U N I V E R S I T E L U M I E R E – L Y O N 2

Ecole Doctorale de Neurosciences et Cognition (NSCo)

Laboratoire sur le Langage, le Cerveau et la Cognition (L2C2), CNRS UMR 5304

Thèse de Doctorat de Sciences Cognitives, Mention Linguistique

Soutenue publiquement le 27 Juin 2013 par :

Armelle Boussidan

DYNAMICS OF SEMANTIC CHANGE

Detecting, analyzing and modeling semantic change in corpus in short diachrony

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

Devant un jury composé de :

Dr. Sabine Ploux : Directrice de thèse	Université de Lyon, L2C2, ISC, Fr.
Pr. Johannes Kabatek : Rapporteur	Université de Tübingen, Allemagne
Pr. Antoinette Renouf : Rapporteur	Université de Birmingham, GB
Dr. Anne Reboul : Examinatrice	Université de Lyon, L2C2, ISC, Fr.
Dr. Christophe Gérard : Examinateur	Université de Strasbourg, LILPA, Fr.
Dr. Pascaline Dury : Examinatrice	Université de Lyon, CRTT, Fr.

Laboratoire sur le Langage, le Cerveau et la Cognition (L2C2)

CNRS, UMR 5304

Institut des Sciences Cognitives

67 boulevard Pinel, 69675 BRON cedex

Remerciements

Bien que la thèse soit une aventure éminemment solitaire, force est de constater, à l'heure des remerciements, qu'elle est également le fruit d'interactions collectives.

En premier lieu, je tiens à remercier Sabine Ploux, qui a dirigé cette thèse, et m'a ouvert la porte de la recherche et des sciences, en relevant à mes côtés le défi d'un travail interdisciplinaire, chose éminemment précieuse.

Je remercie Pascaline Dury, Christophe Gérard, Johannes Kabatek, Anne Reboul et Antoinette Renouf de m'avoir fait l'honneur d'accepter de participer à mon jury et au développement de ma thèse.

Je remercie la Région Rhône-Alpes d'avoir retenu le projet « Modélisation Sémantique Dynamique en TAL et pour le Web » soumis à l'appel « Cible, créativité-innovation » en 2009. Ce financement a permis de réaliser une grande partie de ce travail, et m'a permis d'assister à plusieurs conférences internationales qui ont contribué à façonner ma pensée.

Je remercie l'ATILF (Nancy) et le IULA (Barcelone) pour m'avoir donné accès à des données.

A ma famille, sans le soutien de laquelle cette thèse n'aurait pas été possible, j'adresse ma profonde gratitude. Le psychiatre et la plume n'y sont sûrement pas pour rien dans mon choix étrange de décortiquer la langue et les sens.

Je remercie de tout cœur mon amie Daniela Issa, qui a généreusement passé de nombreuses heures à relire ce manuscrit, à me corriger, à discuter de grammaire anglaise comme de relations internationales, café après café.

Je remercie tous les membres (passés et présents) du L2C2 et de l'ISC qui m'ont accompagnée dans cette thèse. En particulier les ingénieurs Sylvain Lupone et Anne Cheylus qui sont les auteurs de programmes utilisés dans cette thèse, tout comme Charlotte Franco et Eric Koun. Merci à Sylvain Maurin pour le soutien informatique et la bonne humeur. Merci à Patrice Berger de m'avoir guidée et d'avoir obtenu les documents les plus introuvables. Merci à tous les autres étudiants pour ces moments partagés (ils se reconnaîtront). Merci en particulier à Raphaël Fargier, pour son soutien crucial et son amitié, les heures d'ébullition intellectuelle et celles à se casser la tête devant Excel. Merci également à mes colocataires de bureau successifs Guillaume Barbalat et Pia Aravena pour ces heures passées ensemble, à échanger, se marrer ou galérer. Un remerciement tout particulier à Anne-Lyse Renon, pour la richesse de nos collaborations et de nos échanges. Un grand merci à Edmundo Kronmuller pour ses conseils, sa générosité, et les discussions partagées.

J'aimerais remercier également tous les chercheurs que j'ai rencontrés, ici et là, qui m'ont aidée et avec qui j'ai collaboré, discuté, appris ... Merci en particulier à Eyal Sagi, pour ces échanges passionnants.

Merci à tous ceux qui ont parcouru les bibliothèques pour moi.

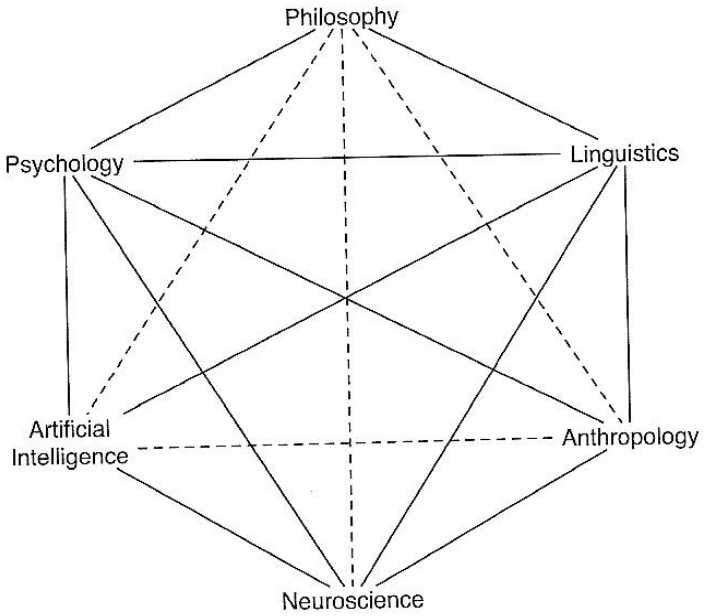
Merci à tous mes amis, qui ont été les témoins des hauts et des bas, et m'ont donné la force de mener à bien ce travail. Merci à mes amis strasbourgeois, à Juliette Steydli, et à mes compères du Lost Room, avec qui j'ai partagé de précieux moments de créativité et d'émotions.

Merci à mon compagnon Frédéric Monnoye, qui a vaillamment traversé les moments les plus difficiles à mes côtés. Merci Fred pour les petits plats, la patience à toute épreuve, les explications sur la vie des ordinateurs, et les idées saugrenues…

Et pour finir, merci à tous les inconnus avec qui j'ai discuté de malbouffe, de démondialisation ou de biopunks. Ils sont le piment de ce parcours.

"Most advances in science have been the result of intermediaries venturing beyond the boundaries of the paradigms of their disciplines, uniting insights which come from different kingdoms of knowledge."

T. Zeldin (1994:160)



Connections among the cognitive sciences, taken from Thagard (2009: 243)

Abstract

This doctoral thesis aims at elucidating the mechanisms 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.

TABLE OF CONTENTS

List of Figures

List of Tables

List of abbreviations and symbols

Languages

En. English
Gr. Greek
Fr. French
Lat. Latin
ME. Middle English
OE. Old English
Sp. Spanish

Dictionaries

OED Oxford English Dictionary
PR Le Petit Robert (2010 edition)
GRLF Le Grand Robert de la langue française (2011 updated edition)
TLFI Trésor de la langue française Informatisé

Models

SA Semantic Atlas
ACOM Automatic Contexonym Model

Note on translated text

All translations are mine, except when the translator's name is given. All remaining mistakes are mine.

INTRODUCTION

I. Objectives

The goal of this thesis is to pave the way and contribute to the understanding, detection, analysis and modeling of semantic change in short diachrony, in a context of mass production of textual data.

This work was mainly conducted with a French press corpus, using distributional semantics methods; however some equivalent data in English and Spanish was used complementarily.

The choice of working with corpora was born out of the recognition that leading a study of semantic change on language is close to an impossible task. In effect, language is a very ungraspable object of study. I have first restricted the research to written language, leaving aside spoken interaction and pragmatics. Nevertheless, spoken language would precisely be the best observatory for language dynamics. To understand dynamics of language one should theoretically anchor their work in dynamic sources of language but studying them imposes a number of technical restrictions and makes it difficult to establish any frame of reference whatsoever. Apart from spoken interaction, we have at our disposal written dynamic sources of language such as the Internet, which is constantly fed with textual data. To carry out a project based on a direct flux corpus from the Internet, there is a need for very complex technical and computational structures. Even though excellent Internet crawlers are operational in both academic and private research, they are still in the course of development and are not yet easy data to access or work with. Moreover, working directly with the Internet would be a project for a large group of people relying on substantial technological resources and multidisciplinary skills. Therefore, I have decided to create the closest simulation of a dynamic language input that can be worked with in the framework of a PhD. The present work has been led on finite corpora to understand some of the fundamental mechanisms of semantic change. The ideas that are developed are beyond the scope of a static corpus exploration, even if the conclusions I have reached are only applicable to this framework.

The methods, as well as computational and theoretical tools that are presented here are meant to be testable at higher scale in future research, and are a small contribution to the scientific

community, in a field that is just starting to unfold in its full force, and just integrating the numerous possibilities that new technologies offer us.

The choice of working with a finite press corpus is therefore motivated by the need to rely on a stable frame of reference. Within this framework, I have set the lines for a prototype of detection, analysis and modeling. This prototype is by no means in a finite state and calls for further experimentation, extensions, and applications. My work is thus purposefully focused on understanding fundamental mechanisms rather than creating large scale detection tools.

Language change is one of the backbones of linguistic theory and is an immense subfield of linguistics in itself. I have more specifically centered my research around semantic change, leaving aside grammatical issues but including some morphology and discourse analysis issues when needed. Semantic change asks how words change meaning over time, and questions both the mechanisms involved and the causes. My approach is both analytical and computational in that it partly relies on a pre-existing mathematical model of representation, the Semantic Atlas, developed by my PhD supervisor Sabine Ploux and numerous contributors, at the L2C2, in the "Modeling language" team[1], as well as on case studies comprising detailed analyses. The latter involve some sociological issues in that they are rooted in real world events. The question of how semantic changes are related to these is raised.

To start, I define what semantic change is and the related terminology.

II. Definitions

Semantic change, language change and linguistic change

From the variety of approaches encountered in the literature and among active researchers in the field, I can attest without doubt that there is no clear agreement on the terminology involved in the treatment of semantic change. The terminological discussion itself is central to seminars, books or projects on the subject (for instance it showed itself central to the *Semaneo*

[1] http://l2c2.isc.cnrs.fr/en/teams/modelang/

colloquium[2] on the topic). This ongoing discussion aims at creating an established terminological framework in the area, and is still subject to numerous disagreements.

Indeed the phrase "semantic change" encompasses a lot of phenomena. First it is distinct from its hyperonyms "language change" and "linguistic change". The wording "language change" includes all that is related to language evolution, from its historical, biological, genetic and cultural aspects down to its linguistic aspects. It deals with historical linguistics at large with a specific focus on causes rather than mechanisms. By contrast, "linguistic change" concentrates more on the form than the external causes, and encompasses all types of changes, phonological, phonetic, grammatical, syntactic and lexical. The above definition is wide and is discussed by several linguists with several degrees of refinement. For instance, "Posner (1997:3)[..] distinguishes between linguistic change (which affects "dynamic systems …[having] their own mechanisms of change") and language change (since "the language of a community, as an entity, can change")" (Joseph and Janda 2005: 84).

Semantic change and neology

Semantic change -or *drift*- looks at how word meaning evolves in time at the level of words and lexical units. As such, it is a topic pertaining to diachronic and historical linguistics. "Semantic change" is sometimes used as a synonym of "semantic neology" however terminologists generally talk about semantic neology when the change becomes lexicalized and attested by lexicography, that is to say when lexicographers judge it right to include the new meaning of a word in the revised version of a dictionary. Semantic neology thus seems to be the result of the process of semantic change, or a specific case of semantic change, that is, when semantic change eventually gives birth to a new meaning-form pair ready to enter the lexicon.

Nevertheless, the question of when semantic change becomes semantic neology is critical on the conceptual point of view. Indeed how does one draw a line between conceptual change and a semantic neology ready to enter the lexicon? This question echoes century long discussions on the nature of concepts, words, and word meaning, and how they are related. It

[2] The colloquium's proceedings are published in : Cahiers de Lexicologie n° 100: Néologie sémantique et analyse de corpus. (Sablayrolles 2012)

more specifically poses the question of where to draw a line to assert that we are dealing with semantic change and no other phenomena. To delineate a threshold, issues regarding time, fixation, diffusion, lexicographic decisions, amplitude of meaning variation, social connotation, event-based semantic bursts and numerous other factors enter into play.

Semantic change is thus tightly linked to semantic neology. Neologies are more precisely defined theoretically than semantic change, even if there is still some amount of disagreement about the different types of neologies. A generally accepted traditional distinction is made between formal neology, grammatical neology and semantic neology.

Formal neology is when a new word enters language, for instance the word *hotspot* related to the development of Internet wireless technology. It is, in a way, the easiest to detect, on the basis of frequency counts.

Grammatical neology is when a word changes grammatical category, for instance when we create the verb *to eternalize* from the adjective *eternal*. The latter too may come out with simple frequency analysis.

Semantic neology, the most complex of the three, is when a given word changes meaning. It is itself dividable into a few major mechanisms: broadening (e.g. *dog* used to refer to a specific breed and shifted to refer to all dogs), narrowing (e.g. *deer* used to refer to wild animals and shifted to "deers" only) metaphor (e.g. *mouse* has changed from describing the animal to describing the technological device as well) metonymy (e.g. *White house*), melioration and pejoration (e.g. *nice* meant "ignorant" at first). These mechanisms are themselves dividable in more sub-mechanisms that should be dealt with in more detail in the course of this work.

"On distingue classiquement deux sortes de néologismes : le néologisme ordinaire, unité pourvue d'une « forme » et d'un « sens » nouveaux, et le « néologisme de sens », acception nouvelle pour une unité déjà constituée. […]La néologie sémantique est un cas particulier de la polysémie, avec un trait diachronique de nouveauté dans l'emploi, donc dans le sens.[…]La néologie sémantique est toujours produite ou repérable par le contexte, le

contexte étroit de la phrase ou du syntagme où s'insère l'unité, le *contexte large* du domaine discursif de référence" [3] (Bastuji 1974: 6).

