Dynamiques du changement sémantique. Détection, analyse et modélisation du changement sémantique en corpus en diachronie courte.

Armelle BOUSSIDAN

CNRS / L2C2 – Institut des Sciences Cognitives
armelle.boussidan@isc.cnrs.fr

Avant-propos

Le texte qui suit est un résumé en français de ma thèse publiée en anglais et intitulée « Dynamics of semantic change. Detecting, analyzing and modeling semantic change in corpus in short diachrony » (2013).

L'angle d'analyse choisi pour ce travail a été la création d'un pont entre les théories linguistiques du changement sémantique et les besoins du Traitement Automatique des Langues (TAL). Ce pont a déjà été partiellement établi au travers des contacts entre le TAL et la linguistique de corpus. Pour le français, le corpus *Frantext*, ainsi que le corpus composé d'archives du journal *Le Monde* font référence. De par leur disponibilité, ces ressources sont largement utilisées en recherche. Le genre d'un corpus détermine par avance la prédominance de certains thèmes, ici la politique et la société. Le corpus *Le Monde,* utilisé dans ce travail, a donné naissance à de nombreuses analyses en lexicométrie politique, souvent complétées par des résultats provenant d'autres journaux, de sites web, ou encore de blogs. Bien que l'analyse politique ne soit pas au centre de ma démarche, celle-ci transparaît dans les données, puisqu'elles sont essentiellement constituées d'articles à vocation politique et sociale. L'analyse de la variation lexicale fait ressortir des thèmes et des évènements à teneur majoritairement socio-politique. Elle montre également que la politisation d'un sujet a un impact sur le réseau sémantique qui y est associé (voir par l'exemple la politisation de l'écologie). Le travail que je présente dans ma thèse croise donc l'ensemble des travaux de lexicométrie, sémiométrie, logométrie ou textométrie en analyse politique, du fait des données utilisées, mais également du fait de l'approche sémantique choisie. La sémantique y est mise au service de l'interprétation des données et de l'analyse du discours, permettant de mettre en lumière des mécanismes du sens ainsi que des mécanismes sociaux et politiques présents dans les données. Cette mise en lumière peut être réalisée selon différents procédés : étiquetage en traits sémantiques, comparaisons entre corpus, représentation des contextes, analyses des fréquences, analyses des fréquences de co-occurrence, etc. L'étiquetage en traits sémantiques ou la création de bases de données lexicales de référence et de filtrage m'ont paru être des processus intéressants mais fastidieux impliquant souvent une part de subjectivité et un temps de mise à jour et d'adaptation aux données conséquents. L'alternative à ces méthodes paraît résider dans l'usage du co-texte comme point de départ à l'analyse, ce qui réduit la subjectivité impliquée dans le processus de catégorisation appliqué par le chercheur. Lorsque l'on utilise le co-texte, cette subjectivité se déplace sur un plan quantitatif : l'interprétation des données sera impactée par le choix de la taille du co-texte, ou sa « fenêtre », et par la nature de la composition de cette fenêtre. Néanmoins, ce choix reste paramétrable, et les données sources restent les mêmes, permettant de comparer des analyses faites avec la même fenêtre. Je suis donc en faveur d'un pré-traitement des données le plus minimal afin de conserver des données sources dont les analyses seront aisément comparables. L'étiquetage tel qu'il est traditionnellement opéré par le TAL ajoute un niveau de traitement catégoriel. Ces niveaux de catégories établis ne permettent pas de comparaison en dehors de leur propre cadre théorique et limitent les possibilités de comparaison avec des résultats obtenus par des variantes méthodologiques. L'annotation en traits sémantiques peut néanmoins fournir des analyses complexes de la néologie sémantique (voir la thèse de C. Reteunauer 2012 dans une perspective d'acquisition semi-automatique). Ces questions méthodologiques expliquent pourquoi l'approche que j'ai privilégiée diverge des méthodes les plus couramment utilisées en TAL et en lexicométrie. En effet, les limites rencontrées appellent à des perspectives exploratoires, à des tâtonnements vers de nouvelles méthodes. Pour que les outils de

traitement du langage et du sens de demain soient le plus en adéquation avec la complexité et la nature dynamique du langage, il est nécessaire que ces outils traitent le langage de la façon la plus directe qui soit, c'est-à-dire en construisant le moins de métalangages possibles. Il est donc intéressant de faire appel à des représentations paramétrables qui vont puiser leur structure dans le corpus directement. La construction d'une représentation reste néanmoins la question centrale de la sémantique, tel un pont entre l'expression et le réel. Le choix d'un modèle de représentation est à la fois une richesse supplémentaire et une limite en sémantique. En choisissant de travailler dans le paradigme des Atlas Sémantiques, je me suis appuyée sur des représentations paramétrables et adaptables, non-totales, l'idée étant de construire une boîte à outil permettant de cartographier le sens plutôt que de le disséquer en catégories. Le recours à un ensemble de paramètres et d'indices est alors lié à une perspective d'observation plus qu'à une perspective d'évaluation. La semi-automatisation des procédés fournit ainsi un matériau d'étude au spécialiste, déplaçant l'intervention de la subjectivité au niveau de l'analyse plutôt qu'au niveau de la structuration des données. Il s'agit donc de créer un prototype modulable et adaptable, en évitant de le structurer à partir d'un jugement de valeur préalable sur la nature du sens.

Abstract : This doctoral thesis aims at elucidating the processes of semantic change in short diachrony (or micro-diachrony) in corpus. To understand, analyze and model the dynamics of these changes and lay the groundwork for dynamic language processing, the corpus is divided in a series of time periods of one month. This work uses H. Ji's ACOM model, which is an extension of the Semantic Atlas, both of which are geometrical models of meaning representation based on correspondence factor analysis and the notion of *cliques*. Language and meaning statistical processing issues as well as modeling and representation issues are dealt with in conjunction with linguistic, psychological and sociological aspects from a holistic multidisciplinary perspective, as conceived by cognitive sciences. An approach of detection and analysis of semantic change is proposed along with case studies which deal both with large scale and precise detailed phenomena, therefore offering several levels of granularity. On the one hand, semantic change is dealt with as the deployment of polysemy in time, and on the other hand as a consequence of communication methods related to the media and the diffusion of such methods. Linguistics, sociology and information sciences all contribute to the study of the making of new meanings and new words. The analysis of the semantic networks of the studied items show the constant reorganization of meanings in time, and captures a few fundamental aspects of this process. The case studies focus primarily on the French term *malbouffe* ("junk food"), and on the semantic change of the element of composition *bio-*, as well as on the connotational drift of the French term *mondialisation* compared to its near-synonym *globalisation* ("globalization"). A prototype has been developed for these case studies as well as future studies.

Key words: Semantic change, neology, diachrony, corpus linguistics, computational semantics, semantic atlas, press, information science, globalization.

Résumé : Cette thèse vise à élucider les mécanismes du changement sémantique en diachronie courte (ou micro-diachronie) dans des corpus. Pour comprendre, analyser et modéliser la dynamique de ces changements et poser les jalons d'un traitement dynamique du langage, le corpus est segmenté en une série de périodes temporelles d'un mois. Ce travail utilise le modèle ACOM (de H. Ji), qui s'inscrit dans le paradigme des Atlas Sémantiques, un modèle géométrique de représentation du sens basé sur l'analyse factorielle des correspondances et la notion de *cliques*. Les questions de traitement statistique du langage et du sens, de modélisation et de représentation sont traitées conjointement aux questions d'ordre linguistique, psychologique, et sociologique, dans la perspective d'une analyse multidisciplinaire unifiée, telle que conçue par les sciences cognitives. Une démarche de détection et d'analyse du changement sémantique est proposée, accompagnée d'études de cas qui portent à la fois sur de la détection large et sur des détails précis, proposant différent niveaux de granularité. Le changement sémantique est traité comme un déploiement de la polysémie d'une part, et comme une conséquence des modes de communication liés aux médias actuels et à la diffusion de ceux-ci. Linguistique, sociologie et sciences de l'information se rencontrent dans l'étude de la fabrique de sens nouveaux et de mots nouveaux. L'analyse des réseaux sémantiques des termes étudiés montre la réorganisation constante des sens dans le temps et en capture quelques aspects fondamentaux. Les études de cas portent notamment sur le terme « malbouffe », sur le changement sémantique de l'élément de composition « bio- » et sur le glissement de sens observé pour le terme « mondialisation » par rapport à son quasi-synonyme « globalisation ». Un prototype informatique a été développé pour permettre le traitement de ces études et d'études futures.

Mots clefs : Changement sémantique, néologie, diachronie, linguistique de corpus, sémantique computationnelle, Atlas Sémantiques, presse, mondialisation, malbouffe, bio.

Le changement sémantique est un sujet vaste qui bénéficie de l'apport théorique de nombreuses disciplines en linguistique. C'est un phénomène qui s'ancre également dans la sociologie en termes d'usage et de diffusion de l'usage. L'usage est éclairé par la pragmatique alors que la diffusion est éclairée par les nouveaux paradigmes de la société de l'information, et donc par le rapport aux technologies. L'acceptation de la nouveauté et du changement sont eux ancrés dans un paradigme psychologique. Ainsi, le changement sémantique est, par nature, au croisement de plusieurs disciplines. De plus, pour évaluer la diffusion et l'usage des termes, le recours à des données réelles est indispensable. Dans ce cadre, il est nécessaire de traiter de vastes ensembles de données, et de caractériser ces données, ici la presse écrite, pour obtenir une forme d'objectivité statistique. Dès lors, des questions statistiques, mathématiques, ainsi que des questions de représentation et de caractérisation des données s'ajoutent aux disciplines précitées. Pour réduire le champ d'analyse de ce vaste ensemble, ce travail traite d'exemples au niveau lexical uniquement, et se concentre sur les mots pleins, laissant de côté les mots de fonction et les verbes qui demandent un traitement différent. L'étude de la diachronie des mots de fonction bénéficie par ailleurs de nombreuses contributions dans le champ de la grammaticalisation.

Les questions principales que pose le changement sémantique sont celles traitées par la sémasiologie et l'onomasiologie. La sémasiologie se demande comment les mots changent de sens et inversement l'onomasiologie se demande comment les sens et les concepts se réorganisent dans le temps, en termes de distribution au travers des unités lexicales et des mots. En s'intéressant au changement sémantique, on s'intéresse à la vie des mots et des concepts et à leur capacité à se modifier dans le temps. En effet, le langage n'est pas un objet figé, mais « vivant », comme en attestent les tentatives d'analyse du langage en tant qu'organisme biologique au XIXème siècle (voir Joseph et Janda 2005). Si cette idée a perdu en crédibilité, le langage n'est pas non plus, par opposition, un objet statique et définissable hors-temps. Ce statut paradoxal est exprimé par Keller (1994) qui qualifie le langage d'objet du troisième type, c'est-à-dire qu'il n'est ni complètement un artefact humain, ni un phénomène naturel, mais qu'il est un peu des deux à la fois. Le langage est par essence dynamique, et c'est cette dynamique à laquelle s'intéresse ce travail, en cherchant à en comprendre les mécanismes fondamentaux au niveau lexical. Il s'agit donc d'extraire ces mécanismes, de les analyser, de les comprendre, pour pouvoir finalement les modéliser. La modélisation des mécanismes de dynamique du langage permettra, à long terme, de se reposer sur des modèles du langage qui respectent sa nature. En effet, l'on observe aujourd'hui les limites posées par les représentations statiques du langage en TAL, en intelligence artificielle et en sciences de l'information. Pour dépasser les pratiques de mise à jour manuelle des outils qui traitent le langage, il est nécessaire de comprendre les mécanismes de dynamique du sens, pour envisager la création d'outils plus proches de la nature plastique du langage, dans un futur proche.

