sous la dir. de Miriam A. LOCHER et Sage L. GRAHAM. Handbooks of Pragmatics. Berlin, New York : Mouton de Gruyter, p. 253–269 (cf. p. 25).

SINCLAIR, John (1966). « Beginning the study of lexis ». In : *In Memory of J.R. Firth*. Sous la dir. de C. E. BAZELL, J. C. CATFORD, M. A. K. HALLIDAY et R. H. ROBINS. London : Longman, p. 410–431 (cf. p. 47).

– (1987). « Collocation : a progress report ». In : *Language Topics : Essays in Honour of Michael Halliday*. Sous la dir. de Ross STEELE et TTerry THREADGOLD. T. 2. Amsterdam : John Benjamins, p. 319–331 (cf. p. 47).

– (1991). *Corpus, Concordance, Collocation*. Describing English language. Oxford : Oxford University Press (cf. p. 47).

SINCLAIR, John et Ronald CARTER (2004). *Trust the Text : Language, Corpus and Discourse*. London : Routledge (cf. p. 47).

SMITH, Nicholas, Sebastian HOFFMANN et Paul RAYSON (2008). « Corpus Tools and Methods, Today and Tomorrow : Incorporating Linguists ? Manual Annotations ». In : *Literary and Linguistic Computing* 23.2, p. 163–180. eprint : `http://llc.oxfordjournals.org/content/23/2/163.full.pdf+html` (cf. p. 113).

SPEELMAN, Dirk (2014). « Logistic regression : A confirmatory technique for comparisons in corpus linguistics ». In : *Corpus Methods for Semantics : Quantitative Studies in Polysemy and Synonymy*. Sous la dir. de Dylan GLYNN et Justyna ROBINSON. Amsterdam : John Benjamins, p. 487–533 (cf. p. 90).

STEELS, Luc (2004). *Fluid Construction Grammars : A brief tutorial*. published (cf. p. 22).

– (2011). *Design Patterns in Fluid Construction Grammar*. John Benjamins Pub. Co. (cf. p. 22).

– (2012). *Computational Issues in Fluid Construction Grammar*. Lecture Notes in Computer Science. Berlin : Springer (cf. p. 22).

STEFANOWITSCH, Anatol (2006). « Negative evidence and the raw frequency fallacy ». In : *Corpus Linguistics and Linguistic Theory* 2.1, p. 61–77 (cf. p. 52).

STEFANOWITSCH, Anatol et Stefan Thomas GRIES (2003). « Collostructions : Investigating the interaction of words and constructions ». In : *International Journal of Corpus Linguistics* 8.2, p. 209–243 (cf. p. 68).

– (2005). « Covarying collexemes ». In : *Corpus Linguistics and Linguistic Theory* 1.1, p. 1–46 (cf. p. 68).

STUBBS, Michael (2001). *Words and Phrases : Corpus Studies of Lexical Semantics*. Oxford : Blackwell (cf. p. 47).

TAGLIAMONTE, Sali et Rachel HUDSON (1999). « *Be like* et al. beyond America : The quotative system in British and Canadian youth ». In : *Journal of Sociolinguistics* 3.2, p. 147–172 (cf. p. 63).

TALMY, Leonard (2007). « Foreword ». In : *Methods in Cognitive Linguistics*. Sous la dir. de Monica GONZALEZ-MARQUEZ. Amsterdam ; Philadelphia : John Benjamins, p. xi–xxi (cf. p. 11, 12).

TAYLOR, John R. (2012). *The Mental Corpus : How Language is Represented in the Mind*. Oxford University Press (cf. p. 47).

TER MEULEN, Alice (1981). « An intensional logic for mass terms ». In : *Philosophical Studies* 40.1, p. 105–125 (cf. p. 32).

TEUBERT, Wolfgang (2005). « My version of corpus linguistics ». In : *International Journal of Corpus Linguistics* 10.1, p. 1–13 (cf. p. 47).

THALER, Verena (2012). « Mitigation as modification of illocutionary force ». In : *Journal of Pragmatics* 44.6–7, p. 907–919 (cf. p. 25).

THE SKETCH ENGINE (2012). *enTenTen12* (cf. p. 53).

TOGNINI-BONELLI, Elena (2001). *Corpus Linguistics at Work*. Amsterdam : John Benjamins (cf. p. 46).

TOMASELLO, Michael et Patricia BROOKS (1998). « Young children's earliest transitive and intransitive constructions ». In : *Cognitive Linguistics* 9.4, p. 379–395 (cf. p. 10).

TUGGY, David (2001). « Linguistic evidence for polysemy in the mind : a response to William Croft and Dominiek Sandra ». In : *Cognitive Linguistics* 10.4, p. 343–368 (cf. p. 8, 10).

TUMMERS, José, Kris HEYLEN et Dirk GEERAERTS (2005). « Usage-based approaches in Cognitive Linguistics : A technical state of the art ». In : *Corpus Linguistics and Linguistic Theory* 1.2, p. 225–261 (cf. p. 45).

VAN DER MAATEN, Laurens et Geoffrey HINTON (2008). « Visualizing Data using t-SNE ». In : *Journal of Machine Learning Research* 9, p. 2579–2605 (cf. p. 118).

WARD, Joe H (1963). « Hierarchical grouping to optimize an objective function ». In : *Journal of the American Statistical Association* 58.301, p. 236–244 (cf. p. 78).

WATTS, Duncan J. et Steven H. STROGATZ (1998). « Collective dynamics of 'small-world' networks ». In : *Nature* 393, p. 440–442 (cf. p. 100).

WATTS, Richard J. (2003). *Politeness*. Cambridge, UK ; New York : Cambridge University Press (cf. p. 3).

WEINREICH, Uriel (1966). « Explorations in Semantic Theory ». In : *Current Trends in Linguistic Theory*. Sous la dir. de Thomas Albert SEBEOK. T. 3. The Hague : Mouton, p. 395–477 (cf. p. 33).

WENGER, Etienne (1998). *Communities of Practice : Learning, Meaning, and Identity*. Cambridge ; New York : Cambridge University Press (cf. p. 38).

WIDDOWSON, Henry G. (2000). « On the limitations of linguistics applied ». In : *Applied Linguistics* 21.1, p. 3–25 (cf. p. 55).

WILLIAMS, Geoffrey (1998). « Collocational networks : interlocking patterns of lexis in a corpus of plant biology research articles ». In : *International Journal of Corpus Linguistics* 3.1, p. 151–171 (cf. p. 104, 105).

– (2001). « Mediating between lexis and texts : collocational networks in specialised corpora ». In : *ASp* 31–33 (cf. p. 104).

– (2002). « ?In search of representativity in specialised corpora : categorisation through collocation ». In : *International Journal of Corpus Linguistics* 7.1, p. 43–64 (cf. p. 104).

YANG, Yingli (2013). « Exploring linguistic and cultural variations in the use of hedges in English and Chinese scientific discourse ». In : *Journal of Pragmatics* 50.1, p. 23–36 (cf. p. 25).

ZIPF, George K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge : Addison-Wesley (cf. p. 53, 87, 100).