This largely accepted view is not universal. For Guilbert (1975: 59), for instance, there are four types of neologies: phonological, syntagmatic, semantic and borrowed:

"[…] nous proposons de prendre en considération quatre formes de néologie : la néologie phonologique, la néologie syntagmatique, la néologie sémantique et la néologie d'emprunt. La première consiste dans la formation de la substance du signifiant et dans sa transcription. La seconde englobe tous les modes de formation qui impliquent la combinaison d'éléments différents ; elle est morpho-syntaxique et rassemble toutes les formes de dérivation indépendamment de la place respective des composants, de la nature formelle de leur relation, qu'elle se présente sous la forme du mot ou de plusieurs mots. La troisième consiste dans la mutation sémantique sans création d'une substance signifiante nouvelle ; elle est du domaine du signifié. La quatrième définit les différents aspects de l'emprunt à une langue étrangère [4]" (Guilbert 1975: 59).

Semantic change primarily refers to changes in the meaning of lexical units, and may also encompass issues related to compositional semantics when strings of words are concerned as in compounds and collocations. Compositional semantics poses further challenges outside of the scope of this work, although we shall tackle a few compound forms encountered while conducting case studies.

---

[3] "We traditionally distinguish two kinds of neologisms: the ordinary neologism, a unit provided with a new "form" and a new "meaning", and the "sense neologism", a new meaning for an already formed unit. [...] Semantic neology is a special case of polysemy, with a diachronic novelty feature in use and therefore in meaning. [...] Semantic neology is always produced or detectable via context, the *narrow context* of the sentence or phrase where the unit is inserted and the *broad context* of the discursive field of reference. "

[4] "we suggest considering four types of neologisms: phonological neologisms, phrase neologisms, semantic neologisms and neologisms from borrowings. The first consists in the forming of the substance of the signifier and its transcription. The second includes all modes of forming that involve the combination of different elements; it is morpho-syntactic and brings together all forms of derivations regardless of the respective roles of the components, and of the formal nature of their relationship, whether existing in the form of a word or several words. The third consists in semantic mutation without the creation of a new substance; it belongs to the domain of the signified. The fourth defines the various aspects of loans from a foreign language

I therefore encompass formal neology, some grammatical neologies (although they are not central to this investigation but rather peripheral), semantic neologies, semantic changes at large, and most importantly the key to all these phenomena: connotational drifts.

Connotational drifts

Connotational drifts are the progressive changes in connotation of a word. Connotation, taken in its largest definition, encompasses all the associated concepts that come with a word, at the symbolical and imaginary level as well as through collective reference. This implies notions such as positive and negative poles and the scalar grades in between them. For instance if one uses the word *fat* to talk about someone, it is negatively connoted, whereas *fleshy* or *corpulent* would be more neutral. Connotation is deeply rooted in context use. A word acquires its connotation in context, and the contexts of use become part of the word's implicit features.

Connotation is generally opposed to *denotation*; the first refers to associated meanings while the second refers to the representation or symbol. Denotation corresponds to the literal meaning as listed by the dictionary, while connotation includes associated concepts. For instance the denotation of the word *rose* is the flower while its connotation is love.

The nature of change

We can only assess change after it has taken place and not during its unfolding. The precise moment at which an object undergoes change is very hard to pick since processes of change are very often gradual. If one were asked to determine at which point black turns into white, or white into black, confronted to various shades of grey, we shall get different answers as there isn't a single answer to this question. Moreover, as Kosko (1993: 8) puts it "*the world is grey but science is black and white*. We talk in zeroes and ones but the truth lies in between." Kosko points out that it is almost inadequate to try and understand change within a binary framework. This remark applies even more to language as the object of change, since language has several subjective and abstract dimensions that a binary framework cannot grasp.

Not only the nature of change is complex but the nature of the object under study, language, is equally complex. To look at language change one must look at the mechanisms of change in general in the first place, and second, at the nature of the object that is observed, language. By

doing so one may look at the nature of language change. In this sense, theories based on continuous paradigms are better adapted to the issue that discontinuous ones.

One of the easiest metaphors to work with, in that context, is the image that language is organic, and therefore that we can analyse its change the way we look at biological processes. The idea of dealing with language change as biological change has fascinated many thinkers in the XIXth century. They viewed language as a living organism and asserted that the causes of language change were contained within language itself (see Joseph and Janda 2005). This view has been radically left aside as language became more and more analyzed as a system, with the advent of structuralism and later systemics, and while Noam Chomsky's generative grammar gained influence. These views also implied that the causes of change were contained within language. However, with structuralism and its philosophical rooting into anthropology, the link between sociology, psychology and language emerged. Sociolinguists look for causes in society while psycholinguists do so within our minds and through our inter-subjective interactions. Both ideas –that causes of change are contained within or outside language- might hold their share of truth. However, it might be in the way those two levels interact that most processes take place. Moreover, it is worth noting that the main concern of language change research has shifted from being the quest and analysis of causes, to being more concerned with mechanisms, and dynamics. This is related to systemics, but also to new discoveries in other areas of science, especially physics. To analyse change in language we may consider that language is a system and can further wonder whether dynamics of change established by physics, and quantum physics, especially thermodynamics, apply. However it is quite jarring to determine whether language should be considered an open or a closed system, since it is both. Indeed in the case of a corpus, we are dealing with a finite set of sentences; however this set is produced within a context, and includes the reflection of changes happening at the level of society.

Considering that change is an extremely delicate issue on the philosophical point of view in the same way that time is, there is no pre-established assumption about the nature of change in this research.

Some of the cited approaches take language as a system, but also as its own referential context. However, not only do the direct, linguistic contexts of use bear an important impact on word meaning change, but meaning evolution taken as a whole is also rooted in

sociological and historical contexts. Therefore, in a press corpus study, two types of contexts have to be considered: the direct linguistic context of a word (the sentences or documents it appears in) and the global social context in which the language production is rooted in.

Context

Social context: a paradigm shift

Nowadays, Internet has become the most popular tool to access information in Western societies. In this context, a means to access information has in itself an impact on meaning; to such an extent that the linguist David Crystal talks about "Internet linguistics" (Crystal 2005). Internet makes language evolve more quickly and also provides an observation spot for this evolution (see Kilgariff and Grefenstette, 2003). This new need demands a vast and fundamental research into the nature of language dynamics, in the context of its newly acquired scale of production and diffusion. This scale itself is constantly bursting, and by doing so creates the need for extremely adaptable tools in terms of the quantity of data being dealt with.

Internet is not only a place for retrieving novelties but is also a great place to store and retrieve knowledge, old and new. In other words, Internet has a great potential to become the interactive archive of human knowledge. With the computerization of documents and books, Internet is already on the path of becoming an electronic library. However, do we know how to navigate within this library properly? One of the challenges of an efficient navigation relies on understanding and encoding as precisely as we can the relationships between words and concepts.

Language and language change are confronted with a major paradigm shift, which can be compared to the impact of the printing press, the telephone, TV, the radio or the cinema, and as such it bears new challenges. This research is embedded in that paradigm shift and cannot answer to its complexity, but tries to take a very small step towards understanding language dynamics in its context of mass production and flux. Not only are we dealing with a shift regarding the nature of language, as established by Crystal (2005) but this shift is part of a larger sociological shift to do with the velocity of exchanges, may they be economic, political, sociological, geographic, or verbal. The time needed to communicate is constantly being reduced, reaching the apex of instantaneity and sometimes going faster than the human mind

can handle, along with a highly computerized society, of which the "social network" is the utmost symbol.

Previous analyses of change, before that shift, were based on the world influence linguistic systems were exposed to; the time of writing, publishing, acquiring, sharing, etc. those data. These have been overtaken by high paced language contact, making dialectal, cross-linguistic borrowings and calques rocket. Moreover, these phenomena are embedded in a context of immediate international cultural contact with the influence of modern media (see McLuhan 1994).

Presently, we do not have the necessary hindsight to analyse the impacts of that paradigm shift on language. However, we can try and look into the process as it unfolds. This may be done at macro- and micro- levels. Phenomena under the linguistic microscope may help understand the organization of the structure as a whole. Micro- and macro- lenses seem necessary to grasp patterns in such large amounts of data, as Zeldin puts it:

"What to do with too much information is the great riddle of our time. My solution is to look at the facts through two lenses simultaneously, both through a microscope, choosing details that illuminate life in those aspects that touch people most closely, and through a telescope, surveying large problems from a great distance" (Zeldin 1994).

The challenge, therefore, is to explore how language evolves within the instantaneous. I have tried to isolate some phenomena, with the perspective in mind that these phenomena might reveal some of the fundamental structure of semantic change, and might provide frameworks to further test whether they replicate at higher levels and across different types of corpora.

Scientific context

Traditional vs. computational linguistics.

Scientific contributions have also greatly evolved, since the tools to analyse language have diversified and become more accessible, with the growing introduction of computers and computerized models to carry out linguistic research. Operational linguistic tools have clearly become necessary to deal with the spectacular growth in quantity of available electronic information. Moreover, to process language with machines, one needs to work with a representation of language. Models of language add another layer of complexity: in the same

way that traditional linguists were stuck in the cycle of creating metalanguages based on the object of study, and could not further distance themselves from the object at stake, language engineering creates models that organize language in ways that impact and bias the conclusions the models might provide. Whether the representation is a metalanguage, a vector space, or a rule-based learning framework, it always colours the object under study.

The understanding of processes of change is necessary, while tools societies develop have to be in tune with the new dynamics of language at the heart of a major paradigm shift of language production, but is also necessary to the understanding of the results of that production. Therefore there are two needs: one related to producing future communication tools and one related to the analysis of what is produced. The latter is anchored in a socio-linguistic approach and provides an interdisciplinary analytical tool, while the former is anchored in a mechanistic approach that involves a computational view as well as modeling issues. This work shall try and bridge those two aspects. A traditional linguistic analysis would be impossible to carry out on great amounts of data, while a purely computational analysis could encompass those data but miss out on several fine-grained aspects of language that it cannot yet encode and deal with. Indeed, in the past decades the advances in NLP have allowed for mass treatment of linguistic data and large corpora, creating new possibilities for research in this area. Automatic translation, data mining, automatic summary production, question answering software, sentiment detection … are all born out of the great burst of computational linguistics which provide powerful and fast tools. However, as technology helps us dealing with more data, faster and more efficiently, it also misses the fine details that only a careful human analysis can provide. In the process of dealing with large amounts of data, how do we take into account the more subjective, abstract or poetic side of language?

This has generated a dichotomy within the scientific community between the advocates of a linguistically centered approach and the advocates of a computationally centered approach. The former claim a lot of the meaning contained the data is lost, at the profit of figures and measurements that cannot, by nature, encompass pragmatic and sociological aspects such as intention, irony, cultural reference, … or even idiosyncratic aspects. The latter claim that with the high paced changes language is undergoing within an extremely technological society, we have to adapt to dealing with the language production we are confronted with, and there is no time to delve into extremely detailed theoretical considerations. While the first insist on the

fundamental need for a strong theoretical understanding, the second concentrates on providing applicable tools that answer the needs of society and science. Although tremendous progress has been achieved in computational linguistics and semantics, we cannot yet obtain the results that we obtain through a more classical and applied analysis with a computational approach in terms of theoretical understanding. This is because the mystery of language rests in its ungraspable nature and its constantly evolving identity. Language can never be looked at as an isolated system as it is in constant interaction with the societies and people that produce it. It therefore contains a certain amount of subjectivity and relationship to world events that cannot be made sense of in a purely computational approach, unless that approach makes room for it. This difficulty applies not only to linguistics but to most cognitive sciences which are faced with the ungraspable nature of human experience. Moreover, that difficulty applies as regards a static object of study but it applies even more to a dynamic object of study.

This question is in itself a subject of study and my position is to find a safe balance between the computational and analytical approaches. The type of data under study cannot benefit at this stage from a fully automatic treatment. However, semi-automatic analysis may be conducted safely providing human analysis is substantial enough. Therefore, I shall rely on computational tools to explore large amounts of data and extract items that seem interesting. Models are used to provide a general view of the organization of data, and then items of interest are isolated to analyse them in detail with a more traditional linguistic, philological and sociological approach. The detailed analysis of specific cases of semantic change at a fine-grained level is therefore coupled to a global view of semantic trends related to a whole corpus in the third part of this work.

The representation of language, from static to dynamic

Moreover whether the analysis is manual or semi-automatic, it most of the time relies on a static picture of language. I postulate that this static picture is a fundamental flaw. Not only do we need to integrate technology for scale and velocity reasons, but we also have to try and work toward dynamic models rather than static models, to fully respect the nature of language. In semantics, models and computational tools largely depend on the availability of ontologies and thesauri which index the meaning of words and concepts, either for language at large or in a specialized domain. However, these ontologies and thesauri, when they are available, fail to take into account one of the fundamental characteristics of semantics: word

meaning change and dynamism. All types of modern communication and information management interfaces using language have to be in tune with their environment. The interactive society-language cycle is in ever changing transformation. Consequently, tools that try to grasp it need to be dynamic. To create functional dynamic tools, one has to understand how the dynamics work first.

Most of the available language software relies on ontologies and thesauri, and these tools are now overtaken by the dazzling evolution of language. Indeed, the very nature of these tools relies on a conceptual organization that has to be pre-determined. This organization then needs to be maintained even if observed changes do not fit into pre-established semantic categories or relationships. These tools thus lack the flexibility to truly adapt to a changing paradigm. Therefore, dynamic tools must be created, in which the nature of dynamics is taken into account. To create those dynamic tools in NLP in the future, we must first understand the nature of language change dynamics.

Surprisingly, word dynamics has been widely overlooked by scientists until recently (at the exception of pioneer work such as Nerlich and Clarke 1988; and Clarke and Nerlich 1991), since modern concerns have been centered on the immediacy of communication. Information has become available on the spur of the moment, in all types of format.