Pour cela, de nombreuses disciplines ont proposé des outils et des approches théoriques, dont j'ai tenté de faire la synthèse, sans faire de choix d'école préalable. Les appellations « changement sémantique » ou « glissement sémantique » ou encore « déplacement sémantique » proviennent généralement de la linguistique anglo-saxonne. En Europe, des questions du même ordre se posent sous la dénomination « néologie sémantique » ou « néologie de sens ». Ces différences terminologiques peuvent partiellement expliquer pourquoi des disciplines qui posent des questions similaires n'ont pas encore formé un champ disciplinaire à part entière. En effet, des éléments d'analyse du changement sémantique sont fournis par la linguistique historique et la diachronie, la lexicologie, la lexicographie et la terminologie, ainsi que les versants quantitatifs de ces disciplines. S'y intéressent également la sémantique structurale et interprétative, la pragmatique et l'analyse du discours ainsi que la sémantique et la linguistique cognitives, dont la théorie des prototypes. Cette liste pourrait être complétée de nombreuses sous-disciplines, dont l'un des aboutissements récents est la formation d'une sémantique computationnelle diachronique historique et cognitive.

La néologie possède son propre cadre théorique, de la même manière que le « changement sémantique » s'inscrit dans un cadre théorique qui englobe changement linguistique et linguistique historique. La néologie sémantique est alors une forme de néologie, ni formelle (nouveau mot), ni grammaticale (un mot qui change de catégorie) mais sémantique, c'est-à-dire lorsqu'un mot change de sens. Ce changement ne se fait pas seul, car les concepts en jeu se redistribuent au travers de réseaux de mots et de réseaux d'associations. C'est pourquoi

l'approche expérimentale choisie se concentre majoritairement sur l'analyse de réseaux de mots et de sens au travers de leurs contextes d'emploi, de sorte à respecter la dimension de l'usage.

La question du changement sémantique a été posée depuis des siècles en termes de « nature » et de « causes », séparant ainsi les phénomènes internes (linguistiques) et externes (sociologiques et psychologiques) au langage. Cela a donné lieu à de nombreuses classifications typologiques et taxonomiques de leurs causes (politiques, sociales, psychologiques…), de leur nature morphologique et de leur nature rhétorique et stylistique (voir entre autres Bloomfield 1933 ; Ullmann 1953 ; Sablayrolles 1996), cette dernière catégorie étant la plus documentée.

Le point de vue adopté ici est plutôt celui d'une observation des mécanismes en jeu, non pas en vue de les classer mais de décrypter leurs interactions aux niveaux macroscopiques et microscopiques. Quelques approches cherchent à établir des ponts entre facteurs internes et externes, notamment en linguistique cognitive et en pragmatique historique, néanmoins cet aspect fondamental de l'articulation des sens est resté relativement inexploré. Cela peut être dû aux clivages entre disciplines qui perçoivent le langage soit comme un système indépendant (biologique, structural, mathématique), soit comme le produit d'interactions sociales entre individus (en psychologie, psycholinguistique et pragmatique, sociologie, sociolinguistique et socio-pragmatique). Ce clivage repose en partie sur la distinction entre *langue* et *parole* établie par la linguistique de Saussure. Néanmoins, l'étude du changement se trouve limité par cette distinction, puisque le changement s'exprime dans ces deux niveaux et dans l'interaction entre ces niveaux, qui sont avant tout des entités théoriques. De la même manière, la distinction entre synchronie et diachronie peut être remise en question, puisque ces concepts théoriques ne sont pas applicables dans un cadre expérimental. En effet, il n'y a pas d'unité pour définir la synchronie, qui n'est qu'un état statique théorique de la langue. Saussure affirme que les changements n'affectent que la langue en diachronie mais non en synchronie, ce qui va à l'encontre de sa propre théorie selon laquelle le système se réorganise dans son ensemble à chaque fois qu'un de ses éléments est modifié, puisque chaque élément n'existe que par opposition à un autre élément du système. Cette contradiction est notée dès les années 1950 par Coseriu (1958).

L'idée de mise à jour d'un inventaire des causes des changements sémantiques mérite donc d'être remise en question. Néanmoins, ces classifications sont utiles car elles permettent de décoder en détail les phénomènes observés.

Les facteurs internes du changement sémantique sont soit de nature morphologique soit de nature rhétorique. La morphologie permet d'établir les combinaisons possibles à partir du matériau de la langue, les deux mécanismes principaux étant la dérivation et la composition. Par exemple, le mot *anticapitaliste* est créé par dérivation en ajoutant le préfixe *anti-* à *capitaliste,* et l'anglais *blackboard* est une combinaison de *black* (« noir ») et *board* (« tableau »). Les règles de la morphologie prennent également en compte des facteurs de facilitation ou de limitation. Corbin (1987) parle de « disponibilité » et de « rentabilité » de l'affixe, c'est-à-dire sa capacité à se combiner avec le plus grand nombre de formes. Les combinaisons sont limitées par des facteurs grammaticaux, orthographiques, phonologiques (évitement de la cacophonie), et sémantiques (évitement de la création de synonymes, homonymes, homophones et homographes). Ces règles théoriques ne sont pas toujours respectées, du fait de contingences historiques ou du fait de la créativité. Par exemple, l'anglais compte un grand nombre de quasi-synonymes dont l'un est d'origine germanique et l'autre française, ce qui est dû à l'intégration de nombreux termes français dans le vocabulaire anglais suite à la conquête normande (par exemple *birthday*/*anniversary*). Dans ce cas, l'emprunt n'a pas été bloqué par la synonymie.

Les figures rhétoriques principales qui agissent au sein des procédés du changement sémantique sont la métaphore, la métonymie (ainsi que la synecdoque qui est un cas particulier de la métonymie), les procédés de généralisation et de spécialisation, ainsi que les procédés de mélioration et de péjoration. Cette liste est complétée par de nombreux phénomènes qui s'y rapportent. Voici quelques exemples phares pour chaque catégorie :

Généralisation : le terme anglais *dog* faisait référence aux chiens de chasse uniquement et s'est étendu au travers des siècles pour désigner tous les chiens.

Spécialisation : à l'inverse le terme anglais *deer* faisait référence à tous les animaux sauvages et s'est spécialisé au travers des siècles pour ne faire référence qu'au cerf.

Métaphore : le terme *tête* en français tout comme *head* en anglais, se réfèrent de façon métaphorique à ce qui est en position supérieure, par exemple dans l'expression « *être à la tête d'un groupe* » ou « *head of a company* » (« directeur d'une entreprise »).

Métonymie : si la métaphore se base sur la similarité, la métonymie se base sur la contigüité, comme par exemple lorsque l'on dit « Paris a décidé de … » pour faire référence au gouvernement français.

Mélioration/péjoration : les termes peuvent acquérir une polarité plus ou moins positive ou négative et celle-ci peut changer lorsque le concept est associé à un évènement ou à un changement de domaine (par exemple le terme *niais* : « bébé faucon » → « personne sotte », ici le sens « qui ne connaît rien » subit une péjoration). Ces processus sont parfois extrêmes avec des renversements complets, comme le terme anglais *nice* qui voulait tout d'abord dire « ignorant » avant de prendre une connotation positive de « sympathique, joli ». Ces changements de polarité font l'objet d'études ciblées, par exemple celle de Cook and Stevenson (2010) qui cherchent à les détecter automatiquement en corpus.

Les emprunts entre langues et entre communautés scientifiques méritent d'être traités séparément également, car ils sont au cœur de dimensions historiques, culturelles et sociologiques complexes. Par exemple, l'intégration des anglicismes dans la langue française fait l'objet d'une opposition et d'un contrôle politique (voir la loi dite « Toubon » de 1994 ainsi que la loi relative à l'enrichissement de la langue française de 1999[1]). Ainsi, pour chaque anglicisme, les autorités françaises s'efforcent de proposer des équivalents, qui ne sont que rarement acceptés dans l'usage, à l'instar d'un terme comme *baladodiffusion* proposé pour remplacer le terme *podcast*, déjà en vogue parmi les locuteurs du français.

Tous ces facteurs dits « linguistiques » ou « internes » font l'objet de nombreuses études et de nombreux désaccords. Ils s'opposent traditionnellement aux facteurs dits « externes ». Les facteurs externes, aussi appelés causes ou motivations, sont de nature plus abstraites. Ils incluent des processus émotionnels, psychologiques, historiques, sociologiques et anthropologiques. Ces facteurs sont beaucoup moins bien documentés, car bien plus délicats à aborder, du fait des dimensions contextuelles en jeu (culture, époque, cadre politique et social, etc.). Les mots et les sens sont alors étudiés comme des objets historiques, comme le recommande Bréal (1899), comme des entités sociales, comme les perçoit Meillet (1906) ou comme des véhicules émotionnels, comme les aborde Stern (1931). Ils sont parfois des outils de propagande politique, comme les envisage Orwell (1949) dans sa fiction « 1984 », ou Klemperer (1975) dans son analyse philologique de l'allemand du troisième Reich. Dans ces contextes, l'évolution de la connotation ainsi que les changements de domaine sont des procédés centraux. Par exemple, Hughes (1992) note que le terme anglais *noble* acquiert avec le temps une dimension morale, alors qu'à l'origine il n'était lié qu'à la classe sociale (cette observation tient aussi pour l'équivalent français).

L'idée de « réanalyse » permet de faire un pont entre les dimensions de la langue et de l'usage, et donc entre l'interne et l'externe théoriques. La réanalyse se produit lorsqu'un mot ou une expression sont employés de façon nouvelle et qu'ils sont compris dans ce sens. Cela peut se produire au niveau d'un seul emploi créatif, auquel cas il s'agit d'une idiosyncrasie (aussi appelée *néologisme subjectif*), ou au niveau d'une communauté de locuteurs (*néologisme objectif*), auquel cas il peut s'agir d'un emploi terminologique spécialisé ou communautaire (groupe défini par la géographie, l'appartenance ethnique ou sociale, groupe professionnel, groupe d'intérêt, etc.). Lorsqu'un nouvel emploi se diffuse dans une communauté, il devient alors candidat à une diffusion plus large, puis candidat à la lexicalisation, et à l'intégration lexicographique.

---

[1] Loi nº 94-665 du 4 août 1994, et le décret no 96-602 du 3 juillet 1996.