II. Methods and frameworks

Corpus

Internal vs. external causes of change

As stated previously, semantic changes have more to do with the societies, communities and cultural backgrounds they are rooted in than with purely intra-linguistic phenomena. Such intra-linguistic phenomena are traditionally analyzed in finding the pathway between an old meaning A and a new meaning B by applying rules internal to language taken as a system.

However, the real engine behind change might rather be found in a social event shaking society at some point, of which the intra linguistic system can say nothing. As such, press corpus exploration radically differs from a more Saussurian perspective in which language bears the reasons of its own behaviour (although some corpus studies are led accordingly to

Saussurian theory). The corpus anchors the analysis in society and makes this work a piece of sociolinguistics rather than pure or computational linguistics.

As Forston (2005) notes in his conclusion, it is necessary to distinguish between the process of semantic change, its result and its diffusion. Corpus use provides a picture of a quite advanced state of diffusion in the linguistic community. It captures an early stage of large scale diffusion but we do not know yet whether this diffusion will reach further, and become robust and sustainable.

However, a corpus restricts the analysis to its scope. No corpus may represent the state of a language at any given point in time. It only represents one specific sample of it, and induces cultural, linguistic and stylistic restrictions.

Stylistics and specialized terminology

Therefore, a corpus stands as the closest simulation of a linguistic community one can work with in a written format for large scale exploitation. The corpus stands as the image of a specific linguistic community in a time frame. A well structured corpus also possesses a certain stylistic homogeneity. This characteristic allows for a better control over the detected phenomena. Stylistic variation in corpus is in itself a field of study which involves discourse analysis at large scale levels. It involves several layers of analysis, such as the relationships between different styles, and the ones between style and genre, language and genre, style and idiolect (idiosyncratic style), text, genre and style, and so forth, as described by Gérard (2011). In a stylistically homogenous corpus, it is easier to distinguish idiosyncratic innovations from ones that spread, or to detect when an idiosyncrasy acquires the power to spread to a language community. Other niche corpora may be used such as specific magazines, Internet blogs and forums designed for sub-communities in music, fashion, or the professional world. Using reference corpora in comparison with specialized corpora to detect how words and expressions transfer from a small linguistic community to a more general acceptance pertains partly to specialized terminology. The methods used in specialized terminology are partly similar to the methods implemented here, and partly different in that they include more sociological considerations and analysis of the chosen specialized area. A strong example of this type of work may be found in (Picton 2009).

Specialized press and communication possesses its own stylistic framework. The advantage of press corpora is that they are stylistically quite homogenous, since journalistic style is well defined by a set of rules. It is therefore a good compromise to deal with language "at large" in a written format. The journalistic style, however, possesses its own rules and involves editorial decisions. It most of the time uses a "politically correct" language, and as such provides a good picture of what changes have become "acceptable" enough to enter journalistic writing. However, this attitude is too shy to include innovations quickly, and the press corpus is not the first place to encounter them, it is rather the place that tells us they have become widespread enough so that most people can decode them.

An overview of the main approaches

Long vs. short diachrony

Most of the referential frameworks and methods of study developed by previous contributors have been developed to tackle long diachrony issues at the level of centuries. They rely on large scale corpora (such as *Frantext*[5] in French or the *Project Gutenberg* in English[6]) including a variety of styles (mostly literary and journalistic), as well as dialectal and orthographic variations. To include these contributions, it is necessary to question the adaptability of the methods of long diachrony to short diachrony.

Are the observable phenomena in long and short diachrony similar? Can short diachrony mechanisms constitute a basis for prediction of long term changes? At the present time, this question still needs further contributions from the scientific community. This work encompasses approaches in detection and analysis and does not offer a framework for prediction; however it paves the way towards prediction possibilities. Detection and prediction are very much corpus related, and therefore the question of the applicability of methods across different types of corpora adds up to the initial issue. Predictive hypotheses may be formulated with the help of probabilistic tools that shall be mentioned in Part. I. With very recent corpora, we are looking at a process as it unfolds, and have almost no way to assess whether what we are looking at will sustain itself. The main difference between long

[5] http://www.*Frantext*.fr/

[6] http://www.gutenberg.org/

and short diachrony is that short diachrony studies in corpus extract a lot of event-based phenomenon as well as other types of bursty phenomenon which are smoothed out in a long diachrony approach. This has a double impact: on the one hand one can analyse trends and events and their possible impact on language, and therefore provide tools for trend watching, on the other hand it makes it harder to delineate the linguistic from the sociological and to set aside effects which are intrinsic to the corpus at stake. However, event-based changes show a particular profile, either seasonal or bursty and can thus be identified and dealt with separately.

To conduct short diachrony studies, what are the frameworks established by long diachrony studies that are relevant to be applied in short diachrony too?

Linguistic approaches & frameworks for analysis

To conduct a study on semantic change, we have at our disposal the contributions of several theoretical schools. Some of them provide purely theoretical tools that were developed without working on corpora and have been applied to corpora more recently. Traditional (non-computational) diachronic approaches include semasiology and onomasiology, typological works, structural analysis, and prototype theory analyses in cognitive linguistics. With NLP, recent approaches rely on ontologies or language models, within which I shall mainly look at vector space models.

In lexical semantics, semasiology and onomasiology have been the frameworks of understanding word meaning change since the XIXth century. Onomasiology starts from concepts and asks how they can be named, in synchrony or diachrony. Conversely, semasiology takes the word as a starting point and asks what are the concepts related to it.

In structuralism, in the spirit of De Saussure (1916) the classical framework of analysis is based on the distinction between the *signified,* which equals the concept referred to and the *referent*, the real world entity the word refers to, associated to the *signifier,* the word.

In later structural semantics, especially in European works, semantic change can be understood as the change in the distribution of *features*.

Word meaning change can also be analyzed in terms of changes in i*ntension* (roughly the properties and features of a word) and *extension* (roughly the referents it points to), following Frege's (1948) theory and subsequent applications by numerous linguists.

In the framework of prototype theory, it can be modeled as a shift in *category*, in continuation with Rosch's theory (1973 ;1978), or as a reorganization in conceptual distance between words as is posited by Wittgenstein (1953). There are numerous contributions in cognitive linguistics relying on this framework and extending it, such as (Geeraerts 1997). It can also be understood in change in the *qualia structure* within Pustejovsky's (1995) theory as done by Adelstein (2007).

With statistical tools in corpus, several neological observatories and research teams rely on exclusion or reference corpora, which may be compilations of dictionaries to which they compare text corpora to extract neologies, as demonstrated by works conducted at the ATILF[7] or the Observatory of Neology of the group IULAterm[8]

One of the main aspects to differentiate the scope of all the cited tools, rather than their intrinsic value, is how they are applied and to which object. In modern linguistics, the central object to apply them is a corpus. Work on made-up examples is discarded.

Typological approaches& beyond

Even though there is a multiplicity of possible approaches, most of the research in semantic change has been of a typological kind. Typologies are generally not based on corpus studies. They try and elaborate lists of causes and mechanisms to classify types of change. Many of them are available, offering different classifications, and leaving doubt as to which of them stands out. One of the major limits of typologies is that they mostly look at the results of change rather than its process, or reanalysis, as stated by Forston (2005: 650) :

[7] « Analyse et Traitement de la Langue Française » (analysis and processing of the French language) a CNRS laboratory located at the University of Nancy.

[8] University Pompeu Fabra, Barcelona.

> " …typologies themselves are beside the point. The reason is that they refer to the *results* of change; they leave entirely untouched the reanalyses (innovations) that are the true changes and that are of primary interest."

Most of the listed types can be combined and extended and there is no agreed list of them between linguists. Moreover, many cases of semantic change cannot be described by the most accepted typological sets. An example of this is the change of the verb *to realize* meaning "to bring to fruition" and then "to understand" cited by Forston (*ibid*: 652), that does not fit into any established category.

Typologies do not cover the spectrum of semantic phenomena we are interested in. However, they are still discussed, extended and redefined among linguists. The lack of agreement over typologies is rooted in the fact that semantic change as reanalysis is rooted in use, in particular community use, and when it becomes robust in a given linguistic community, it can expand to other communities. As such, no textbook may grasp it, as change takes on as many forms as speakers create. The power of interaction on innovation is central in Traugott and Dasher's theory (2002) in which change primarily emerges through reinterpretation within each interaction. This theory integrates a pragmatic dimension where typological approaches lacked rooting in social interaction and the relationship of language with the world.

Multidisciplinary approaches

To look at how words change meaning, one has to encompass the systems in which these words are embedded. Cognitive sciences and previously systemics have set a path toward this type of analysis, bridging research between areas as diverse as linguistics, philosophy, anthropology, sociology, brain sciences, mathematics, computer sciences, and modeling at large.

Within this interdisciplinary ground I have chosen to bring together semantics, socio-linguistics and language modeling and representation, using computational tools to allow for treatment of large amounts of data. The computational tools intervene to enable dealing with big corpora, at the detection, or extraction stage. They allow for context variation extraction on the basis of indicators such as frequency, as shall be detailed later on.

In synchrony, words appear in a multiplicity of contexts in a somehow stable pattern. The stability of that pattern may be disrupted by a trend according to which one of the uses of a word becomes pregnant. While observing this phenomenon in short diachrony we are dealing with the same process if the word simply undergoes a temporary trend or if the context shift marks the beginning of a deeper connotational drift that will eventually lead to semantic change. At this stage, the need for human analysis arises, since computational tools cannot yet take into account the linguistic background and native intuition of whether a word "sounds new" or not, and whether it may be related to a specific event.

Therefore, the chosen approach is to combine the computational data with applied analysis. I shall introduce a few available computational models that may be used in the detection and modeling part.

Available models in NLP

With the advent of NLP, semantic change can be measured in terms of semantic distance or similarity as in distributional semantics, working with vector space models, as in Van de Cruys (2010). Measures of density (Sagi, Kaufmann, and Clark 2009) or volatility (Holz and Teresniak 2010) as well as change in polarity (Cook and Stevenson 2010) have been offered very recently to specifically evaluate meaning variation.

A few recent studies in distributional semantics that implement semantic change detection in corpus also provide graphical outputs, with very similar approaches as the one chosen here, such as (Holz and Teresniak 2010).

Indeed, among the computational tools available in NLP a few models are more adapted to the modeling of unfolding processes of change than others. As mentioned earlier the static ontologies are too limited for this purpose, therefore widespread tools such as Wordnet (Fellbaum et.al 1998) should be left aside. However, models based on graph theory and on vector spaces are of interest since they provide spaces that may constantly be redefined. There are a plethora of models of this kind such as Latent Semantic Analysis[9] (LSA) (Landauer et al. 2007), Hyperspace Analog to Language (HAL)(Burgess and Lund 1997), or other types of

[9] http://lsa.colorado.edu/

word vector space described in full length in (Sahlgren 2006) and (Van de Cruys 2010). As noted by Utsumi (2010) these spaces are excellent grounds to represent different types of semantic relations. Considering they can be represented synchronically and they have better flexibility than other models to encode and represent changes, vector space models seem to be among the best tools at our disposal to model semantic change. They have already been used for very similar studies in lexical variation and related phenomenon, as in (Peirsman and Speelman 2009). Lexical variation shows extremely related processes to semantic change, at the difference that the first unfolds across dialects in synchrony, and the second within one given language in diachrony.

One of the major flaws of a certain number of vector space models though is that they associate one single vector per word. This is a major limitation, since semantic change is rooted in the diachronic restructuration of polysemy. By contrast the Semantic Atlas model relies on correspondence analysis[10] (Benzécri 1980) and attributes a several vectors per word since the space is built on the basis of a unit smaller than the word: the clique.

Chosen approach

The model used in this study is ACOM (Automatic Contexonym Organizing Model), developed by H. Ji (Ji and Ploux 2003; Ji, Ploux, and Wehrli 2003), an extension of the Semantic Atlas created by S. Ploux (1997). The Semantic Atlases were originally created to handle synonymy relations, and H. Ji extended their use by handling co-occurrence relations in raw non annotated corpora. The original model provides maps obtained via a correspondence factor analysis, offering three degrees of granularity: the level of the word, and above the one of clusters of words (or related thematic sets) and under the one of *cliques* which are fine grained infra linguistic sub-units of meaning. The maps involve a new interpretative tool in that they bring possible graphic readings, as well as interactive navigation.

The chosen approach here is exploratory rather than typological. This choice relies mainly on the idea that the different phenomena we are looking at may be interconnected. We shall miss this interconnection if we try and look at types separately.

[10] In French "correspondence factor analysis"

The focus of this work is to detect, analyze and model semantic change processes in corpora with the help of ACOM and new extensions based on it. Primary work was conducted on the French press corpus *Le Monde* (1997-2007).

The chosen approach is to explore the corpus relying on the model and adding tools to it to answer the needs of this exploration.

To detect context change, I used word co-occurrence, following the principles of distributional semantics. This is rooted in the idea that "you shall know a word by the company it keeps" (Firth 1957: 11). On the basis of Ferdinand de Saussure's (1916) structuralist view of language, in which the meaning of a sign has no intrinsic value but acquires one via its existence within a set of other signs, Zellig Harris (1954) asserted that semantic meaning may be studied on the basis of the distribution of signs in a given system.

Networks

Using networks is necessary to grasp the complexity of one word's changes. Indeed one cannot isolate words as if they were independent items for two main reasons:

1) Most words are polysemous and changes in context use is a natural consequence of polysemy in synchrony or diachrony.

2) Words are part of a language which evolves as a whole, and that language itself is part of a complex cycle involving sociological and psychological aspects.

Word co-occurrence

Word co-occurrence studies apply basic or advanced statistics to word use in corpus, relying mainly on the frequency patterns, as explained in (Baayen 2008). Word co-occurrence looks at the words that come before or after the target word that is chosen for study. Co-occurrence can be extracted according to different windows: one can choose to retain a certain number of words coming before and/or after the target word (n-grams), or the whole sentence or paragraph the word is part of. This choice is a syntagmatic one; however, one may choose to include paradigmatic information too, or to mix syntagmatic and paradigmatic levels.