Cette approche donne une place prépondérante au locuteur, ainsi qu'à l'énonciation. Elle est en résonance avec la théorie de Traugott et Dasher (2002) dont l'argument central est la réanalyse du sens à chaque interaction. Les auteurs se basent sur la théorie des inférences développée en pragmatique. Pour ces derniers, le passage d'une inférence individuelle à une inférence généralisée est le mécanisme central du changement sémantique, au cœur de l'énonciation.

En onomasiologie diachronique et cognitive, cette dimension est prise en compte dans la classification des mécanismes du changement sémantique. Les travaux de Blank (1999 ; 2003) proposent ainsi une typologie complexe et tentent d'abroger la distinction entre facteurs internes et externes (ou encore nature/cause/conséquence) établie par Ullmann (1953 ; 1962) et reprise jusque-là par de nombreux auteurs. Une des idées centrales sur laquelle reposent les travaux de Blank est le transfert de sens. Ce transfert peut se faire au niveau paradigmatique (métaphore, étymologie populaire) ou syntagmatique (métonymie, ellipse). Cela entre en résonance avec les théories de la métaphore qui se basent sur le transfert (ou « mapping ») de sens d'un domaine à l'autre, comme décrit dans l'exemple connu donné par Lakoff and Johnson (1980) : ARGUMENT is WAR. Cet exemple montre que nous utilisons le vocabulaire de la guerre pour décrire un débat ou une dispute, par exemple « défendre sa position ». Dans ce cas, nous opérons un transfert entre un domaine source (la guerre) et un domaine cible (le débat). Ces mécanismes donnent lieu à des changements sémantiques au travers de l'usage.

Enfin, la théorie des prototypes propose une analyse du changement sémantique qui donne un rôle central à l'expressivité (c'est-à-dire la créativité) ainsi qu'à l'efficacité communicationnelle. Cette efficacité rejoint l'idée que les sens sont groupés selon un principe d'efficacité et d'économie, que postule la théorie des prototypes. Ces principes s'appliquent au niveau formel et au niveau conceptuel. L'efficacité formelle est liée à la stratégie pragmatique et l'expressivité formelle est liée à la productivité morphologique ainsi qu'à des facteurs idéologiques et sociaux. L'expressivité conceptuelle est influencée par des facteurs sociaux et par la dimension individuelle. Enfin, l'efficacité conceptuelle est ancrée dans les mécanismes de métaphore et de métonymie au sein des catégories prototypiques. Cette théorie, comme le souligne Rastier (1999), n'explique pas comment de nouveaux prototypes voient le jour, car tous les changements semblent être décrits à l'intérieur d'un système fermé. Cette critique rejoint toutes les critiques adressées aux défenseurs de théories limitées à des explications intra-linguistiques. En effet, l'idée que tous les changements se déploient dans un système fermé ne prend pas en compte la dimension de l'usage, dimension qui ne peut être ignorée.

Suite à ce survol des théories applicables au changement sémantique, laquelle faut-il retenir ? Il semble que la classification des phénomènes du langage rencontre un problème de circularité, car le langage est lui-même par essence une classification du réel. Ainsi, le linguiste crée un langage pour décrire le langage, ou métalangage, et ce métalangage s'oppose à d'autres métalangages, construits sur d'autres vues théoriques. J'ai donc choisi de ne pas m'inscrire dans la création d'un métalangage, mais au contraire de m'en extraire, tout en sachant qu'une approche totalement dénuée de théorie est impossible. Néanmoins, j'ai fait le choix d'observer et de décrire les phénomènes sans leur imposer de grille de lecture préalable, mais en faisant appel à des grilles si nécessaire. Parmi les grilles utiles, on peut retenir la décomposition du phénomène en trois étapes décrite par Lüdtke (1999 : 50) :

étape 1 : OUTSET ("début") [innovation/créativité]

étape 2 : INTERMEDIATE ("intermédaire") [diffusion/imitation]

étape 3 : OUTCOME ("résultat") [résultat/différence]

Ce cadre permet de positionner les données dont nous disposons, ici des archives de presse. La presse constitue un lieu relativement avancé de diffusion puisque ses codes d'écriture doivent toucher la population dans son ensemble. C'est donc un lieu de diffusion (étape 2) plus qu'un lieu de créativité, ce qui garantit une certaine homogénéité des données. Néanmoins, et fort heureusement, il reste une petite marge de créativité subjective dans le style journalistique.

Le principe phare au cœur de la créativité (linguistique, subjective, personnelle et collective) et de la productivité (morphologique et sémantique) est celui de la polysémie. En effet, il est rare qu'un changement sémantique se base sur un sens complètement nouveau, auquel cas une néologie formelle se crée généralement. Le changement sémantique s'ancre sur une multiplicité de sens déjà présents, mais dont les relations structurelles changent. Elles changent en termes de réseaux d'associations, de connexions entre ces réseaux, et de degrés de prépondérance (« *salience* ») des réseaux les uns par rapport aux autres. On considère donc qu'un changement sémantique est, par définition, une réorganisation structurale de la polysémie dans le temps. C'est l'analyse de cette réorganisation structurale qui retiendra mon attention.

Toutes les distinctions précitées (langue/parole, synchronie/diachronie, facteurs internes/externes) sont mises à l'épreuve par un changement de paradigme conséquent qui modifie l'échelle et la nature de la créativité, du contact, de la diffusion, et jusqu'à la nature même de la langue. Ce changement de paradigme est intimement lié au développement d'internet et des technologies mobiles dans nos sociétés, au raccourcissement extrême des délais nécessaires à la communication et à l'abolition virtuelle des distances. L'accès à ces technologies se démocratise dans les années 2000 dans les pays industrialisés. Cette transition est souvent résumée par l'appellation « société de l'information » par opposition à la société industrielle qui l'a précédée.

Ce changement de contexte touche la langue mais également les outils à notre disposition pour l'analyser. Ces dernières décennies, le développement de puissants outils de calcul a permis de traiter des ensembles de données de plus en plus conséquents et ceci de plus en plus rapidement. Dans ce contexte, la linguistique de corpus a bénéficié d'un regain d'intérêt au contact du TAL. Les données réelles sont ainsi redevenues centrales à la linguistique, la rapprochant de fait de la sociologie. Le nouvel essor de la linguistique de corpus est couplé à l'utilisation de statistiques. Ainsi, l'un des indicateurs les plus simples pour étudier un corpus est la fréquence d'apparition des mots. La fréquence ne se suffit pas à elle-même mais cet outil, aussi simple soit-il, permet d'explorer de nombreux aspects du langage. Par exemple, Pagel, Atkinson, et Meade (2007) montrent que les mots à forte fréquence évoluent plus lentement que les mots à basse fréquence dans l'ensemble des langues indo-européennes en diachronie longue. Pour ces auteurs, la fréquence est un indice du taux de remplacement des mots dans le temps. Au-delà de la simple fréquence du mot, la mise en relation avec le contexte linguistique (aussi appelé co-texte) ouvre des perspectives de traitement sémantique. Néanmoins, les concepts sont plus ou moins dépendants des contextes dans lesquels ils apparaissent, comme le souligne Barsalou (1982) et il convient donc de prendre des précautions dans l'analyse des chiffres obtenus avec de telles méthodes.

Ainsi, les techniques d'analyse statistique textuelle utilisées à des fins d'analyse littéraire et du média ont permis d'aborder la néologie et la sémantique avec de nouveaux outils. Par exemple, deux corpus peuvent être comparés, permettant de confronter un corpus de référence et un corpus test, ou des corpus de langues différentes. En utilisant des corpus dits d'exclusion, de nombreux travaux de terminologie ont détecté les néologies formelles dans la presse ou dans des textes spécialisés, pour le français (voir la base de données Bornéo par exemple[2]), l'espagnol (voir les travaux de l'observatoire de néologie de l'université Pompeu Fabra par exemple) et l'anglais entre autres (voir par exemple Renouf, 2007 pour la presse). Des logiciels de détection de la néologie formelle apparaissent également. En France et au Québec, les travaux de la sémiométrie (voir Lebart, Piron, et Steiner 2003), de la textométrie ou encore de la logométrie s'intéressent à des textes provenant du média, de la politique, et proposent des outils d'analyse. Le média lui-même s'en empare, publiant des analyses de son propre contenu, comme notamment les analyses proposées par Jean Véronis[3]. Il se crée donc un nouveau rapport entre la linguistique de corpus et le média, ainsi que l'entreprise, qui elle aussi s'est rapidement tournée vers les outils de statistique sémantique pour développer ses outils de communication et les évaluer. On trouve donc dans la presse des articles comparant les fréquences d'emploi de termes pour des candidats politiques ou des analyses semi-automatiques des champs sémantiques utilisés par des partis politiques. Cette nouvelle relation pousse les acteurs de la sphère médiatique et politique à s'adapter aux analyses qui sont faite de leur usage de la langue, ce qui modifie en soi le matériau étudié.

---

[2] http://www.atilf.fr/borneo/
[3] http://blog.veronis.fr/

Dans ce contexte, les contacts entre les communautés linguistiques spécialisées (groupes professionnels, de style, d'intérêt, etc.) et le langage formellement accepté s'accélèrent. Ces communautés démontrent un taux de créativité plus élevé, puisque l'identité verbale fait partie des outils à leur disposition pour asseoir leur identité sociale et culturelle. Elles forment des niches sociologiques (le concept de niche est emprunté à la biologie). Altmann, Pierrehumbert, et Motter (2011) affirment que la niche est plus déterminante que la fréquence en diachronie courte, en analysant deux corpus constitués à partir de forums en ligne, l'un s'intéressant à la musique et l'autre à Linux. Deux mesures sont employées : l'indexicalité, qui correspond à la dissémination d'un terme par utilisateur, et la topicalité, qui correspond à la dissémination d'un terme par entrée (ou fil d'actualité). Ainsi, le rapport entre la fréquence d'emploi générale et la fréquence d'emploi par utilisateur permet d'éclairer les processus de diffusion.

Le langage spécialisé (par exemple : mouvement musical) est généralement bien moins formel que le langage dit « général » ou alors, à l'inverse, il est extrêmement technique (par exemple : nouvelles technologies). Dans cette veine, Chesley (2011) analyse l'apprentissage du *slang* afro-américain au travers des paroles du hip-hop et trace l'intégration de termes en Américain Standard (par exemple, l'emploi de *bad*, « mauvais » pour exprimer l'opposé, *cool*).

Le remplacement des termes obsolètes se trouve également accéléré lorsqu'un domaine spécialisé devient d'intérêt général ou qu'il se politise. C'est le cas de l'écologie, qu'étudient notamment Dury et Drouin (2009).

Les outils utilisés par la terminologie spécialisée recoupent les besoins des approches variationnistes, et de telles études existent également pour évaluer la pénétration de termes provenant de dialectes ou de langues en contact, comme par exemple l'Espagnol et le Catalan, le Flamand de Belgique et le Néerlandais, ou encore le Français Québécois et le Français de France.