Connotational drift

Semantic change is rooted in one main process: connotational drift, and more rarely on a radical change in connotation. However, all words undergo connotational drifts at some level and it is jarring to draw a line as to when they do undergo semantic change *per se*. Connotational drift is based in context change and is the technical door used here to detect semantic change. Sometimes, connotation change may happen without further impact on the target word. However, this temporary phenomenon has its interest as it will impact the (re-) distribution of meanings in the semantic network related to the target word.

However, even though context use and connotation are strongly related, they are intrinsically two different things. One more question adds up: when does context use variation point to connotational drift?

Most obviously, I have not answered the theoretical question of how to draw temporal, lexicographic, connotational or typological lines, and as far as this doctoral thesis is concerned, I have chosen to deal with semantic change at large. I have focused on connotational drift as the main indicator of change to try and draw the line of words being subject to meaning drift. I shall consider two major interrelated dimensions to try and draw that line: the temporal and lexicographic dimensions. However, how to set a threshold for each of these is still an open question.

Time

Semantic changes can happen without becoming fixed, and therefore without bearing semantic neologies. Do we still call this semantic change at all? Or are we dealing with an epiphenomenon of change processes? How much time do we need to assume that the new meaning(s) have become fixed? Moreover, is this time limit itself changing with the paradigm shift language is undergoing along with technological innovations?

In this sense very short diachrony is a very ambiguous timeframe to deal with. Moreover, as linguists we do not have serious scientific references to rely on as there are very few studies in short diachrony (on decades) conducted with modern computational tools and the modern data sets. In traditional diachrony, the time unit is at the level of the century, in modern short diachrony however, time units are still very subjective.

Lexicography

The only reference we have up to now is the one of lexicography. In a nutshell, we can say that when a new meaning enters the dictionary, it becomes attested. However the dictionary alone is not a sufficient reference in our study, since lexicographers generally wait a few years before they can judge whether the word has properly entered usage. Some words even undergo entering and exiting dictionary editions over years. Therefore, while working with a very short temporal window, most of the cases we look at have not yet entered the lexicographers' datasets. In addition, different dictionaries choose to include different items, following their own policies, either progressive or conservative, making the strongest reference we have –the dictionary- rather blurry.

However, the Internet stands as a new database of lexical knowledge, and new terms that have not entered dictionaries yet will often have entered the Internet through forums, discussions or even reference sites such as Wikipedia. The value of a word's presence on the Internet is not yet rated as such by the academic community; however it is taken as a point of reference by most researchers.

Multidisciplinary explorations

Since the model provides a graphical output, I have tried to integrate the graphical tool too in the research, to further explore how graphs could be used to decipher data, at the level of reading, or to make data more readable, at the level of producing them. This aspect has benefited from the precious input of another student who is a graphic designer. Together, we have formulated a few propositions to lay the path for a graphically dynamic model based on interpolation. Interpolation issues are still in the course of study and have generated works in the team, such as Xia (2011)'s study on the mathematical aspects of interpolation.

IV. Applications and implications

Who cares?

One of the questions PhD students often have to answer is: what is the use of this research? Understanding semantic change mechanisms is a topic pertaining to the realm of fundamental research, however it also opens paths in applicative technologies, sciences and thinking.

Across disciplines, past and present works have shown this, but the issue's understanding also bears the seed of promising future applications.

Indeed, this topic has impact in the academic world, the media, sociology, history, philosophy and politics, in the corporate world and at state level.

The fundamental aspect of this understanding is more relevant to cognitive sciences, psychology, philosophy and evidently NLP in that it contributes to understanding mental structure and processes, and therefore to modeling and applicative issues. It indirectly impacts Artificial Intelligence.

The purely applicative aspect however is relevant to a wider set of disciplines, academic or professional. The centrality of understanding the diffusion on concepts, ideas and related expressions is heightened by the recently acquired importance of social networking.

Therefore, understanding the fundamental mechanics of semantic change can benefit the academic world, including NLP, lexicology, lexicography, terminology and translation and reach beyond the academic world, for all software based on natural languages. It also impacts sociology and history, and all critical approaches of society and the media based on language.

Sociological, philosophical and political analysis of the media, in reality and fiction

Semantic change analysis is a powerful tool to understand our societies. Deciphering the semantics of a society provides a picture of it and also a ground for predicting its future evolution. Such deciphering was conducted in societies submitted to drastic changes and have provided a solid frame of reference to detect and analyze the changes that were taking place. For example, Klemperer (1975) studied the German language in the context of the creation of a totalitarian language in Hitler's times. This study on totalitarian language and how it progressively included meaning drifts to make some concepts acceptable by the masses may be completed by the linguistic insights provided by Orwell in his fiction "1984" (1949), which comprises a linguistic appendix on the principles of meaning drifts in language. To unfold his theories, Orwell created a whole fictive language called 'Newspeak' built on the basis of 'Oldspeak' (the English language) according to a set of semantic and morphological rules. The controversial idea behind this fictive language was to show how the impoverishment of language made constructed analytical thought unavailable to the mind.

Presently, the position assumed by Klemperer (1975) in observing on a daily basis the evolution of word meaning, has been rendered more complex but also richer with the advent of the Internet and the corresponding statistical tools necessary to browse it. Therefore various researchers conduct similar studies in a sociologically and politically critical perspective. One of the outstanding blogs of the kind in the French language is the one of Jean Veronis[11].

The French *textométrie*, *lexicométrie* and *sémiometrie* [12] apply statistical methods to texts to extract politically and sociologically relevant information. They were in the first place created for marketing and communication issues (see Lebart, Piron, and Steiner 2003).

These approaches are relevant for journalists, may they need to characterize a politician's speech or may they want to find information that is harder to access. In the same way, they are interesting for people who analyse the media, either in a purely sociological perspective, or as a basis for decision-making involving financial investments or more simply branding, brand awareness and verbal identity. Historians could also benefit from such approaches to scan the evolution of ideas and the connections between events in written texts.

The corporate world and governing bodies

Most evidently these needs arise across the corporate world as a whole. Companies need to take language issues into account to decide on their verbal identity, marketing, advertising and communication line. More than anyone else, they need to be closely in tune with language evolution to communicate efficiently. They also need tools in database management, data mining, forecasting and trend analysis. The former three may be used at the technological, strategic, social, economic and media levels in competitive sectors, or by governing bodies as well as security and intelligence agencies. These applications are also linked with the language policies at governmental level as the treatment of terminological evolution is a concern for terminological commissions in countries such as France or Spain. Since the issue is central to very different areas, the developed frameworks and tools in research need to be as flexible and adaptable as possible.

[11] http://blog.veronis.fr/

[12] Literally "textometrics, lexicometrics and semiometrics". These disciplines apply statistical methods to texts to extract politically relevant information.

V. Structure of the work

In the first part, I outline the state of the art in semantic change research, in historical linguistics, structural, cognitive and corpus linguistics as well as NLP. The state of the art is divided in three sections: theoretical approaches, the transition from pure theory to corpus-based analysis in a context of social, technological and scientific paradigm shift, and finally studies based on context models.

In the second part, I outline the methodology, the tools relied on and especially the model and how it can contribute to the issue. First, I describe the SA and ACOM models and explain why I have chosen them to deal with this topic, mainly due to their strength regarding the treatment of polysemy and granularity, as well as their capacity to handle raw text with very little input from the researcher, allowing for a truly exploratory approach with the minimum amount of theory imposed on data as a starting point. In this sense, the SA paradigm provides an empirical framework that is as little as possible rooted in theoretical assumptions that could bias the obtained results. Then, I describe the extensions that have been created on its basis and how they contribute to the issue. These extensions are then brought together into a model of semantic change detection, analysis and modeling, in the form of a methodology giving birth to a preliminary prototype. The adaptability of this prototype to create a dynamic model of semantic change is discussed.

In the third part, I present tools and indices followed by applied case studies in formal and semantic neology, semantic change and connotational drifts along with the theoretical questions they raised. These case studies are intrinsically related to the methodologies presented earlier, since they allowed for its creation. Rather than testing pre-established cases of semantic change, I have tried to browse the corpora to extract trends, and zoom in on items that showed highly representative of these trends. In the course of analyzing those items, a certain number of hypotheses emerged: First, I postulate that a semantic change is rooted in a process involving whole networks and semantic fields rather than isolated units. Secondly, the nature of these units has an impact on the possibility of a change. For instance, some words are more “plastic” than others, in that they have a higher degree of polysemy and a greater number of idiomatic and contextual uses in synchrony. Not only the words as items may be more or less “plastic” but the networks they are embedded in may also be characterized in terms of plasticity. The hypothesis is that these words along with the networks they are part of

are more prone to semantic change than less “plastic” ones. Moreover, in the co-occurrence networks some words seem to act as “pivots” around which meanings reorganize. Pivots are key co-occurrent words around which meanings transfer and end up entering the target word to modify its meaning or create a formal neology built on it. Formal neology productivity itself may act retroactively on the target word, redefining its scope in the light of the newly established network extension. Therefore, I shall demonstrate through a few case studies how semantic networks and fields, pivot mechanisms, plasticity and nature of the target words, along with retroactivity processes, show central to the understanding of the processes underlying semantic change. These hypotheses are discussed in the light of the examples. Finally, the conclusion deals with possible applications and future perspectives.

PART I SEMANTIC CHANGE: A STATE OF THE ART

Chapter I.1 : Semantic change in linguistics: theoretical approaches

"la lengua cambia *para seguir funcionand*o como tal."[13] (Coseriu 1958: 17)

Theoreticians have been dealing with the topic of semantic change since the XIXth century. Initially, the main focus of study was set on classifying the reasons for semantic evolution. The central questions for linguists were the causes and forces, or processes, underlying change. The works of Paul (1880), Bréal (1899), Darmesteter (1887), Littré (1888), to cite only a few notable ones, and later (Bloomfield 1933) offer typologies that list stylistic devices and processes involved in semantic change and establish inventories of neologisms resulting from it. These typologies of nature and causes have been built upon and refined by many linguists until today. Theoreticians looking for semantic rules are still concerned with the causes, consequences and the functioning of the ever renewing signified and signifier elements of the Saussurian sign, however Saussurian theory has been largely questioned and refined, since the Saussurian framework showed weaknesses regarding the treatment of diachrony and its relationship to synchrony, as will be detailed thereafter.

The inventory inherited from typological works was completed by a reflection about the role of the individual in society, through the works of Meillet (1906) who analyzed semantic change as the result of social variation. This perspective, strongly influenced by the works of Durkheim (1907), nourished later sociolinguistic approaches.

In the XXth century, historical pragmatics, discourse analysis and cognitive sciences completed and modified these approaches, in particular through the works of Geeraerts (1997; 1983), Traugott and Dasher (2002) and the ones gathered by Blank and Koch (1999). The focus shifted to issues of regularity and continuity in semantic change in a cross-linguistic perspective, as well as directionality and conceptual mapping, giving central importance to metaphorical and metonymical processes as well as polysemy.

In the first section of this chapter, I try to lay the theoretical foundations and define the key questions of semantic change and provide an overview of the frameworks that address them.

[13] "Language changes to keep operating as it is."

The second section gives an overview of typologies classifying factors internal to language. It is followed by a third section reviewing causal approaches rooted in sociology and psychology. The fourth section is devoted to approaches that try to bridge internal and external factors, with the contributions of cognitive linguistics, in particular diachronic prototype theory and the larger spectrum of metaphor theory, as well as historical pragmatics and discourse analysis. The final section discusses a suitable theoretical framework for the purpose of exploratory corpus-based analysis.

1.1.1. Theoretical framework

There are several theoretical frameworks that look at semantic change. First, the term *semantic change* is generally associated with Anglo-Saxon approaches, while most European approaches prefer the term *semantic neology*. However, semantic neology, or how known words acquire or lose some of their meaning, is part of semantic change which encompasses other types of phenomena. Dealing with semantic change in the framework of *neology* involves the associated theoretical framework for analysis. Semantic change also encompasses conceptual change or how new or obsolete concepts are attached to a known or new word.

Semantic neology is one of the phenomena that cannot yet be studied in an agreed framework of semi-automatic treatment. It is the most complex to grasp among types of neologies. Trying to grasp the challenge and bring elements of answer cannot be done without first studying the theoretical frameworks in linguistics that deal with this problem, whether they consider semantic change as a neology or not. These frameworks often encompass other related phenomena: grammatical neology, formal neology (new words) also referred to as *lexical entrenchment* (Chesley 2011a) including borrowings across languages or across linguistic communities, rhetorical processes by which change takes place, grammatical, syntactical, structural and combinatorial processes, as well as psychological and social processes. The type of chosen framework induces not only specific terminologies but also specific types of classifications, whether they take the form of theoretical models or typologies. The latter show some degree of similarity across approaches as well as some degree of divergence. This largely depends on the chosen perspective of analysis.

We may focus on language itself as a system (organic, structural, mathematical), or on the interaction of language and people (psychology, psycholinguistics, pragmatics) and/or society (sociology, sociolinguistics, socio-pragmatics). In the first case, mechanisms internal to language are at stake, in the second, external causes are, as well as interactions. From my point of view semantic change analysis does not benefit from a radical distinction between language and interaction, or *langue* and *parole* ("language" and "speech") in Saussurian terminology, since change, by essence, is rooted in both. However, this work does not delve into speaker interaction since the chosen approach is textual.

At the linguistic level, there are two distinct questions: How does language change? and, what are the reasons for change? They lead to the analysis of mechanisms and causes of linguistic change and semantic change. I shall leave aside linguistic change, as it encompasses grammatical, syntactical, and combinatorial issues to a large extent, to focus specifically on semantic change *per se*, which is more concerned with meaning. Therefore, this work looks at word meaning change and meaning change at the conceptual level. They are also known as semantic *shifts* or semantic *drifts*, to refer to subtle changes at the level of connotation. In the terminology of neology, they are referred to as meaning or sense neologies (fr. "néologies de sens").