Parmi les nombreux outils développés, certains s'appuient sur des approches mixtes, incluant une dimension grammaticale (lemmatisation), syntaxique (analyseurs syntaxiques) ou actantielle (encodages sous forme ontologiques de la transitivité). Les comparaisons des données issues de telles études deviennent donc complexes. De ce fait, il me semble que l'usage de corpus sous forme de texte brut (sans annotation et sans recours à une ontologie) est une piste intéressante pour pouvoir comparer les résultats produits, sans qu'un débat sur une norme de traitement ne vienne limiter les disciplines impliquées.

Le traitement statistique des textes s'enrichit également d'outils puisés dans la théorie des graphes, avec une particularité française, celle de l'utilisation de l'AFC (Analyse Factorielle des Correspondances) faite par les sciences humaines, dans la lignée des travaux de Benzécri (1980) puis de Bourdieu (1979). Ailleurs, des méthodes similaires sont employées pour produire des représentations vectorielles (comme le SVD, « décomposition en valeurs singulières »[4] ou le MDS « positionnement multidimensionnel »). Avec ces outils, commence l'avènement des modèles de représentation vectoriels en linguistique. Il faut un temps avant que ces derniers ne soient appliqués en sémantique.

En modélisation, les travaux de Nerlich et Clarke (1988 ; 1999 ; Clarke et Nerlich 1991) posent la première pierre d'un modèle du changement sémantique réalisé de façon informatisée. L'idée que la prédominance d'un mot pour un concept donné est en constante compétition avec ses quasi-synonymes (par exemple *group* vs. *band* en anglais pour désigner un groupe de musique) est modélisée, et il en ressort un schéma de vagues. Les facteurs utilisés dans cette modélisation sont la fréquence, l'expressivité et l'accessibilité mémorielle.

Les possibilités de modélisation sémantique sont aujourd'hui bien plus développées, et le regain d'intérêt porté aux modèles vectoriels permet de poursuivre ces idées. Les modèles vectoriels utilisés en sémantique

[4] SVD : « singular value decomposition » ; MDS : « multi- dimensional scaling »

lexicale reposent sur l'hypothèse distributionnelle selon laquelle on peut accéder au sens d'un mot en observant les contextes d'emploi dans lesquels il apparaît (Firth 1957). Les modèles vectoriels attribuent des coordonnées aux mots d'un texte et il est possible de mesurer la similarité sémantique de ces mots en fonction des distances qui séparent les points. Ces méthodes sont très proches de l'AFC et permettent également de traiter des relations sémantiques (voir Utsumi 2010 par exemple), des relations grammaticales, syntaxiques et sémantiques (voir Padó et Lapata 2007 ; Mitchell et Lapata 2008) voire des relations entre plusieurs langues. LSA[5] (Landauer et al. 2007), LDA[6] (Blei, Ng, et Jordan 2003) et HAL[7](Lund et Burgess 1996), sont les modèles les plus utilisés ainsi que de nombreuses variantes de ces derniers. Ces modèles font concurrence aux approches ontologiques, dont la plus connue et la plus documentée est Wordnet. Wordnet dépend d'une structuration et d'une mise à jour manuelle. La version anglaise est assez fiable, alors que les versions dans d'autres langues demandent encore des travaux conséquents, car, à l'origine, ils n'étaient obtenus que par traduction de l'anglais. La question de la variation des structures ontologiques d'une langue à l'autre pose une limite que les modèles vectoriels ne rencontrent pas, s'ils n'ont pas recours à une ontologie pour former une représentation.

Les modèles vectoriels ont fleuri dans les années 1960 dans le sillon des sciences cognitives. L'enthousiasme provoqué par ces modèles a rapidement diminué en linguistique, laissant place au courant de la grammaire générative transformationnelle (voir les ouvrages de Chomsky, par exemple Chomsky 1965) et aux approches à dominante syntaxique en générale. Néanmoins ces modèles sont revenus sur le devant de la scène scientifique ces vingt dernières années, notamment via leur emploi en psycholinguistique, discipline qui cherche à valider des modèles de la cognition, de l'organisation des connaissances et de l'apprentissage. Ces modèles servent également les sciences de l'information et le TAL.

Dans le contexte de l'étude semi-automatisée de la sémantique avec des modèles vectoriels, le principe de co-occurrence déjà utilisé en statistique textuelle est la source principale de structuration des données. Il y a différentes façons de traiter la co-occurrence.

La fenêtre de co-occurrence correspond à l'unité de découpage choisie : on peut prendre pour fenêtre la phrase, le paragraphe, le document, la tranche temporelle, ou choisir de sélectionner *n* mots avant et après le mot cible, comme cela est le cas dans les corpus dits n-grams, bi-grams, tri-grams, etc. Les modèles qui se basent sur la relation terme cible – document sont appelés « topic models » (ou « modèle de sujet »). La co-occurrence peut également être étendue (pour obtenir une co-occurrence dite de second degré), pour sortir de la contrainte syntagmatique et inclure des données paradigmatiques. Pour cela, on analyse la co-occurrence des mots co-occurrents du mot cible pour élargir le réseau.

A partir de ces données, on construit une matrice, qui possède des entrées terme-document, ou terme-contexte, selon l'approche choisie. Cette matrice permet de construire un espace multidimensionnel en attribuant des coordonnées vectorielles aux mots ou aux ensembles de mots.

Quelques études ont appliqué ces méthodes au changement sémantique. Les développements dans ce domaine sont très récents, ainsi il se peut que de très récentes contributions aient échappé à ce recensement. Les travaux de Cook et Stevenson (2010) mentionnés plus haut, mesurent la polarité (positive ou négative) des mots en diachronie longue dans des corpus pour extraire des mouvements de péjoration et de mélioration. Sagi, Kaufmann, et Clark (2009) se basent sur Infomap, un dérivé de LSA, pour mesurer la variabilité et la densité des mots en diachronie longue. Ils s'appuient sur des cas démontrés par les linguistes préalablement et génèrent des topographies en MDS qui permettent de visualiser les processus de spécialisation et de généralisation. La densité est calculée en fonction de l'angle moyen entre les différents vecteurs pris dans le temps pour le même mot.

---

[5] Latent Semantic Analysis : http://lsa.colorado.edu/
[6] Latent Dirichlet Allocation
[7] Hyperspace Analog to Language

Heyer, Holz, et Teresniak (2009) et Holz et Teresniak (2010) s'appuient sur les changements de sujets dans la presse récente (le *New York Times* pour l'anglais et le *Wortschatz* pour l'allemand) et proposent une mesure de volatilité, inspirée de l'économétrie. Cette mesure s'appuie sur la moyenne du coefficient de variation du rang des mots co-occurrents (leur position par ordre d'importance). Avec cette mesure, les auteurs affirment éviter les biais liés à la fréquence. Ils proposent également une interface visuelle sous forme de graphique permettant d'afficher les courbes de fréquence et de volatilité pour les données étudiées.

Rohrdantz et al. (2011) proposent également un outil de visualisation. Les auteurs se servent d'un modèle LDA pour analyser le corpus presse du *New York Times*, et s'intéressent en particulier au vocabulaire de la technologie. Les données sont libellées en termes de catégorie contextuelle ainsi qu'avec un marqueur temporel. Les auteurs montrent, par exemple, comment l'emploi de *surf* a remplacé celui de *browse* en anglais, de la même manière qu'en français l'expression « surfer sur le web » a remplacé celle de « naviguer ». Ce changement est lié notamment à la disparition du Netscape Navigator, extrêmement répandu dans les années 1990 et dont le nom a marqué l'histoire du vocabulaire lié à Internet. Nous avons traité le même exemple (sans nous concerter) en français et trouvé de hautes fréquences dans le contexte technologique entre 1997 et 2001 dans le corpus *Le Monde*, et un déclin total par la suite. Nous notons par ailleurs que ce sens n'était toujours pas inclus dans le dictionnaire de référence du TLFI[8] (voir Boussidan et Ploux 2011).

Gulordava et Baroni (2011) ont appliqué des méthodes similaires à un corpus web, composé de livres au format électronique. Ils comparent deux périodes, les années 1960 et les années 1990. Ils utilisent une mesure de similarité sémantique pour comparer les termes de ces deux périodes. Les résultats sont évalués par des humains.

Les travaux que j'ai menés avec le modèle ACOM (« Automatic Contexonym Organizing Model ») sont assez proches de ces méthodes. Le modèle ACOM a été développé par H. Ji (2005) sur la base du modèle des Atlas Sémantiques développé par S. Ploux et son équipe.

Le modèle des Atlas Sémantiques (AS) est originellement construit sur un corpus de synonymes et génère des cartes du sens basées sur l'AFC (Benzécri, 1980). Un algorithme de hiérarchie permet une division en clusters thématiques. L'ACOM est dérivé des AS et traite des relations de co-occurrence entre les mots (« contexonymes ») à partir de n'importe quel corpus, dans le même paradigme géométrique. Ces modèles produisent des cartes sémantiques multidimensionnelles qui sont navigables de façon interactive et peuvent être consultés librement à l'adresse http://dico.isc.cnrs.fr/fr/index.html.

L'originalité du modèle initial repose sur l'unité de clique, unité mathématique et infra-linguistique. Les cliques sont des sous-unités de sens composées par des listes de termes à l'intérieur desquelles tous les mots sont liés par une relation : celle de synonymie dans les AS et celle de co-occurrence dans l'ACOM. Les cliques sont organisées dans un paradigme continu. Le passage d'une clique à la suivante s'effectue par un terme commun ; ce terme peut être considéré comme un pivot sémantique autour duquel s'articule le passage d'une valeur sémantique à une autre. Cette notion est fondamentale en diachronie car je cherche à observer ces pivots non plus comme des points d'articulation figés mais comme de potentiels points d'articulation actifs dans le temps.

On construit une matrice dont les lignes sont tous les mots cibles et les colonnes les cliques y correspondant. A partir de cette matrice, le modèle génère une visualisation multidimensionnelle en appliquant l'AFC. Cette visualisation permet trois niveaux de navigation conceptuelle : le niveau des cliques (représentées par des points), le niveau des mots (représentés par des enveloppes) et le niveau des ensembles conceptuels ou clusters thématiques (représentés par de plus larges enveloppes). L'intérêt de cette méthode est de construire l'espace en partant des unités lexicales et donc de la polysémie et non en partant du mot. Les clusters sont ensuite obtenus à partir d'un algorithme de classification hiérarchique. Ces clusters peuvent être composés à partir des cliques ou à partir du centre de gravité des enveloppes. Le nombre de clusters est paramétrable.

---

[8] Trésor de la Langue Française Informatisé

Voici un exemple de carte sémantique obtenue avec l'AS, pour représenter la polysémie du mot anglais *bright* (voir Figure 1). Ce terme exprime à la fois la luminosité et l'intelligence, et possède une connotation positive liée au bonheur. La carte montre la séparation de ces valeurs sémantiques en quatre clusters : en jaune un cluster lié à la clarté, en rouge un cluster lié à l'intelligence, en bleu un cluster lié au bonheur et enfin en vert un cluster lié à la luminosité. La transition entre clusters s'opère dans un paradigme continu, par exemple pour passer de *happy* (« heureux ») à *luminous* (« lumineux »), la transition s'opère au travers de concepts comme *beaming* (« rayonnant »). Ces transitions sont indiquées par des dégradés de couleurs au sein des enveloppes des mots les plus représentatifs des ensembles.

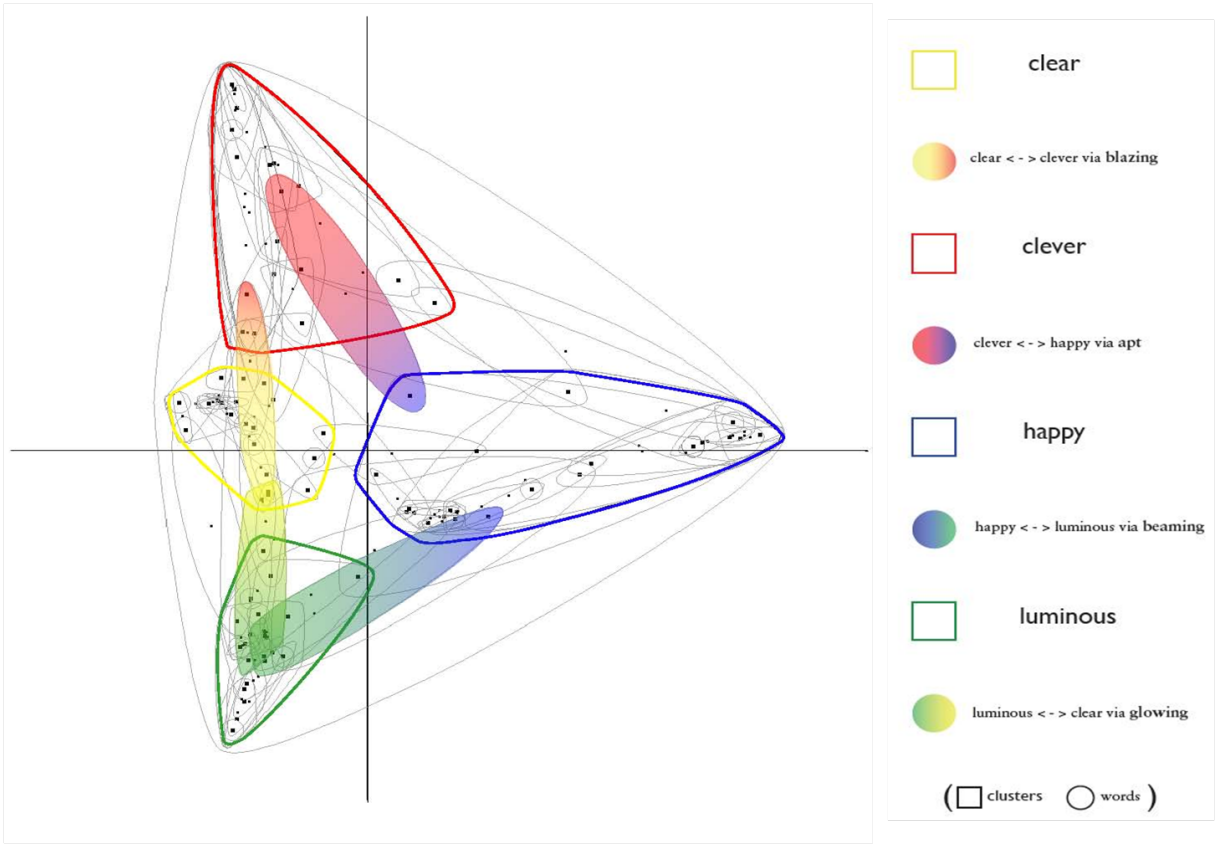

Figure 1 Représentation de la polysémie de *bright,* réalisée en collaboration avec Anne-Lyse Renon (Design graphique). Cette carte a été publiée dans (Ploux, Boussidan, et Ji 2010)

La continuité des sens observée avec les AS est transposée dans l'ACOM en terme de contextes. C'est la capacité du modèle à donner une place prépondérante à la polysémie, ainsi que sa capacité à représenter des vastes ensembles de données tout en conservant un niveau de détail extrêmement précis, qui en font un outil de choix pour observer le déploiement de la polysémie dans le temps. Néanmoins cette capacité est limitée à la représentation d'un à cinq mots sur une carte, car la représentation d'ensembles plus grands génère des cartes saturées et donc illisibles.

Pour adapter l'ACOM à l'étude de la diachronie, quelques ajustements ont été nécessaires. Le corpus est d'abord découpé en tranches temporelles d'un mois. Il est ensuite exploité au format lemmatisé ainsi qu'au format non-lemmatisé. Deux indices majeurs permettent ensuite de repérer les candidats au changement sémantique : la variation de fréquence des mots-entrées, ainsi que la variation de leur réseau de co-occurrence.

Pour ces études, le corpus presse *Le Monde 1997-2007* a servi de corpus test, néanmoins le corpus du *New York Times 1987-2007* ainsi que le corpus *Frantext*[9] et un corpus espagnol composé de *El Pais* et *Lavanguardia*[10] ont été utilisés de façon complémentaire.

Pour explorer ces corpus, de nombreux indices ont été testés. Il en résulte que les indices suivants se sont révélés pertinents :

INDICES MATHÉMATIQUES

Variation de fréquence

- La variation des fréquences brutes ou normalisées pour chaque mot-entrée. La fréquence normalisée est obtenue en divisant les fréquences brutes par le nombre total de mots dans la tranche temporelle et en multipliant le résultat par le nombre moyen de mots dans la tranche temporelle.

Lorsque la variation de fréquence est inhabituelle, cela indique que le mot étudié peut être affecté par un évènement ou qu'il est au centre d'une mode.

- Le coefficient de variation des fréquences. Ce coefficient correspond au rapport entre l'écart type et la moyenne.

- Le coefficient de régression linéaire (ou pente de la régression linéaire) pour chaque mot cible.

Structure des réseaux de co-occurrence

- La fréquence d'apparition d'un mot co-occurrent avec le mot-cible, indépendamment de ce dernier.

- La fréquence d'apparition d'un mot co-occurrent avec le mot-cible, relativement à ce dernier.

- La fréquence de co-occurrence normalisée : La fréquence du mot co-occurrent est divisée par celle du mot cible, puis multipliée par la fréquence moyenne de ce dernier.

- Les réseaux de co-occurrences triés hiérarchiquement pour un mot cible. La fréquence moyenne de co-occurrence est calculée pour chaque mot co-occurrent et les mots sont classés hiérarchiquement dans un tableau, avec la moyenne par mois pour chaque mot.

- Les rangs. Ils montrent l'évolution de l'importance hiérarchique des mots co-occurrents dans le réseau, par tranche temporelle. Ils font ainsi émerger la structure interne du réseau de contextes.

- L'indice de cohésion et de densité. Cet indice prend pour support le modèle ACOM et a été créé lors de cette thèse. Il se base sur le rapport entre le nombre de cliques générées et le nombre de mots co-occurrents. Ce rapport met en lumière la nécessité de créer une continuité entre des contextes existants et de nouveaux contextes. Pour que les nouveaux contextes entrent dans le réseau de façon cohérente, de nouvelles cliques sont créées. Plus ces cliques sont nombreuses par rapport au nombre de termes ajoutés, plus il y a de nouveaux contextes en jeu. Si les cliques sont peu nombreuses, la densité est élevée.

---

[9] Partiellement mis à disposition par l'ATILF que je remercie.
[10] Mis à disposition par le IULA que je remercie.

- Tous les indices précités peuvent être réappliqués suite à un redécoupage des tranches temporelles effectué en fonction des fréquences observées pour le mot cible.

INDICES LINGUISTIQUES

- Catégorie grammaticale

-Ponctuation et orthographe (un mot nouveau apparaît souvent entre guillemets, ou suivi d'une explication entre parenthèses, son orthographe varie souvent, comportant parfois un trait d'union).

-La productivité morphologique (préfixation, suffixation, composition), ainsi que la rétroactivité de la production morphologique sur le terme source (par exemple lorsque l'on crée le terme *anti-mondialisation*, cela a un effet sémantique rétroactif sur *mondialisation*). La productivité morphologique est un indice de la vie du mot cible, soit actif, soit en dormance.

-La compétition synonymique. Lorsque de nouveaux concepts sont en jeu, il existe plusieurs mots candidats pour y faire référence. Lorsque l'un des candidats « gagne » la compétition, les sens en jeu se redistribuent au travers des différents candidats.

INDICES EXTRA-LINGUISTIQUES

-Les informations sur le sujet, l'auteur, la rubrique, le genre, le style…

-Les informations supplémentaires, dites « encyclopédiques » sur les concepts en jeu, récoltées manuellement au travers de recherches bibliographiques, sur Internet, dans la presse, etc.

Ces indices ont été testés lors d'études de cas. Ces études comportent deux volets : un volet d'évaluation des fortes tendances dans les corpus, et un volet d'études détaillées d'exemples précis.

En appliquant les coefficients de régression linéaire à tous les mots pour chaque corpus, on observe les tendances d'emploi. Les chiffres obtenus sont soit positifs, montrant que l'emploi augmente, soit négatifs, montrant que l'emploi diminue. Par exemple, dans le corpus *Le Monde*, l'on voit que les acteurs politiques changent (le coefficient est positif pour *Sarkozy* mais négatif pour *Jospin*) et la prédominance de sujets politiques (comme les élections, avec des termes comme *candidat* dans les valeurs positives) par rapport à la diminution de sujets de société (avec des valeurs négatives pour *travail* ou *temps*). Les corpus en anglais et en espagnol montrent des tendances similaires. Le *New York Times* a des valeurs positives pour les termes liés à la guerre (*death*, « mort », *Irak, attack*, « attaque »), à l'avènement d'Internet (*online, web, internet, site..)* et à la vie de famille (*family*, « famille », *wife*, « épouse »), et des valeurs négatives liées au monde de l'entreprise et de l'argent (*company* « entreprise », *sale* « vente », *earn* « gagner »). Ces thèmes se retrouvent en espagnol, où l'on trouve également des valeurs négatives dans le domaine de l'impression, par opposition à Internet (*imprimir*, « imprimer », *edición*, « édition »). Ces rapports sont exacerbés dans *Frantext*, qui couvre plusieurs siècles de données, et pour lequel les résultats ont une forte valeur sociale et stylistique, avec des valeurs positives pour des termes comme *train* ou *guerre* et des valeurs négatives pour *cœur, dieu, roi* ou *âme*.