At the semantic level, we may ask "Comment les mots changent de sens[14]?"(Littré 1888; Meillet 1906), what is the "development of meanings associated with a form"? (Nerlich and Clarke, 1999: 181) as semasiology does, or we may ask "how do concept repartition across words change?", as onomasiology does or "how do people attribute new meanings to words?" as socio-pragmatics does. These distinctions rely on whether one chooses to ground the question in word dynamics (terminology, semasiology) or in concept dynamics (onomasiology, cognitive linguistics and psycholinguistics). However these approaches are not completely incompatible, and may be combined. Notably, onomasiology and semasiology are often combined in cognitive linguistics:

"Combining the onomasiological approach with a well-founded semasiological typology of diachronic semantic processes will enable us to understand, in a sort of

[14] "How do words change meaning?"

"panchronic" perspective, the basic cognitive patterns of how man conceives the world." (Blank and Koch 1999: 11)

Word dynamics include all the semantic processes of word formation, such as morphological productivity, whereas concept dynamics include all the mechanics of conceptual evolution and redistribution. Both views are then confronted with the issue of *diffusion*, or how a new word or meaning is relayed in communities of speakers until it becomes part of language, either at the level of unspoken consensus or at the level of lexicology.

1.1.1.1. Reanalysis and speaker innovation

Do words change meanings intrinsically? For many authors, they do not do so by essence, but rather through speaker interaction:

"words don't change their meanings, speakers just use them differently in similar and or different contexts according to various communicative needs" (Nerlich & Clarke 1988: 76).

The idea that linguistic change is speaker-based lays the foundations for the concept of *reanalysis*:

"If one deduces a different underlying form or rule for producing something that a speaker or the speakers round about are producing, then one has made a reanalysis." (Forston, 2005: 650)

Thus, new words and meanings may emerge from individual innovation first but their existence is ultimately rooted in a community, for diffusion matters. Indeed, new meanings may be said to come to existence via *language use*, as stated by Traugott & Dasher (2002). Speaker-based semantic change is often referred to as an *innovation*, or a *speaker innovation*, and becomes a *change*, or *linguistic change*, when it has been adopted by other members of a community. This distinction is discussed by historical linguists such as Milroy (1992) and Shapiro (1991). For the latter, change is a *social fact*. This rooting in individual innovation also implies that a lot of changes emerge from colloquial language before being absorbed into more formal language. This reinforces the discrepancy between spoken language and written language, and the distinction between the study of recorded spoken data and textual data,

since the written medium is generally more formal. A form like fr. *ouais* (which may be translated as "yep"" or "yeah") for instance, instead of *oui* ("yes"), is still considered informal, but had been employed in spoken language for decades, before it entered realist literature that aimed at reproducing street language.

Works focusing on reanalysis deal with how the processes of change unfold and whether they show some kind of systematicity, rather than classifying mechanisms.

1.1.1.2. Productivity and diffusion issues

How are new words created? Can we predict new words? At what point in diffusion do we consider that a new word has entered language?

Meaning change at the lexical level relies on both semantic motivation (dynamics of polysemy, semantic shifts, and semantic association networks) and morphological motivation (possible derivational and compositional patterns). The probabilistic study of the set of possible morphological derivations and compositions, or a words' *productivity* has been largely studied, for instance by Baayen (1991). This approach does not look at the diffusion process following the possible creation allowed by morphological productivity. Moreover, a series of semantic change cases cannot be captured by quantitative studies of morphological productivity. Diffusion issues are one step after productivity issues, looking at the *conventionalization* of new meanings and words at the social level, or *lexical entrenchment*:

> "Lexical entrenchment refers to the process by which a new word becomes part of the lexicon, or vocabulary, of a language. In other words, given that a new word occurs, how likely is it that the word will become part of the lexical stock of the language, as opposed to a fleeting lexical item? One central question for those studying lexical entrenchment is how lexical innovations are diffused throughout a speech community. Another question is the role individual speakers' memories play in lexical entrenchment. Lexical entrenchment differs from morphological productivity, sometimes referred to simply as productivity, because the study of the latter seeks to generate, or predict, possible words, either with rule-based approaches (see, for example, Aronoff 1976, Selkirk 1982, Halle and Marantz 1993, and Ussishkin 2005) or by assigning a probability of productivity to a lexical item or lexical process (Baayen 1992). For example, productivity deals with whether or not a form like *hateable* is possible, or how likely we are to see this form. Entrenchment deals with how likely it is that this word will recur, and how often it will recur, given that we have already seen the word *hateable*. In this sense, lexical entrenchment takes up where productivity leaves off in the study of new words." (Chesley 2011a)

Chesley's (2011a) doctoral thesis looks at lexical entrenchment in terms of diffusion through linguistic communities and includes a psychological aspect. She reaches beyond morphological productivity and predictability issues at the morphological level to tackle the next stage of probability in the diffusion process. For Forston (2005) as well, there should be a distinction between the study of reanalysis and its diffusion, which is a separate sociolinguistic issue.

These approaches look at the morphological formation of new words and at their diffusion; however they do not explore in detail the mechanisms of meaning change.

1.1.1.3. Polysemy and meaning saliency

Indeed, a fundamental question of semantic change is: What are the mechanisms of word meaning change?

One of the central mechanisms underlying semantic change is the dynamics of polysemy. The way meanings are organized in a word's polysemy is referred to as its *internal structure*. Most words have several meanings and each meaning/form couple is referred to as a *lexical unit*. These lexical units each possess a degree of saliency, and may be connected to the other lexical units that compose the whole meaning by semantic relations or not, as they may differ referentially. In case of referential difference, the question of whether these are two homonymous words rather than one word with several lexical units may be raised. The change in saliency of lexical units is the most accessible tool to assess semantic change. Ontologies, such as Wordnet (Fellbaum et.al 1998), list "lexical concepts" in hierarchical order, and in a relational framework; however, all ontologies to date require manual updating if the saliency order has substantially changed, integrated or lost lexical units.

In the dynamic (re-) organization of lexical units, shifts may take place within the same semantic field or across several semantic fields, reflecting intrafield, or interfield semantic shifts. For instance, the fact that the word *mouse* came to refer to the technological device we use with a computer additionally to the reference to the animal, is an interfield change. But the fact that *dog* changed from referring to all kinds of animals to a single species is an intrafield change.

The aforementioned questions are dealt with in several frameworks that use different tools to answer them.

1.1.1.4. Different frameworks, different questions

The two most central questions seem to be:

1) What are the mechanisms of word meaning change?

2) How are new words created?

In **lexical semantics**, the traditional distinction between **semasiology** and **onomasiology** aims at answering separately the two following questions:

"How do words change meaning?"

And "how do concepts repartition across words change?"

Thus, words and concepts are distinguished as separate entities. Two paths of analysis are considered, from word to concept or from concept to word.

Lexical semantics is akin to **lexicology**, however lexicology embraces all aspects of vocabulary whereas lexical semantics is more particularly focused on its content, and therefore on meaning.

Terminology is an area of applied lexicology. It tries to unveil relationship between specialized and non-specialized language. Terminology goes as far as asking: "At what point in diffusion do we consider that a new word has entered language?" This question of conventionalization includes aspects of diffusion. **Lexicography** questions conventionalization; it comes one step after lexicology and can be defined as lexicology applied to dictionary issues. Lexicology studies the morphological mechanisms of word formation to answer the question of how new words are created. **Quantitative lexicology** does the same thing with a quantitative morphological approach and furthermore asks: Can we predict new words?

Structural semantics (e.g. Hjelmslev 1959, Greimas 1965) and **interpretative semantics** (Rastier 1987) try to answer those questions with a componential approach (word meaning is

understood in terms of sets of features, referred to as "sèmes" "components" or "markers" depending on the chosen terminology). Structural semantics also includes morphology and mechanisms based on rhetorical figures. Structural semantics largely focus on synchrony, and notions of signifier and signified in the continuation of De Saussure (1916), at the exception of Coseriu's (1964) diachronic structural semantics, and authors following his view. The centrality of synchrony in structural semantics led to a large lack of contributions to diachrony; however we are witnessing a revival of diachrony over the past decades, notably due to the influence of historical linguistics.

Structural semantics also relies on semantic fields (or domains) by grouping items which share similar features into sets. The idea of semantic domain is shared by cognitive semantics, except that their definition does not rely on componential analysis. The main point of disagreement between structural semantics and cognitive semantics is the relationship of words with concepts. In the structuralist approach, "the meaning of a word becomes nothing other than the set of relations that the word contracts with other lexical items" (Taylor 1999: 38). Contrarily, cognitive semantics and cognitive grammar (see: Langacker 1987), as well as related theories such as Jackendoff's (1985), defend conceptualist semantics, in which concepts are intrinsically encoded in language.

"European structural semantics has pleaded for a strict theoretical separation of *encyclopedic knowledge* from *language-specific semantic features* and has determined the latter to be the only object of linguistic semantics. In contrast to this, cognitive linguistics has strongly emphasized the importance of encyclopedic knowledge for semantics" (Blank and Koch 1999:4).

Cognitive semantics tries to answer the questions surrounding semantic change in terms of conceptual mapping between items and prototypical organisation, and in terms of speaker interaction when combined with **pragmatics** and **discourse analysis**. These two disciplines give central preponderance to speaker interaction and socio-cultural constraints while structural semantics focuses on language-internal mechanisms in quasi isolation from interaction.

At the heart of cognitive semantics, is **prototype theory** (Rosch 1973**;** 1978) , which dissects meanings into conceptual categories. **Diachronic prototype semantics** (Geeraerts 1997)

analyses semantic change in this framework. Rastier (2001; 1987) disagrees with the mentalist conception underlying these theories, and with the idea that the content of a word is directly associated with the corresponding concept.

All these branches of linguistic semantics are initially anchored in synchronic theories and benefit from the contributions of **historical semantics** when they focus on diachronic issues. Diachronic lexical semantics, diachronic structural semantics, and diachronic cognitive semantics therefore encompass historical semantics issues. Another emerging branch of historical linguistics is computational historical linguistics (see Allan and Robinson 2011) as will be detailed in Chapter 1.3. Historical semantics is an area of historical linguistics, of which one of the prime concerns has been the origin(s) of language. This gave birth to cross-linguistic research in diachrony, trying to extract some degree of universality in language mechanisms. Most of these works rely on *cognates*, or words that have the same etymological root across languages from the same family. The search for cognates across languages is itself a subfield of semantic change in cognitive linguistics (see for instance Wilkins 1996). It therefore includes theoretical tools from comparative linguistics. This is not the approach chosen here, even if several languages may be looked at. Instead, I study whether a model that gives good results in a language may be applied to another language from the same family.

Historical linguistics is not only a branch of linguistics but also one of history. Intrinsically, it includes social considerations, and studies the connections between language and real-world events in history. Indeed, semantic change and language are deeply intertwined with history, both witnessing and making it. In this spirit, Coseriu (1958:12) posits:

"Una lengua, en el sentido corriente del término (lengua española, lengua francesa, etc.) es por su naturaleza un 'objeto histórico.'"[15]

Historical linguistics looks at language evolution across centuries and across languages. Long diachrony issues are treated and include language change. Depending on the theoretical point of view, some studies in historical linguistics shed light on the similarity of processes across languages (universalism), while others tend to highlight the differences (relativity). Historical

[15] "A language, in the current sense of the term (Spanish language, French language, etc.) is by nature a "historical object""

linguistics relying on corpora is faced with orthographic changes and dialectal variations and only has at its disposal a series of incomplete records. Historical linguistics includes a sociological dimension in its framework while capturing language as a historical object. This discipline is deeply anchored in the history of vernaculars and their relationship with mainstream language. And, by extension in granularity, it deals with idiosyncratic innovation, and therefore reanalysis.

1.1.1.5. Past Saussurian theory: short diachrony, long diachrony and the nature of meaning

The study of semantic change posed a certain number of questions to the field of structural linguistics at large. In fact, its original paradigm, the now classic theory of Saussure (1916), did not provide a satisfactory framework for diachrony. I agree with Coseriu (1958, esp. chapt VII, 135-161) who states that the study of semantic change cannot be carried out in terms of the Saussurian categories of "language" and "speech", and "synchrony" and "diachrony" as he defined them :

"En los últimos tiempos se ha señalado a menudo la necesidad de reducir la rigidez de las dicotomías saussureanas. Se ha dicho, con razón, que hay que colmar el abismo excavado por de Saussure entre *langue* y *parole*. Y, por lo que concierne a la "lengua", se ha insistido en la necesidad de colmar el abismo entre *sincronía* y *diacronía*.[16]"(Coseriu 1958: 9)

For De Saussure (1916) change is exterior to the system. Change happens in "parole" ("speech") and not in "langue" ("language") and therefore do not affect the system (of "langue") that is immutable. Indeed, for Saussure, "langue" as a system is synchronic by nature, and diachrony only affects "parole". The changes are isolated phenomena but the structural relations stay. However, Coseriu argues that this is going against Saussure's own theory in which elements exist through a contrasting relationship ("un juego de opposiciones"). Saussure admits that diachronic changes affect words, but not language in its

[16] "Lately, the necessity of reducing the rigidity of Saussurian dichotomies has been progressively signaled. It has been rightly said, that the abyss created by De Saussure between langue and parole has to be bridged. And, with regards to "language", the necessity of bridging the abyss between synchrony and diachrony has been insisted upon."

synchronic state as a system. Yet, if the elements of the system are modified by an external cause, the entire system is in turn modified, following Saussure's principle.

This theoretical weakness is what leads linguists, and especially cognitive linguists, to question the very distinction between "langue" and "parole" ("language and speech"), and between synchrony and diachrony. Indeed the system must be modified in diachrony as there cannot be changes in language that solely affect one of its theoretical states ("langue" or "parole"). However, cognitive linguistics is far from refuting all of Saussurian theory, as it develops some of its aspects. It considers that the sign is arbitrary in that it is the product of conventionalization. Moreover, "Cognitive Grammar is strongly committed to the symbolic nature of language, and in this respect is profoundly Saussurian in spirit." (Taylor 1999 : 19)

The idea that the system's structure is immutable clashes with the one of the dynamic nature of language, and creates a theoretical paradox, or "tension between norm and evolution" as Rémi-Giraud (2000:37) terms it:

"Le changement sémantique soulève en effet une énigme fondamentale qui tient à la contradiction apparente entre la dimension normative des significations et leur dimension évolutive[17]" (Rémi-Giraud 2000:37).