On applique également le calcul des coefficients de variation à ces corpus. Alors que le coefficient de régression linéaire permet d'extraire les tendances positives ou négatives, le coefficient de variation permet d'extraire les emplois instables. Ces emplois renvoient à des idées, des thèmes, des objets, des personnes ou encore des lieux. Par exemple, un thème typiquement instable par nature est celui de l'*éclipse*, mot dont le coefficient de variation est extrêmement élevé. Une fois de plus le domaine politique vient en première place. Puis, apparaissent les thèmes-choc, comme l'*anthrax*, le débat sur le *voile* islamique, les *caricatures* de Mohammed, la *canicule*, le *tsunami,* la *dioxine,* les *intermittents* du spectacle, l'*immigration*, les *virus*… Le thème du terrorisme est aussi très présent (*terrorisme, bombardement, attentat*). Les résultats sont similaires en anglais, complétés bien sûr par des scandales locaux comme l'affaire *Lewinsky,* et des évènements marquants comme les ouragans *(hurricane*) qui

touchent le continent. En espagnol, alors que le corpus est plus réduit, on voit une réelle spécificité locale, avec des thèmes liés à la dictature, au capitalisme, à la résistance ou encore à l'interdit (*dictadura, capitalismo, resistir, prohibido*).

Ces indices portent donc sur des thèmes à forte tendance négative ou positive ou à forte instabilité. Ils sélectionnent aussi des thèmes dont nous savons qu'ils ne sont pas candidats au changement sémantique. Ces thèmes sont filtrés. Par exemple, parmi les thèmes instables, se trouvent tous les mots liés à une apparition régulière dans le calendrier ou dont la programmation est saisonnière, comme les noms de jour, de mois, les jeux olympiques ou les élections. Ces mots sont donc sujets à des pics de fréquence très forts puis à des fréquences très basses voire nulles. Néanmoins, un terme comme *terroriste* est également sujet à un pic de fréquence à la période des attentats du 11 septembre 2001. Son comportement ne se distingue que par une augmentation de la fréquence moyenne après ce pic et par une productivité morphologique sans précédent après ce pic (*bioterroriste, islamo-terroriste, cyber-terroriste*, etc.). Il est donc délicat de ne se baser que sur les indices, et l'analyse manuelle est nécessaire après avoir appliqué ces indices comme filtres. La deuxième difficulté, bien plus complexe, réside dans la distinction entre la variabilité naturelle (aussi appelée fluctuation) et la variabilité liée à un changement sémantique. Les modèles mathématiques mettent ces deux phénomènes sur le même plan (à moins qu'ils n'aient recours à une ontologie ou à une base de données encyclopédique comme Wikipédia). En effet, les mots sont sujets à des emplois polysémiques mais également à des emplois idiomatiques variés. Plus un terme est riche, plus il est malléable ou « plastique ». C'est le cas par exemple du terme *bouquet*, dont le sens « bouquet de chaînes télévisées » s'impose progressivement, alors que les sens « bouquet » de fleurs, de senteurs, ou les emplois idiomatiques comme « c'est le bouquet » restent stable. Pour faire apparaître ce processus, il est nécessaire de calculer les rangs pour chaque contexte. En choisissant les indices pertinents pour réaliser un filtrage, on gagne en précision à chaque étape.

Cette méthode a été appliquée lors d'études de cas détaillées, portant sur le terme *malbouffe*, ainsi que sur la productivité en *mal-*, puis sur la productivité et l'ambigüité sémantique d'éléments de composition comme *crypto-, cyber-* et *bio-*, et enfin sur l'analyse de la compétition synonymique entre *mondialisation* et *globalisation*, ainsi que le changement sémantique de *mondialisation* lié à l'évolution de sa connotation. L'étude complète sur la *malbouffe* est publiée dans (Boussidan, Lupone, et Ploux 2009), et celle sur *mondialisation* dans (Boussidan et al. 2012).

Le terme *malbouffe* est un néologisme qui a été choisi dans une liste de mots nouveaux intégrés par les dictionnaires *Le Larousse* et *Le Petit Robert*, établie par Martinez (2009). Ce terme est intégré en un temps record par la presse et les dictionnaires et subit un changement sémantique au moment de son intégration. Le premier sens de *malbouffe*, alors orthographié *mal bouffe*, ou encore *mal-bouffe*, est lié à la diététique. Ce sens est celui donné par les auteurs Rosnay et Rosnay (1979) dans leur ouvrage de recettes diététiques. Il s'agit alors d'éviter les aliments trop gras, trop sucrés, etc. Ce sens perdure alors qu'un second sens s'y superpose : celui des conditions de production des aliments, lié à la société industrielle. C'est au travers du discours de José Bové que ce sens se répand. En effet les fréquences d'emploi de *José Bové* et de *malbouffe* sont très liées. On observe que les réseaux sémantiques de ces deux mots-cibles sont partagés. J. Bové inscrit sa « lutte contre la malbouffe » au sein d'une lutte idéologique plus globale, se dressant contre le consumérisme, les Etats-Unis, ou encore les OGM. Il s'agit pour lui de consommer de façon engagée. Ce sens est intégré par le dictionnaire tout en préservant le sens diététique. La médiatisation des actions de J. Bové participe alors à sa diffusion. Ainsi, on peut dire qu'il est le « père » de ce nouveau sens. Ce phénomène est renforcé par la médiatisation de la crise alimentaire, générant des inquiétudes relatives à la qualité de la nourriture et à sa toxicité potentielle (voir les débats autour des OGM, de la dioxine, de la vache folle, etc.). Comme le note Barraud (2008), *malbouffe* est une construction en *mal-* avec un substantif familier. Selon cette dernière, ce néologisme déclenche une « nouvelle vague de mots composés » en *mal-*. Cette hypothèse est étudiée. La Figure 2 montre tous les noms (en bleu) et les adjectifs (en vert) en *mal-* dans le corpus *Le Monde*, de nature nouvelle (non attestée par le dictionnaire ou très récemment attestée). La somme des fréquences normalisées des 172 adjectifs et noms (en rouge) montre une claire augmentation de cette tendance. L'entrée dans le dictionnaire du terme *mal-logement* en 2006 confirme l'hypothèse de lexicalisation des ces innovations. On trouve de nombreuses formes comme *mal-vivre*, ou *malbonheur*, au sein desquelles *mal-* subit un glissement sémantique. Il ne s'agit pas de quelque chose de mauvais, mais de ne pas faire les choses

« correctement », comme s'il y avait une « bonne » manière de vivre par opposition à une manière incorrecte. Cela s'applique à des termes comme *mal-gouvernance*, ou *maladministration,* reflets d'un mécontentement général.

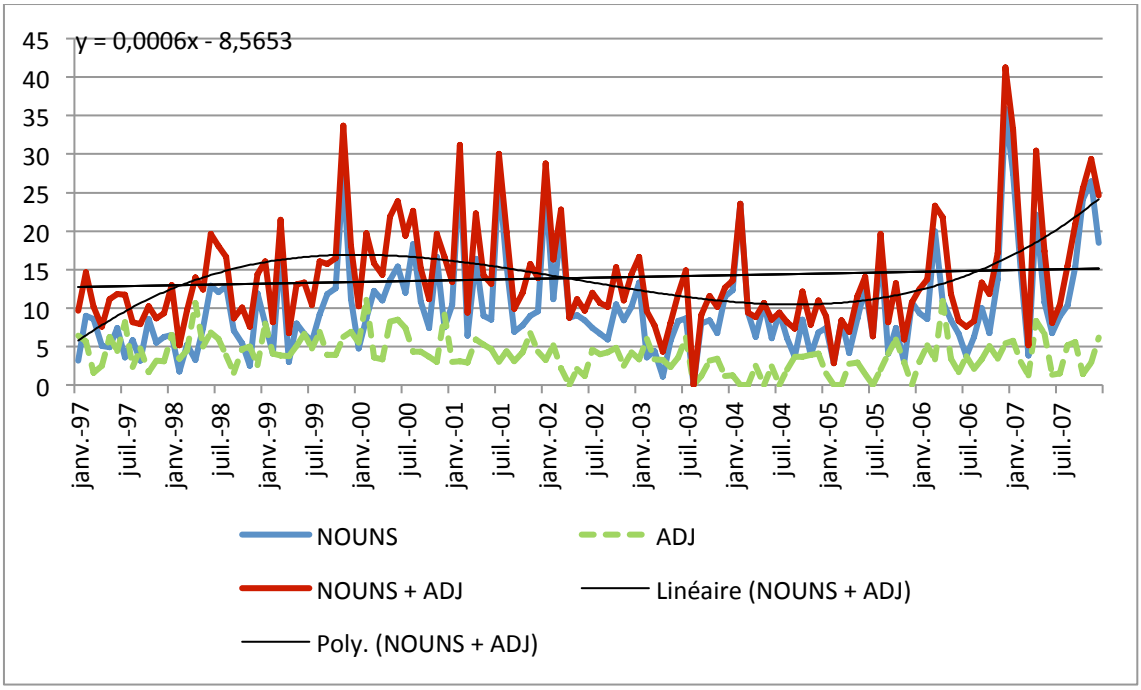

Figure 2 Somme des fréquences normalisées de tous les termes nouveaux en *mal-* dans le corpus *Le Monde* 1997-2007

Ce phénomène ne se limite pas à *mal-*. L'analyse de *crypto-* montre que les éléments de composition peuvent subir des changements sémantiques, se reportant alors sur les productions morphologiques dans lesquelles ils interviennent. *Crypto-* change de domaine, en passant de la biologie à la politique. Ce changement s'observe à la fois en français et en anglais, et il est délicat d'évaluer l'influence exacte d'une langue sur l'autre. Néanmoins, les innovations entrent plus rapidement dans les dictionnaires anglophones, ne rencontrant pas de discussion terminologique aussi poussée que celles qui ont lieu en France, au travers des commissions de néologie et de terminologie. De la politique, les associations s'étendent aux catégories sociales et religieuses. On trouve ainsi dans le corpus *Le Monde*, des termes comme *crypto-communiste* (attesté sans le trait d'union), *crypto-juif*, ou encore *crypto-gay* ou *crypto-gothique*. Parallèlement, les usages en biologie deviennent obsolètes, tout comme en anglais. *Crypto-* est utilisé dans le sens de « caché » comme dans *cryptobiose* (dit d'un organisme vivant ne montrant pas de signe de vie). Néanmoins, les règles morphologiques sont celles de la composition savante en biologie (composition basée sur un élément du grec ou du latin dans des domaines spécialisés), mais cela n'est pas le cas dans les domaines politiques, religieux et sociaux, pour lesquels le procédé de composition savante est imité mais non respecté.