Contrary to Saussure's focus on synchrony and stability of the system, most of XIXth century philology was strongly focused on diachrony. Paul (1880), for instance, stated that the synchronic view alone was a reduction of linguistics.

"Linguistics in the nineteenth century was [hence] tightly interwoven with the humanities. The study of a language was always about the language as a historical entity, as an object that could be understood against the background of its development in time. What is true for the discipline as a whole holds true for its subparts as well: the investigation of semantics took its origin (and even the term *semantics* itself was coined) in the investigation of meaning change" (Eckardt, Von Heusinger, and Schwarze 2003).

[17] "Semantic change does indeed bring out a fundamental enigma that lies in the apparent contradiction between the normative dimension of senses and their evolutionary dimension."

What happened between the times in which "panchrony" was studied in a holistic philological fashion and the times in which synchrony and diachrony have come to be perceived as two different theoretical branches of linguistics?

"Il fut un temps, il y a bien longtemps de cela…, bien avant la naissance des différents courants de linguistique générale et structuraliste qu'a connu le XX^{e} siècle avec Ferdinand de Saussure, Gustave Guillaume, Chomsky, etc., l'étude d'une langue se concevait de manière globale : on ne distinguait pas l'aspect synchronique et diachronique des langues. […] Des micro-conflits internes, mais aux dimensions scientifiques existentielles, ont naturellement conduit à opposer des micro-domaines à d'autres micro-domaines. C'est dans ce contexte que les Écoles opposant la linguistique diachronique et la linguistique synchronique sont nées. [18] " (Guillaume 2010: 2-3)

The initial "panchronic" approach seemed philosophically more adequate for language as an object of study. Philology subsequently focused mainly on long diachrony issues. However, we now face further dichotomies within diachrony, as short diachrony is often considered separately from long diachrony. Short diachrony is oriented toward real time observation and therefore toward directly applicable tools. It is in tune with the society producing a new language embedded in a new paradigm of diffusion. For Picton (2009), short diachrony typologies differ from long diachrony ones, and she adds that:

"…on peut dire que l'évolution relève rarement de la néologie (ou de l'obsolescence de concepts) en diachronie courte [19]" (Picton, 2009 : 263).

In a way, the rare neologies that may be detected in short diachrony are remarkable in terms of how quickly they appear and develop and in terms of their conceptual richness and strength. If their number seems poor, pure neologies are also few across long diachrony

[18] There was a time, long ago... long before the birth of diverse currents of general and structuralist linguistics witnessed in the twentieth century, with Ferdinand de Saussure, Gustave Guillaume, Chomsky, etc.., when the study of language was conceived holistically: synchronic and diachronic aspects of language were not distinguished. [...] Internal micro- conflicts, with existential scientific dimensions have naturally led to oppose micro-domains to other micro-domains. It is in this context that the Schools opposing diachronic linguistics to synchronic linguistics were born.

[19] "… we can say that evolution in short diachrony is rarely of the type of neology (or of the obsolescence of concepts)."

corpora and there is no consistent statistical study yet to show that their proportion is similar or different with respect to the time span corpora cover.

It is surprising that the study of neology is rarely taken out of the framework of long vs. short diachrony. The question of where the time scale limit is set between those two receives a largely vague approximation. Nevertheless, in the literature dealing with *semantic change* rather than *neology*, this distinction is challenged. Nerlich & Clarke (1988) have challenged it, at two levels. First, they question the suitability of opposing diachrony and synchrony to study change, arguing that "the nature of meaning is change" (Nerlich & Clarke 1988:73) and that

"the explanation of meaning should be based on the fact that meaning is dynamic and ever-changing, "something" of which change seems to be an inherent property. That is, one can only explain what meaning is, by explaining how it changes". (*ibid.)*

This argument is strong since it places the study of semantic change at the heart of the study of meaning. Second, instead of viewing short and long diachrony, they consider that semantic change takes place in two temporal and causal phases: micro- and macro- dynamics:

"Micro-dynamics or short term semantic change is related to the actual speech-event, macro-dynamic or long term semantic change to its long-term consequences." (*ibid.)*

Micro-dynamics refer to the level of the idiosyncratic innovation, or creativity, within a speech-event whereas macro-dynamics refer to shape and structure factors in language. Coseriu (1958) also makes this distinction. This theoretical posture sets the framework to formulate a hypothesis to delineate the dynamics of language evolution:

"Our hypothesis is that there exists two fundamental movements in the evolution of language, one toward *arbitrariness* (related to macro-dynamics), the other toward *remotivation* (related to micro-dynamics). Their interaction structures long-term semantic change." (Nerlich & Clarke 1988: 80)

Now that the major differences between the theoretical frameworks have been outlaid, I am going to look into more detail at the contributions offered by each theoretical school to answer the question of semantic change.

To start, the following section describes the contributions of structural semantics, including lexicological and terminological approaches, and focuses specifically on typologies of semantic change established in those perspectives.

1.1.2. Typologies of mechanisms internal to language

Typology aims at classifying linguistic patterns, and tries to unveil some degree of systematicity in these patterns, either cross-linguistically or within one given language. In this section, I deal with typology in the sense of a structural classification, and not in terms of language typology. (The latter establishes relationships between different languages and organizes them into families.)

It is useful to distinguish lexical typology from semantic typology. Lexical typology is the "cross linguistic and typological dimension of lexicology" according to Koptjevskaja-Tamm (2008: 5). It is concerned with how the lexicon absorbs semantic material into words and how this material is organized. Lexical typology includes phonological, morphological, syntactical and cross-linguistic considerations, while semantic typology is more restricted to the study of meaning patterns (Robert 2008).

The literature abounds with different types of classifications. I have chosen to gather approaches that study factors internal to language (sometimes called the "nature" of semantic change) in this section, as opposed to approaches that study external factors (either called "causes" or "motivations"). Some of the listed typologies of language internal factors also include studies of external factors, but very rarely bridge internal and external factors dynamically. The contributions in terms of external factors are dealt with separately in section 1.1.4. Choice of focus and theoretical divergences bring about four types of language–internal typologies:

-Typologies of morphological productivity, focusing on the rules and mechanisms of word formation

-Typologies of neologisms, in which I focus more particularly on formal and semantic neology (creation of new words and meaning shifts) rather than grammatical ones. These typologies strongly focus on rhetorical figures governing the processes by which word meaning changes.

-Those two types can combine, giving a broader approach including morphology as well as rhetorical mechanisms.

-Classifications of semantic change mechanisms.

Typologies of neology encompass more that semantic change issues since they include work on morphology at a larger scale and include grammatical aspects (notably *grammaticalisation,* also referred to as *grammaticization,* the mechanism by which a content word becomes a function word) and syntactical, cross-linguistic, as well as wider lexicological considerations which are not directly related to the object of this study. Works on grammaticalisation support the idea of *directionality*, since lexical items become grammatical ones but not the other way round. I shall only briefly mention those phenomena and focus on elements of these theories that are directly useful to the purpose of this work, that is to say theoretical aspects that provide insight as to lexical semantics meaning shifts. In fact, a great amount of research on semantic change focuses on the functional lexicon (function words) and I focus more specifically on the substantial lexicon (either called "plain" or "content" words).

1.1.2.1. Typologies of morphological productivity

There is no total agreement about word formation, or how new words are coined in terms of meaning-form association relying on pre-existing words. However, there are a series of processes that most linguists agree on, as is the case for most mechanisms described by Bloomfield (1933).

In this area again, the distinction between synchrony and diachrony may be questioned, since what is at stake is a process unfolding in time. Fernández-Domínguez (2009) analyses productivity in English word formation, relying on Marchand (1966) who has a "synchronic-diachronic" view, otherwise called "panchronic" by authors such as Tournier (1985) and Rastier (1999). This view, inclusive of both synchrony and diachrony in a common paradigm seems satisfactory since morphological processes of word formation are similar in synchrony and diachrony. The distinction does not seem to bring further insights to this field of study. It therefore seems unproductive to forcefully apply it to the phenomena for theoretical purposes.

1.1.2.1.1. *Major mechanisms of morphological productivity*

There are two major morphological mechanisms involved in the creation of new words (or *lexical innovation*): derivation of an existing word (by affixation of bound morphemes: prefixation and suffixation) and compounding (by association of existing words or parts of them).

Derivations are built on a pre-existing word, with the adjunction of an affix, which can be either a prefix or a suffix. For instance *happiness* is derived by suffixation in *–ness* from *happy*, and *unhappy* is derived by prefixation in *–un* from *happy*. Derivations may also be applied to proper nouns. Examples are often political, as in French *Sarkozyste*, to say pro-Sarkozy.

Compounding combines two or more pre-existing words. *Blackboard* is an instance of primary compounding (two nouns combined), whereas *peacemaker* is an instance of synthetic compounding (a noun and a verbal element combined).

Other important processes are conversion, back formation and blending.

Conversion is a shift from a lexical category (part-of-speech) to another, for instance when a noun or adjective produces a verb (as in *an email* → *to email*).

Back-formation is the process by which a word looses a morphological element. This intuitively goes against the more traditional process by which morphological derivations add a free morpheme to a base. Examples of back-formation raise a series of disagreements. (Sablayrolles 2000: 219) mentions that for some linguists, the chronology does not seem to matter as much as length does, and therefore the phenomenon is analyzed backwards. Backformation calques its process on other known processes, for instance when creating a verb from a noun ending in –ion in English, such as the verb *acculturate* being created on the basis of the noun *acculturation*. This process imitates the opposite usual path, in which verbs give birth to nouns by adjunction of the suffix –ion as in *insert* → *insertion*.

Blending refers to the creation of a new word composed of parts of pre-existing words. A famous example is the word *brunch*, based on the br- of *breakfast*, and –unch of *lunch*. The press is a goldmine of innovative blending processes, for instance *Le Monde* uses forms such as *hacktiviste*, based on *hacker* and *activiste* (“activist”), or *beurgeoisie* based on *beur* (a

slang word to refer to second generation North-African immigration in France) and *bourgeoisie*[20]. These innovations mostly have a stylistic purpose and may be adopted or not by the readers. Journalists greatly participate to the coining of such blends.

1.1.2.1.2. *Factors and constraints in morphological productivity*

Morphological productivity studies most importantly include processes of word formation, but also other factors that impact productivity such as enhancing or constraining factors, as well as issues of availability and profitability, gradation within productivity, and ultimately, measurement. Their scope is the coining of form-meaning correspondences by users via a morphological process. Possibilities of coining are theoretically infinite. Coining depends on user creativity, but the latter has to abide by the rules and constraints of productivity. Therefore productivity may be seen as a set of possibilities of word formation while creativity is the deliberate action of coining a new meaning-form available in this set. However this view is largely contested, since some authors use the terms "productivity" and "creativity" synonymously while others distinguish between "rule governed coinages [and] …items typical of literary language or of the press" (Fernández-Domínguez 2009: 56)

Corbin introduces the notions of availability and profitability (fr. "disponibilité et rentabilité") that regulate the possibility of coining and its "economic" value in the language system. For Corbin (1987: 17),

"La productivité désigne en fait à la fois la régularité des produits de la règle […] la disponibilité de l'affixe, c'est-à-dire précisément la possibilité de construire les dérivés non attestés, de combler les lacunes de lexique attesté, et la rentabilité, c'est-à-dire la possibilité de s'appliquer à un grand nombre de bases et/ou de produire un grand nombre de dérivés attestés"[21]

[20] In the newpapers *Le Monde*, there are 13 occurrences of the word *beurgeoisie* between 1997 and May 2012, and 6 occurrences of *hacktiviste* between 1999 and may 2012. Source: lemonde.fr

[21] "Productivity indicates in fact at the same time the availability of the affix, i.e. precisely the possibility to build non-attested derivatives to fill the gaps of the attested lexicon, and profitability, i.e the possibility to apply to a great number of bases and/or to produce a great number of attested derivatives."(translation by Fernández-Domínguez 2009: 59-60)

These factors tend to condition coining, while other types of factors may facilitate or constrain word formation. Factors facilitating word formation are: cognitive simplicity, easy access (in terms of memory), fitting with known (universal) patterns, intuitiveness (how "natural" the new item is) and semantic coherence (clear meaning association between root word and target word), also referred to as transparency. There is a gradation in transparency as some words may seem to be of obscure composition when the process of formation is not visible any more while appreciating the form. For historical reasons, some words may keep a relatively high transparency over time while others do not. For instance, *unhappiness* is transparent, while *gospel* is opaque[22] (*gospel* comes from the OE. compound *gōdspel* "good spell" and not from "God's spell", a folk etymology misinterpretation[23]).

Conversely, there are factors which constrain productivity either cross-linguistically or within a given language. Within so-called "universal" (cross-linguistic) constraints Fernández-Domínguez (2009: Chapt 3) lists blocking, complexity base ordering and choice of the base. For English, the author lists complex structural constraints (at the level of phonology, morphology, syntax and semantics) which affect the form of neologisms, as well as pragmatic and argument constraints, which affect their use. These constraints are themselves classified differently in various typologies of constraints, however the main mechanisms described below are acknowledged by several recognized authors in the field.

Blocking (also referred to as *pre-emption*) is the process by which a potential new word will not occur if there exists an exact synonym or homonym in the given language. Fernández-Domínguez (2009:73) cites the example given by Plag (1999:50) of the item **liver*, potentially derivable from *to live*, being blocked by the homonym *liver*, the organ.

However, the history of language contacts has shown the weakness of this rule, as regards synonyms, since it does not seem to have applied to borrowings. In English, for instance, numerous words have synonymic equivalents of Germanic and Romance roots, following the Norman conquest of England. It is considered today that up to 30 percent of the English vocabulary is of French origin. However, in most cases both Germanic and Romance

[22] The example of gospel is given by Fernández-Domínguez (2009 : 68) citing Dressler (2005:271).

[23] Source : Steinmetz (2008: 210-211)

synonyms have survived, specializing only in terms of register, as items of Romance origin are generally perceived as more formal or refined, while items of Germanic origin are associated with common language. Examples are numerous, such as *birthday/anniversary*, *give up/renounce*, *brotherly/fraternal*, *job/profession*, or *quick/rapid* (the first of the pair being of Germanic origin, the second of Romance origin).

Moreover while dealing with near synonyms, the items may specialize over time and coexist with related but ultimately different meanings. (For instance the borrowings *geek* and *nerd* in French were quasi synonymous, until a positive connotation became attached to the first while the second gained a negative connotation attached to certain social behaviours. Finally, *geek* gained in frequency of use compared to its competitor.)

The second constraint, complexity base ordering, refers to affix restrictions in terms of word class (see Aronoff 1976 for the Unitary Base Hypothesis in generative grammar) and choice of the base refers to how frequent the base is (the more frequent, the more productive), as well as its length.

Additionally, complex phonological restrictions apply in morphology, to avoid difficulties of pronunciation, for instance the suffix *–ity* cannot be added to a word ending in *–ous*. As for syntactic constraints, an example is that the suffix *–able* can only be added to transitive verbs. This rule also applies to French, for instance, we say "*de jour en jour*" ("from day to day") but not "**d'an en an*" which poses pronunciation problems and ultimately results in cacophony; therefore it has been replaced by "*d'année en année*" even if this goes against the division between *jour/journée*, *an/année*, in which the first item of the pair is purely chronological while the second refers to time unfolding (example given by Lüdtke 1999 : 55-56).

Semantic constraints may be the most abstract of all structural constraints, which may explain why they are often described together with other structural constraints, or pragmatic ones.

All the aforementioned rules, like all rules in linguistics, are subject to exceptions. Moreover, they often do not cover the spectrums of language contact issues and borrowings. Borrowings, by nature, may require unusual modifications to bear inflections in a target language. For instance, when Swiss, Canadian and Belgian French borrow the verb *to park* from English,

and create the verb *se parquer*, they calque its reflexivity onto *se garer* while *to park* is not reflexive in the first place. Moreover, there is no verb in French ending in *–ker*, and again a calque takes place, based on the French endings in *–quer* (as in *bloquer*, to block, for instance) such that the form resembles other known forms. Additionally, this is primarily an exact synonym of *se garer* and this goes against the aforementioned blocking rule. Speakers of Canadian, Swiss and Belgian French, however, tend to use *se garer* to refer to the maneuver while they use *se parquer* to refer to the finished state of the maneuver, therefore introducing difference in terms of process and state. Process-state semantic differences are numerous in English but few in French. The calque here goes from English to French. They also tend to use the *se garer* when they do so in a *garage*, by opposition to *se parquer* in a *parking* (French borrowing to refer to *parking lots*, otherwise named *car parks*). However, these semantic shifts do not seem to be attested by authoritative lexicographic references yet. The TLFI gives this information:

"Empl. pronom. réfl., fam., rare. Mettre sa voiture en stationnement. Synon. *se garer*. (Dict.xx^{e} s.).[24]"

It does not, however, tackle the subtle semantic shifts taking place. Numerous examples of this kind show that there is a lot of room for creativity that morphological productivity may only encode as new processes appear. It is striking that with the growing use of the Internet and a generally enhanced state of language contacts, these rules tend to be overlooked by creative freedom. It must also be noted that English has more loose rules of idiosyncratic creativity than most European languages, especially French, which is conversely, much more restrictive and conservative. One may wonder whether languages are dependent upon change-driven societies that encourage idiosyncratic creativity or conservative ones that encourage the preservation of language from external influences. Moreover, most of purely morphological rule-based understandings (apart from some that include semantic-pragmatic or cognitive levels) tend to simply ignore external factors. These factors may be of sociological and historical nature, for instance trends and politically imposed language

[24] *Rare and familiar reflexive pronominal use*. To put one's car in parking. Synonymous of *se garer*. http://atilf.atilf.fr/dendien/scripts/tlfiv5/affart.exe?19;s=3826989510;?b=0; last accessed on the 16/04/2012

contacts or restrictions having to do with "politically correct" language attitudes. In my opinion, those factors are extremely powerful and simply cannot be ignored.

1.1.2.2. Typologies of rhetorical figures

The groundbreaking work of Bréal (1899) is one of the earliest typological approaches of semantic change, that has been followed and expanded on by most research in the field. It encompasses lexical change but also morphological typology, as well as an approach of *motivations* (causes). However, the work is largely centered on language-internal processes. The author lists two major types of onomasiological changes: specialization (fr. "restriction du sens", or restriction/narrowing depending on the terminologies) and differentiation ("élargissement du sens", or extension/broadening). Specialization is when one word becomes the strongest representative of a meaning, differentiation is when two synonyms or near synonyms diverge and specialize.

Bréal's examples of specialization mostly involve long term evolutions from Latin or Greek to French. For example, Lat. *fenum*, meaning "product", results in fr. *foin* ("hay") (Bréal 1899: 121).

The term specialization is used by later authors to depict evolutions in shorter spans. For instance, the word *hound* has specialized over time. It used to refer to any type of dog and specialized to refer to hunting dogs specifically. However, the lexical unit referring to any dog is still considered archaic, as the OED[25] attests:

"1. A dog, generally. (Now only *arch.* or *poetic*.)

2. a. *spec.* A dog kept or used for the chase, usually one hunting by scent. Now esp. applied to a foxhound; also to a harrier; (the) hounds, a pack of foxhounds."

Differentiation, for Bréal, is the opposite of specialization: it is the result of historical events. For instance, Fr. *gain,* (*gaïn* in the original spelling) used to refer to the harvest, and widened to mean all types of earnings obtained from work (Bréal 1899:129).

[25] Oxford English Dictionary, accessed online 09/ 05/2012

For semasiological change, the major mechanisms the author describes are pejoration/amelioration, restriction/expansion, and metaphor/metonymy. He also underlines the importance of polysemy, which he describes as a sign that a culture is advanced (see Bréal 1899, chapter IV).

Across the many typologies in the field, the ones put forward by Bloomfield (1933) and Blank and Koch (1999b) seem to stand out. They also expand on Bréal's (1899) classification. According to Forston's synthesis (2005), the central rhetorical figures that are considered the engines of semantic changes across various theories are metaphoric extension, metonymic extension, broadening, narrowing, melioration and pejoration. I have grouped them with other terms that refer to the same mechanisms for clarity:

1. Metaphoric extension, or metaphor. (A word gets a new referent that has common characteristics or shows similarity with the previous referents. Mapping of one concept onto another.) e.g. *Head of a community*
2. Metonymic extension, or metonymy, also referred to as contagion. (A word gets a new referent in contiguity with the previous referent.) e.g. *White house*. This often implies a cultural dimension.
3. Broadening, also referred to as widening, generalization or expansion (a word goes up from a level of subordination to one of superordination) e.g. *Dog* (a specific breed → all dogs)
4. Narrowing also referred to as specialization or restriction (a word goes down from a level of superordination to one of subordination) e.g. *deer* (wild animal→deer)
5. Melioration & pejoration. (a word gains positive or negative connotation) e.g. *nice:* "ignorant, foolish"→"nice, pleasant"/ *silly* "happy, blessed, blissful" → "stupid").

In English, one of the most accepted and widespread typologies is the one established by Bloomfield (1933). In addition to the five figures listed above, he includes synecdoche, or change anchored in a whole-part relation, such as the use of *Paris* to refer to the French government. Semantic change via part-whole transfer of senses, or synecdoche has been acknowledged since Reisig (1839) and Darmesteter (1887) even if the latter does not use the term "synecdoche," preferring to deal with it under the headings "narrowing" and "widening." Synecdoche is a type of metonymy; however, the exact relationship between synecdoche and metonymy has been thoroughly discussed theoretically. Synecdoche may apply at the

taxonomical level, in a genus-species relationship, or more largely at a part-whole level, where its definitions collides with metonymy or meronymy. Nerlich and Clarke (1999), in a historical summary of the evolution of the definition of synecdoche, explain that only the genus-specie type of definition has been kept, while the whole-part type became solely defined by metonymy.

For clarity, synecdoche is classified here as a subtype of metonymy, and is therefore included in the five types above. Similarly, cases of meiosis and hyperbole, or changes in intensity towards weaker and stronger meanings may be included in the melioration/pejoration category. These may reflect euphemistic uses, according to Bréal (1899). For some authors, such as Nazar (2011a), euphemism stands as a category on its own.

Indeed, whether strengthening and weakening are figures by themselves or whether they should be subdivided in lists of more precise phenomena is questionable. For Blank (1999) and Blank and Koch (1999), in continuation with Bréal, the question is rather whether words undergo specialization or generalization. On the taxonomical plane, they add co-hyponymic transfer (horizontal shift in a taxonomy)[26]. However co-hyponymic transfer may be classified under the same umbrella as generalization, as it often refers to vague denotation obtained through the resemblance of referents. To this, Bloomfield (1933) adds antiphrasis, or change anchored in the use of the opposite meaning (often sarcastic or ironical), auto-antonymy, or change of a word's sense to its reverse meaning, e.g., *bad* in the slang sense of "good," as well as auto-converse, the lexical expression of a relationship by the two extremes of the respective relationship, e.g., *take* in the dialectal use as "give."[27] In etymology, meaning reversal patterns are found across languages and centuries, such as the Lat. *perdere* which meant “to give completely” and gave birth to Sp. *perder* and Fr. *perdre*, among others, meaning “to lose”. The reversal relationship is sometimes complex as it includes logical continuum from a state to another. For instance, one who has given something away completely has therefore lost it, in the sense that it is not available to them anymore. For simplification purposes, the aforementioned rhetorical relations may be gathered under a single heading:

[26] For instance the word *pinecone* is used to refer to all types of cones from any conifer in British Columbia. Source :http://www.odlt.org/ballast/cohyponymic_transfer.html, accessed on 09/05/2012

[27] http://en.wikipedia.org/wiki/Semantic_change

6. Contrast relations: antiphrasis, auto-antonymy and auto-converse

One last process is folk etymology, also called popular etymology, by which people spontaneously attribute a meaning to a word by analogy, often because it intuitively sounds like another, even if their meanings are not really related. This can occur either within the given language or through resemblance with another language that is in contact. An example of this is the Dutch word *hangmat* ("hammock," but literally "hanging mat"), of which the form was associated to French *hamac*, while its constituent parts compose a transparent form semantically.

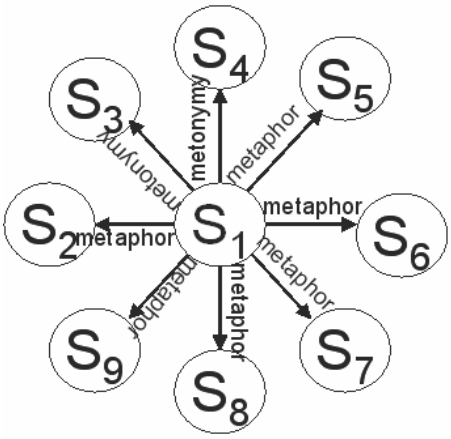
To summarize the cited rhetorical figures, Grzega and Schöner (2007) classify them in terms of relations:

1. "similar to" relation (metaphor)
2. "neighbor of" relation (metonymy)
3. "part of" relation (synecdoche)
4. "kind of" relation (generalization/specialization)
5. "sibling of" relation (co-hyponimic transfer)
6. "contrast to" relation (antiphrasis, auto-antonymy, auto-converse)

They also note that these relations may be combined.

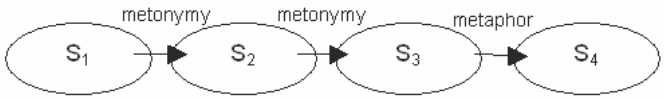
1.1.2.2.1. ***Combined processes***

Metaphor and metonymy may intervene according to different schemes and combinations. Two types of combination networks may be distinguished: *radial* shifts and *chaining* shifts. Balbachan (2006) illustrates these two processes with the example of *head* for radial shift, and *volume* for chaining shift:



sememe	Meaning 'head'
S1	**upper part of human body**
S2	seat of intellect
S3	life (*cf. it cost him his head*)
S4	image of head on one side of coin
S5	knobbed end of nail
S6	foam on top of liquor
S7	top of page
S8	fully developed part of boil
S9	end of table occupied by host

Figure 1 Radial structure of the semantics of *head* according to Balbachan (2006)



sememe	Meaning 'volume'
S1	**roll of parchment** *(disappeared)*
S2	book tome
S3	size, bulk of a book
S4	size, bulk of other things

Figure 2 Chaining structure of semantic change for *volume* (ibid.)

Both shifts involve combined processes of metonymy and metaphor, however in the first case they all derive from S1 whereas in the second S4 derives from S3 which derives from S2, itself deriving from S1.

1.1.2.2.2. *Typologies of neologisms mixing morphological processes and rhetorical figures*

Sablayrolles (1996), in his exhaustive typology of typologies for neologisms (mostly in French) asserts that he came across more than a hundred different ones. However, these typologies include phenomena of neology at large, whereas this doctoral thesis is focusing on semantic neology. This variety is explained by the author by the fact that different criteria or lenses are used in each approach: morphological productivity and rules of word formation, semantic criteria, etymological and functional criteria. Moreover, these typologies may prefer a lexicological, psycholinguistic or a sociological perspective or a literary, historical,

terminological or discourse analysis one. They may also combine some of these aspects to some extent. This makes synthesis extremely difficult. Another layer of complexity is added by the choice of content, whether the authors try and grasp rules for a given language or several, or for a given domain, specialised or not (street talk, specialized vocabulary in professional environments, etc.). These typologies look at *how* word meaning changes. Sablayrolles (1996) adapts the typology of Tournier (1991;1985) with some additions and changes, and chooses to organize the classification in terms of lexical productivity, that is to say, the processes that lead to the formation of words. By doing so, he integrates morphological rules of word formation and processes anchored in rhetorical figures in the same paradigm.