L'étude de *bio-* montre que les nombreux sens portés par cet élément de composition peuvent être ambigus voir entrer en conflit. En effet, le sens premier provient du grec *bios*, c'est-à-dire la vie. Ce sens s'étend au travers de la biologie, et donc de l'étude de la vie. Puis il s'étend encore pour désigner le rapport entre la biologie et les autres disciplines. Le rapport entre la biologie et la technologie donne naissance à la *biotechnologie*, et *bio-* est parfois employé comme abréviation de *biotechnologie*, alors que des emplois de *biotech-* et de *biotechno-* persistent. Parallèlement à cela, un sens non attesté se crée sur la base de l'abréviation de l'adjectif *biologique*, dont le sens est lié aux conditions de production des aliments et des biens dans le respect de la nature. *Bio-* est également employé dans le sens d'*écologique*, là où l'élément *éco-* serait attendu. Cela fait résonance au terme *biodégradable* qui constitue un mot pivot entre la biologie, le biologique et l'écologie. Il en résulte que certains

termes peuvent posséder un sens ou son contraire, c'est-à-dire *bio-* pour « naturel et écologique », ou *bio-* pour « produit de la biotechnologie ». La Figure 3 résume les transferts de sens en jeu. Sur cette Figure, les sens liés par le cercle rouge sont attestés et les autres sens sont des extensions réalisées à partir de ce cœur. Le terme *bioplastique* atteste des conflits de sens qui se créent. Il est employé dans le corpus de la même façon que son équivalent anglais *bioplastic*, pour faire référence à des plastiques (partiellement) biodégradables, et donc plus écologiques que les plastiques classiques. Ce sens est attesté en anglais mais ne l'est pas (encore) en français. Les deux langues possèdent un adjectif homonyme plus ancien, provenant de la biologie, et qui désigne la capacité des cellules à se régénérer. Il s'écrit *bioplastique* en français. Ce cas de figure correspond à une compétition homonymique, qui, selon les règles de la morphologie devrait être « bloquée ». Néanmoins, l'usage semble faire fi de ces règles ainsi que des règles lexicographiques.

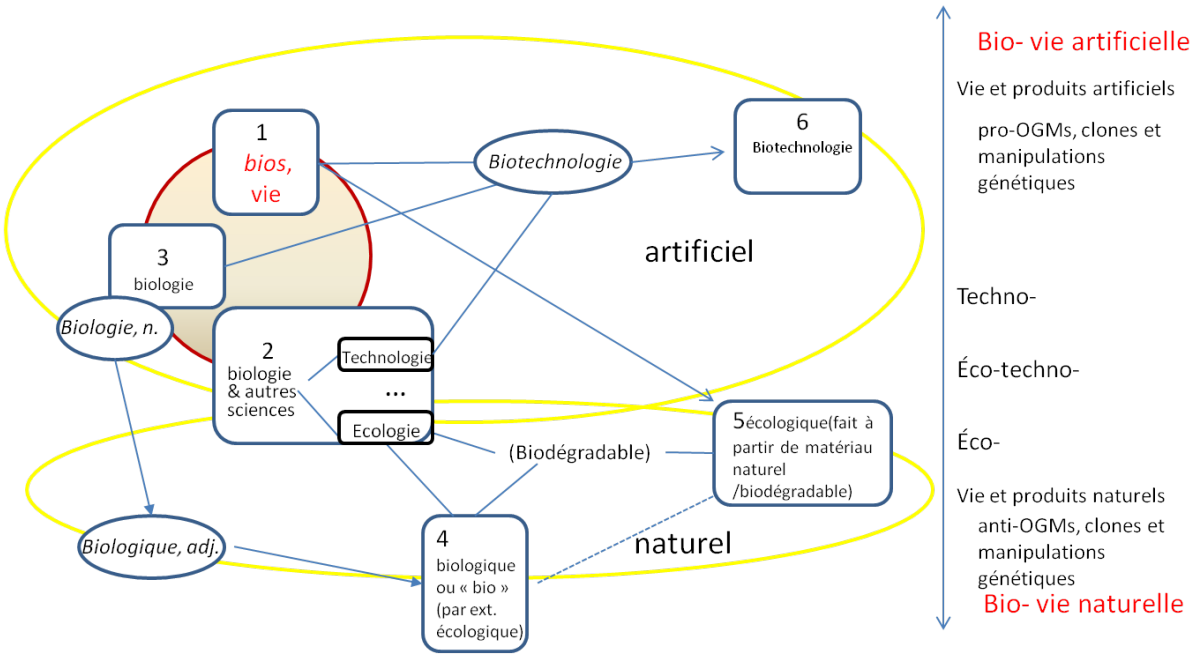

Figure 3 Sens hypothétiques détectés en corpus pour l'élément de composition *bio-* et chemins entre les sens.

Il existe encore un sens supplémentaire à *bio-*, dérivé de la biotechnologie dans son usage en anglais. C'est celui du rapport homme-machine. On le trouve dans des termes comme *biopunk* (dans les deux langues) ou *bionumérique*. On peut poser la question de l'usage de *cyber-* en lieu de *bio-*, puisque des composés comme *cyberpunk* existent. Il semble que l'élément *techno-* présent dans *biotechnologie* se soit partiellement transféré à *bio-* (au travers de *biotech-* et *biotechno-)* et que *cyber-* devienne progressivement obsolète (comme le montre la courbe de fréquence normalisée de la somme de tous les termes en *cyber-* dans le corpus).

Après avoir étudié la plasticité sémantique d'éléments de composition, je m'intéresse au procédé de compétition synonymique entre deux noms : *mondialisation* et *globalisation*. Le premier subit un glissement de connotation lié à l'usage alors que le second, qui est emprunté à l'anglais, se spécialise tandis que son usage est moins populaire. Ce procédé correspond à ce que Bréal (1899) appelle la *différenciation* (lorsque deux quasi-synonymes divergent et se spécialisent.) Alors que les deux termes rivalisent dans les domaines de l'économie, de la finance et de la politique, *mondialisation* se colore d'un usage politique avec une mineure économique, et *globalisation* d'un usage économique et financier. *Mondialisation* a une forte productivité morphologique (en *anti-* et en *alter-*, produisant *antimondialisation, antimondialiste, altermondialisme, altermondialisation et altermondialiste*, tous attestés, dans cet ordre) alors que *globalisation* ne produit que *antiglobalisation*, qui n'est pas attesté, mais imite l'anglais *antiglobalization*. Cette productivité est liée à la médiatisation d'un mouvement anti- mondialisation, qui ne définit pas cette dernière tout à fait de la même façon que ceux qui parlent de « mondialisation économique »

ou de « mondialisation financière ». Il semble que deux camps d'opinion se dessinent : ceux qui perçoivent la mondialisation comme un défi positif à relever, synonyme de progrès, et ceux qui la perçoivent comme une menace, contre laquelle il faut lutter, puis à laquelle il faut proposer une alternative (*alter-*). En générant des cartes ACOM par tranche temporelle, et en appliquant une interpolation, on obtient une visualisation dynamique de ces phénomènes, et on peut voir ces camps d'opinion se structurer, autour de termes clefs (ou pivots) comme *défi*, *menace* et *progrès*. Les cartes montrent une restructuration de l'espace et une hausse conséquente de la densité. La visualisation dynamique des cartes est consultable en ligne ici :

http://dico.isc.cnrs.fr/fr/diachro.html [11] Pour obtenir cette visualisation, un travail d'équipe a été nécessaire pour adapter l'ACOM aux besoins de la diachronie. Les cartes sont donc le produit d'un travail de développement informatique (réalisé par Charlotte Franco et Sylvain Lupone) et d'un travail en design graphique (Anne-Lyse Renon). Lors de cette collaboration, nous avons posé la question de la relation entre les sens et les formes dans une perspective expérimentale. La question de la visualisation des données s'est également posée au travers de l'usage de graphiques pour représenter des données, comme par exemple lors de la création de l'indice de variabilité de la densité et de la cohésion qui s'appuie sur les cliques. La Figure 4 montre le rapport cliques-termes pour les mots *mondialisation* (en rouge) et *globalisation* (en bleu) et montre deux phases de restructuration pour *mondialisation*, entre 1997 et 2001. Ces phases correspondent à l'appropriation du terme *mondialisation* par des groupes *anti-* et à leur médiatisation, rythmée par l'apparition des néologismes en *anti-* et *alter-*. Le premier terme co-occurrent de *mondialisation* est alors *contre*, dont le rang augmente globalement entre 1999 et 2004. Le terme est alors brandi comme slogan dans de nombreux évènements à partir de la fin de l'année 1999 (comme le Forum Social Mondial).

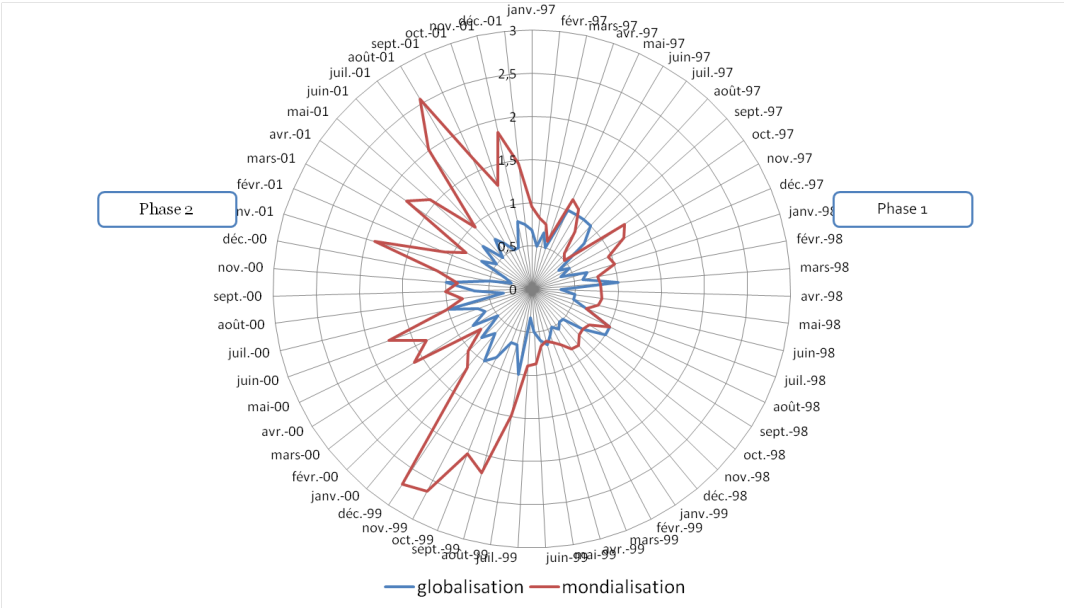

Figure 4 Indice de variabilité de la densité et de la cohésion pour *mondialisation* et *globalisation*.

Dans la visualisation dynamique (http://dico.isc.cnrs.fr/fr/diachro.html), la carte de l'année 2000 montre une prévalence de ce type d'emploi. *Mondialisation* correspond à ce que Hughes (1992) appelle un « mediated change » c'est-à-dire un changement sémantique diffusé par le média. Les glissements de sens et les transferts de domaine sont liés à une compétition synonymique ainsi qu'à une production morphologique, tous en lien avec des mouvements d'opinions portés par des évènements, qui colorent de façon péjorative ou méliorative les termes, et en lien avec la structuration de l'identité verbale de ces opinions. Ce changement sémantique, comme beaucoup d'autres, offre un miroir de la société dans le reflet du média.

---

[11] Cette visualisation est accompagnée d'un texte explicatif, consultable en anglais: http://dico.isc.cnrs.fr/en/diachro.html et en espagnol : http://dico.isc.cnrs.fr/es/diachro.html

Les procédés décrits par la théorie sont bien présents, mais toujours dans des combinaisons différentes. Ainsi, il ressort que l'étude de l'articulation et de la dynamique combinatoire des procédés constitue une poursuite intéressante aux typologies et aux analyses produites par la linguistique. L'approche choisie permet de faire ressortir plusieurs niveaux de granularité : le niveau des unités lexicales et des mouvements conceptuels internes au mot, le niveau du mot, et le niveau des ensembles conceptuels formés par des réseaux de mots et d'associations contextuelles et conceptuelles. Le lien entre ces niveaux semble s'effectuer par le biais de termes pivots, autour desquels les glissements de sens s'articulent.

Les études de cas présentées confirment la nécessité d'adapter les paradigmes d'analyse de l'observation du changement sémantique en corpus à sa vitesse et à sa complexité actuelles. Ces études ont servi à développer un prototype informatique de détection et d'analyse de ces phénomènes. Ce prototype n'est en soi qu'une première pierre apportée à un domaine encore naissant. Il ouvre de nombreuses perspectives, comme la création d'interfaces dynamiques de l'étude du changement sémantique permettant aux utilisateurs d'extraire des données à partir de corpus, et la possibilité d'appliquer ces méthodes à des corpus web. Les méthodes décrites ici peuvent également être transférées à d'autres outils du TAL et de la gestion d'information dans le futur.

Bibliographie

Altmann, E. G., J. B. Pierrehumbert, and A. E. Motter. 2011. "Niche as a Determinant of Word Fate in Online Groups." *PLoS ONE* 6 (5).

Barraud, C. 2008. "La Malcomposition." In *Nomen Exempli et Exemplum Vitae: Studia in Honorem Sapientisimi Iohannis Didaci Atauriensis*, by J. A. Pascual, Sasgo Ediciones. Madrid.

Barsalou, L.W. 1982. "Context-independent and Context-dependent Information in Concepts." *Memory & Cognition* 10 (1): 82–93.

Benzécri, J. P. 1980. *L'analyse Des Données. II: L'analyse Des Correspondances*. Paris: Bordas.

Blank, A. 1999. "Why Do New Meanings Occur? A Cognitive Typology of the Motivations for Lexical Semantic Change." In *Historical Semantics and Cognition*, by A. Blank and P. Koch, 61–90. Berlin/New York: Mouton de Gruyter.

———. 2003. "Words and Concepts in Time: Towards Diachronic Cognitive Onomasiology." In *Words in Time: Diachronic Semantics from Different Points of View*, edited by R Eckardt, K. Von Heusinger, and C. Schwarze, 143:37–66. Mouton de Gruyter.

Blei, D.M., A.Y. Ng, and M.I. Jordan. 2003. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research* 3: 993–1022.

Bloomfield, L. 1933. *Language*. New York: Allen & Unwin.

Bourdieu, P. 1979. *La Distinction: Critique Sociale Du Jugement*. Éditions de Minuit.

Boussidan, A. 2013. "Dynamics of semantic change. Detecting, analyzing and modeling semantic change in corpus in short diachrony". Thèse de doctorat. Université Lumière Lyon 2.

Boussidan, A., S. Lupone, et S. Ploux. 2009. "La malbouffe: Un cas de néologie et de glissement sémantique fulgurants." *"Du Thème Au Terme, Émergence et Lexicalisation Des Connaissances" 8 Ème Conférence Internationale Terminologie et Intelligence Artificielle*. Toulouse, France.

Boussidan, A., and S. Ploux. 2011. "Using Topic Salience and Connotational Drifts to Detect Candidates to Semantic Change." *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011), Oxford, UK.*

Boussidan, A., A.L. Renon, C. Franco, S. Lupone, et S. Ploux. 2012. "Repérage automatique de la néologie sémantique en corpus à travers des représentations cartographiques évolutives. Vers une méthode de visualisation graphique dynamique de la diachronie des néologies." Edité par J. F. Sablayrolles. *Cahiers de Lexicologie* n°100, Néologie Sémantique et Analyse de Corpus: 117–136. Paris, Classiques Garnier.

Bréal, M. 1899. *Essai de Sémantique*. Paris: Hachette.

Chesley, P. 2011. "You Know What It Is: Learning Words through Listening to Hip-Hop." Edited by P. Holme. *PLoS ONE* 6 (12) (December 21).

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Clarke, D., and B. Nerlich. 1991. "Word-Waves: A Computational Model of Lexical Semantic Change." *Language and Communication* 11 (3): 227–38.

Cook, P., and S. Stevenson. 2010. "Automatically Identifying Changes in the Semantic Orientation of Words." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 28–34.

Corbin, D. 1987. *Morphologie dérivationnelle et structuration du lexique*. Linguistische Arbeiten. Tübingen: M. Niemeyer.

Coseriu, E. 1958. *Sincronía, Diacronía e Historia: El Problema Del Cambio Lingüístico*. Investigaciones y Estudios. Montevideo: Universidad de la republica. Facultad de Humanidades y Ciencias.

Dury, P., et P. Drouin. 2009. "L'obsolescence des termes en langues de spécialité: une étude semi-automatique de la «nécrologie» en corpus informatisés, appliquée au domaine de l'écologie." In *Online Proceedings of the XVII European LSP Symposium*, 2010:1–11.

Firth, JR. 1957. "A Synopsis of Linguistic Theory." *Studies in Linguistic Analysis*.

Gulordava, K., and M. Baroni. 2011. "A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus." In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, 67–71.

Heyer, G., F. Holz, and S. Teresniak. 2009. "Change of Topics over Time and Tracking Topics by Their Change of Meaning." In *KDIR 2009: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*.

Holz, F., and S. Teresniak. 2010. "Towards Automatic Detection and Tracking of Topic Change." *Computational Linguistics and Intelligent Text Processing*: 327–339.

Hughes, G. 1992. "Social factors in the formulation of a typology of semantic change." In *Diachrony within synchrony: language history and cognition. Papers from the International symposium at the University of Duisburg, 26-28 March 1990*, edited by Günter Kellermann and Michael D Morrissey, 107–124. Duisburg Papers on Research in Language and Culture 14. Frankfurt am Main: Peter Lang.

Ji, H. 2005. "Étude d'un modèle computationnel pour la représentation du sens des mots par intégration des relations de contexte". Thèse de doctorat. Institut national polytechnique de Grenoble.

Joseph, B.D., and R.D. Janda. 2005. *The handbook of historical linguistics*. Blackwell handbooks in linguistics. Malden (Mass.): Blackwell Publishers.

Keller, R. 1994. *On language change: the invisible hand in language*. Translated by B. Nerlich. London: Routledge.

Klemperer, V. 1975. *LTI, la langue du IIIe Reich: carnets d'un philologue*. Translated by E. Guillot. Agora. Paris: Pocket.

Lakoff, G., and M. Johnson. 1980. *Metaphors We Live By*. Chicago, London.

Landauer, T.K., D.S. McNamara, S.E. Dennis, and W.E. Kintsch. 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates Publishers.

Lebart, L., M. Piron, and J.F. Steiner. 2003. *La Sémiométrie*. Dunod. Paris.

Lüdtke, H. 1999. "Diachronic semantics: towards a unified theory?" In *Historical semantics and cognition*. Edited by A. Blank and P. Koch, 49–60. Cognitive linguistics research. Berlin: Mouton de Gruyter.

Lund, K., and C. Burgess. 1996. "Producing High-dimensional Semantic Spaces from Lexical Co-occurrence." *Behavior Research Methods Instruments and Computers* 28 (2): 203–208.

Martinez, C. 2009. "L'évolution de l'orthographe dans les Petit Larousse et les Petit Robert 1997-2008: une approche généalogique du texte lexicographique". Thèse de doctorat. Université de Cergy-Pontoise.

Meillet, A. 1906. "Comment Les Mots Changent de Sens." *Linguistique Historique et Linguistique Générale*: 230–271.

Mitchell, J., and M. Lapata. 2008. "Vector-based Models of Semantic Composition." *Proceedings of ACL-08: HLT*: 236–244.

Nerlich, B., and D. Clarke. 1988. "A Dynamic Model of Semantic Change." *Journal of Literary Semantics* 17 (2): 73–90.

———. 1999. "Elements for an Integral Theory of Semantic Change and Semantic Development." In *Meaning Change–Meaning Variation. Workshop Held at Konstanz*, 1:123–134.

Orwell, G. 1949. *1984*. Signet Classic. New American Library.

Padó, S., and M. Lapata. 2007. "Dependency-based Construction of Semantic Space Models." *Computational Linguistics* 33 (2): 161–199.

Pagel, M., Q.D. Atkinson, and A. Meade. 2007. "Frequency of Word-use Predicts Rates of Lexical Evolution Throughout Indo-European History." *Nature* 449 (7163): 717–720.

Ploux, S., A. Boussidan, and H. Ji. 2010. "The Semantic Atlas: An Interactive Model of Lexical Representation." In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*,Valletta, Malte.

Rastier, F. 1999. "Cognitive Semantics and Diachronic Semantics." In *Historical Semantics and Cognition*. Edited by A. Blank and P. Koch, 109–144. Cognitive linguistics research. Berlin: Mouton de Gruyter.

Renouf, A. 2007. "'Tracing Lexical Productivity and Creativity in the British Media:'the Chavs and the Chav-Nots'." In *Lexical Creativity, Texts and Contexts*, edited by J. Munat, John Benjamins Publishing Company, 61–89. Amsterdam/Philadelphia.

Reteunauer, Coralie. 2012. "Vers un traitement automatique de la néosémie. Approche textuelle et statistique". Thèse de doctorat. Université de Lorraine.

Rohrdantz, C., A. Hautli, T. Mayer, M. Butt, D.A. Keim, and F. Plank. 2011. "Towards Tracking Semantic Change by Visual Analytics." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short papers-Volume 2*, 305–310.

Rosnay, Stella, and Joël Rosnay. 1979. *La Malbouffe - Comment Se Nourrir Pour Mieux Vivre*. Olivier Orban.

Sablayrolles, J.F. 1996. "Néologismes, Une Typologie Des Typologies." *Cahiers Du CIEL*: 11–48.

Sagi, E., S. Kaufmann, and B. Clark. 2009. "Semantic Density Analysis: Comparing Word Meaning Across Time and Phonetic Space." In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 104–111.

Stern, G. 1931. *Meaning and Change of Meaning; with Special Reference to the English Language*. Bloomington: Indiana University Press.

Traugott, Elizabeth Closs, and Richard B Dasher. 2002. *Regularity in semantic change*. Cambridge studies in linguistics. Cambridge: Cambridge University Press.

Ullmann, S. 1951. *The Principles of Semantics*. Oxford: Blackwell.

———. 1953. *Descriptive Semantics and Linguistic Typology*.

———. 1962. *Semantics: an introduction to the science of meaning*. Oxford: Basil Blackwell.

Utsumi, A. 2010. "Exploring the Relationship Between Semantic Spaces and Semantic Relations." In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*, Valletta, Malta.