1	Préfixation **Prefixation**	affixation construction **Affixation, construction**	morpho-sémantique **morpho-semantics**	matrices internes **internal matrices**
2	suffixation **21**: flexion **suffixation 21 inflection**	// //	//	//
3	dérivation inverse **reverse derivation**	// //	//	//
4	Parasynthétique **parasynthetic**	// //	//	//
5	composition , **51** synapsie, **52** quasi-morph **composition, 51 synapsie (word composed of several lexical morphemes), 52 quasi-morph**	composition // **composition (compounding)**	//	//
6	mot valise **portmanteau word**	// //	//	//

7	onomatopée, phonesthème **onomatopeia, phonaestheme** **71** fausse coupe et jeu phonique: paronymie **71 fake cut and phonic play : paronymy** **72** déformation graphique **72 graphic deformation**	imitation et déformation **imitation and deformation**	//	//
8	Conversion **conversion**	changement de fonction **change in function**	syntactico-sémantique **syntactic-semantic**	//
9	construction différente **different construction**	//	//	//
10	Métaphore **metaphor**	changement de sens **meaning change**	//	//
11	Métonymie **metonymy**	//	//	//
12	autres figures (restriction, extens de sens) etc. **other figures (restriction, extension, etc.)**	//	//	//
13	Troncation **clipping**	réduction de la forme **reduction of form**	morpholo-gique **morphological**	//
14	Siglaison **Acronym formation**	//	//	//
15	Détournement **diversion**	Pragmatique **pragmatic**	//	
16	Emprunt **borrowing**	matrice externe **external matrix**		

Table 1 Table of word formation processes taken from Sablayrolles (1996), my translations are in bold

1.1.2.3. Word roles in typology

1.1.2.3.1. *Concrete and abstract word use and transitivity in typology*

Some early typological approaches such as Huguet's (1934), who studied word evolution from the XVIth century onwards, give more importance to the concrete and abstract nature of words in use, including idiomatic expressions. Huguet's typology includes the influence of institutions and norms, meaning restriction, weakening and degradation, synonyms, and shifts from abstract to concrete, or concrete to abstract. For instance, a word's "value" can drift from physical (concrete) to moral (abstract). The French word *droiture*, used to mean "materially straight" in the XVIth century and shifted to "morally straight" (Huguet 1934: 212). In a similar way, *divertir* was used to mean "to divert" and had a physical value as well as a moral value in its transitive form (e.g. *divertir une rivière, divertir la douleur*, "to divert a river", "to divert pain") (Huguet 1934: 217-218). Today, *divertir*'s first and stronger meaning is "to entertain/amuse others" while the reflexive form *se divertir* means "to entertain oneself". The verb's previous transitivity with things, concrete and abstract, became obsolete and induced change in meaning, although it survived with restriction in expressions such as *divertir l'attention* ("divert the attention"). Word meaning can also shift from abstract to concrete values as shows the French noun *bâtiment* (spelled *bastiment* in the XVIth Century) which used to refer to the action of building and shifted to refer to its result, "a building" (Huguet 1934:231-232). Huguet explores the relationship between words and ideas, and also depicts transitivity issues although he does not offer a substantial theoretical framework for the analysis of transitivity changes. The latter are well depicted by frame theories and actantial schemes.

1.1.2.3.2. *Actantial schemes*

Indeed, a share of meaning evolution is anchored in the changes in the relationships between words in their direct linguistic context, i.e. their respective roles when they combine. These relationships can be analyzed in terms of frames, or actantial scheme, to determine the changes in the roles that words have with each other. For instance, Rémi-Giraud (2000) studies the inversion of actantial scheme for the French word *air*. In the XVIIth century, someone's *air* referred to the way someone was behaving, in the sense that people were consciously acting to be seen under a certain light. This agentive meaning drifted to a non-agentive meaning in the XXth century, as *air* rather referred to someone's appearance, and to

the idea of being seen in a certain way by other people. In the first sense, there is an intention, in the second there isn't. This is reflected in use, involving a different scheme in grammatically attached co-text.

1.1.2.3.3. ***Pustejovskyan approach***

Adelstein's (2007) doctoral thesis and ongoing works, look at word meaning evolution in a Pustejojskyan perspective, and analyze specialized language, from the economic and financial world, in a variational dimension since the author studies Argentinian Spanish. The interesting point of the Pustejoskyan framework is that it allows for theoretical modeling of partial meaning transfers and mapping. Pustejovsky (1995:76) defines four aspects of word meaning, named *qualia*:

> "CONSTITUTIVE: the relation between an object and its constituent parts;
>
> FORMAL: that which distinguishes it within a larger domain;
>
> TELIC: its purpose and function;
>
> AGENTIVE: factors involved in its origin or "bringing it about". "

Adelstein showed at the "Jornada sobre neología[28]" in her presentation *Neología y variedades lingüísticas: investigaciones sobre el español de la Argentina,* that the Spanish word *pantalla (*both *board* and *screen)* underwent mapping in terms of function (telic qualia) and in terms of object (constitutive qualia). The same telic function applies whether we talk about a *pantalla* as a *board* (*black or white board*) or as a modern screen, since both are used to display information. However, the constitutive qualia differs since the board is made of grey stone, or of coated plastic, whereas the screen is made of complex plastics, and technology such as liquid crystal display. She shows how the integration of this word in the realm of finance in Argentinean Spanish subtly enriches its semantics by adding a new constitutive function. Words may travel from specialized terminology into mainstream language and oppositely, gaining and loosing fine aspects, within or across dialectal variation.

[28] http://www.iula.upf.edu/agenda/age040ca.htm

The above-mentioned works rely on a mostly mechanistic, intra-linguistic perspective. Some of them also include discussions about the "cause" of semantic change or neologisms, while some studies focus exclusively on theses causes.

1.1.3. "Causes" and "motivations:" factors external to language

1.1.3.1. Historical, social and emotional motivations

History and major political and cultural shifts both impact and are impacted by semantic change. Bréal (1899: 124) underlines the necessity for a historical understanding of language in a simple demonstration:

"On doit voir combien il est nécessaire que notre connaissance d'une langue soit étayée sur l'histoire. L'histoire peut seule donner aux mots le degré de précision dont nous avons besoin pour les bien comprendre. Supposons, par exemple, que pour connaître les magistratures romaines nous n'ayons d'autre secours que l'étymologie. Nous aurons : ceux qui siègent ensemble (consules), celui qui marche en avant (prætor), l'homme de la tribu (tribunus), et ainsi de suite. Ces mots ne s'éclairent, ne prennent un sens précis, que grâce au souvenir que nous en avons, pour les avoir vus dans les récits des historiens, dans les discours des orateurs, dans les formules des magistrats. En même temps que l'histoire explique ces mots, elle y fait entrer une quantité de notions accessoires qui ne sont pas exprimées."[29]

Not only it is necessary to analyse language against the background of history, but it is striking to look at the importance given by people to neology in history. For instance, the importance of the study of neologisms was evaluated by Tilby (2009) in XIXth century France as "obsessive," relating it with the need to "rethink social political and cultural norms after the excesses of the Revolution" (Tilby 2009: 676). Indeed, the constant need for an understanding of new meanings may emerge from the desire for newness in society at large,

[29] "We must see how necessary it is that our knowledge of a language be supported by history. Only history can provide words with the precision we need to understand them properly. Suppose, for example, that in order to understand the Roman magistracies we have no other help than etymology. We obtain: those who sit together (consules), the one who walks forward (praetor), man of the tribe (tribunus), and so on. It is only thanks to the memory we have of these words,, having seen them in the accounts of historians, in the speeches of orators, in magistrates' formulas, that these words are shed light upon, and take on a precise meaning. At the same time that history explains these words, it also incorporates a series of accessory notions in them which are not expressed"

at the sociological and individual levels. Sociologically, the change in terminology is highly related to critical historical transitions such as the French Revolution. For instance, the words *monsieur* and *madame* ("sir" and "madam") were –forcefully- replaced by *citoyen* and *citoyenne* ("citizen")[30], after the Revolution; however these changes did not persist and the original terms were quickly re-instituted. These types of imposed vocabulary changes are still happening today in countries in which there are governmental bodies and recognized institutions that have the power to do so. In France and Spain, for instance, numerous attempts at "correcting" attested pejorative connotation have met failure. Let us consider the attempted replacement of the expression *femme de ménage* ("cleaning lady", literally "cleaning woman") by *technicien/ienne de surface* ("cleaner", literally "surface technician") to respond to feminist criticism and value the work status. Although official sources, such as newspapers and TV, tried to adopt the new term for a while, the trend quickly faded, as people kept using *femme de ménage*. A series of compromising expressions became more popular in the meantime: *dame de ménage* ("cleaning lady"), *aide à domicile* ("home help"), among others, showing high hesitancy on the part of speakers to find a politically correct expression while avoiding the one forced upon them. In France, a series of laws relative to the correct use of language and its evolution attest to an effort at governmental level, empowering terminology commissions. For example, the "Commission générale de terminologie et de néologie" provides instructions to institutional bodies as to the vocabulary they should use, which are published in the Journal Officiel, the organ of the French Government. Moreover, there are laws on the protection of the French language which attest to the linguistically conservative attitude of the French government (see laws of 1989, 1994 and 1996).[31] Forceful diffusion can encounter failure, however language is a powerful tool for politics, ideology or propaganda, as brilliantly demonstrated by Klemperer (2003) with real data and Orwell (1949) with fictional data, among others.

---

[30] Source: http://www.tlfq.ulaval.ca/axl/francophonie/HIST_FR_s8_Revolution1789.htm. Last accessed on 23.03.2012

[31] Decree 89-403 , 2nd june1989 instituting a « Conseil supérieur de la langue française et une délégation générale à la langue française » (High Council for the French Language and General Delegation for the French Language). Law 94-665, 4th august 1994 ("Loi Toubon"), on the use of the French language. Law 96-602, 3rd July 1996 on the enrichment of the French Language.

In effect, words intrinsically capture and categorize concepts, ideas and referents in different ways according to and impacting the social and cultural changes. The way we categorize the world is both dependent on language and built by it. Therefore, there is an obvious link between cultural and social preoccupations and the study of "newness," ultimately *changes*, in language, both a mirror of the world and ourselves and a collective social construction.

This interest raised by semantic change studies also took on various forms such as collections; in the same way that people collected butterflies and insects in cabinets of curiosities, others would collect new words, as Littré (1888) did, by systematically isolating 98 words that have undergone radical change over centuries.

The psychology of speakers and their interaction with society and culture is taken into account in some works on semantic change, although the interest it raised has not equaled (yet) the interest for language-internal and structural classifications.

As for Ullmann's typology, Bréal's early works already contained a typology of social and psychological motivations, summed up by Traugott and Dasher (2002: 58,59) as follows:

" (i) avoidance of difficulty

(ii) securing of greater clarity

(iii) taboo and euphemism [….]

(iv) fading and discoloration , or loss of semantic content […]

(v) external factors such as cultural change"

1.1.3.2. Early approaches of psychological and emotional causes

According to Ullmann (1962), Sperber (1923) applied a Freudian approach to semantic change issues. His argument was mainly based on the emotional attachment speakers develop with certain topics, making them "centres of expansion" and "centres of attraction." By "centres of expansion" he means topics which we are more likely to expand on, via metaphorical processes to describe other issues. By "centres of attraction," he means that speakers tend to bridge other topics with the emotionally valued ones, therefore creating analogies and variety. Ullmann shows how fashionable topics in an era tend to produce idioms, such as the plethora of religious-based idioms in XVIth century France (such as *vray*

comme la messe, "as true as mass"), science and medicine derived expressions after the Revolution (such as the "electrifying" effect, or "centrifugal force" of the revolution) or the railway derived idioms in the XIXth century (Ullmann 1962: 202).

Stern's (1931) "Meaning and Change of Meaning" is an attempt at adding a structured psychological and emotional dimension to the study of semantic change. The author devotes as much focus to these elements as he does to typological considerations of causes and mechanisms. He classifies mechanisms of sense change under seven major headings:

- substitution (new meanings for an existing referent),
- analogy (e.g. transfer of a sense from a part of speech to another, for instance the adverbial sense of *fast* is obtained by analogy with the adjectival form : fast 1)"firm", 2) "quick"),
- shortening,
- nomination (transferred name from one referent to the other),
- transfer (based on similarity between referents, e.g. *bed*),
- permutation (e.g. *beads*: "prayers"→ ME[32] "beads", the balls of the rosary by which prayers were counted, Stern 1931: 353),
- adequation (adaptation of meaning to the referent, e.g. *horn* "animal horn" → "musical instrument").

Stern distinguishes between changes in the objective world and changes brought about by the individual's apprehension of the world (see Geeraerts 1983: 230). He makes a distinction between change and "fluctuation", the idiosyncratic variation in use, and notes that fluctuation is one of the origins of long term change. It is striking that such an early work gave central preponderance to both structural classification as well as emotional and psychological issues. In 1931, Stern already dealt with considerations that would become central to modern cognitive science and prototype theory, notably processes of analogy, meaning transfer and regularity issues, as well as the role of vagueness and the central or peripheral nature of meanings. The same may be said of considerations that became central to pragmatics such as speaker's attitude, adequation (cognitive and emotive) and irony. Stern's approach of emotion

[32] Middle English

is inclusive of connotation, at the collective level (encoded in the semantics of a word), at the sociological level (within a group of speakers) and at the individual level (subjectivity).

To underline the importance of one's subjective apprehension of the referent, Stern (1931: 41) states:

"Meaning is essentially personal. What a word means depends also on who uses it, when, where, why, in what circumstances, with what aim, with what success."

Moreover, a person's own psychological associations can influence meaning. One's associative background is "mental context, which, through the medium of the subjective apprehension, may exercise a certain influence on the meaning of the word" (*Op.cit:* 62)

Stern investigates "psychic elements and categories of meaning," and offers a classification of cognitive and emotive elements. Although a few theoretical assumptions are clearly outdated, the study of psychological elements seem a necessary addition to structural typology and is something which has not been dealt with by most other works on the subject.

Stern classifies cognitive and emotive elements as follows:

- The referent has emotional value (the emotive element is permanent and therefore it is a semantic element), either
    - emotion is the referent,
    - or emotional elements are central or peripheral (as will be developed later by cognitive linguistics).
- The emotive element is incidental (as will be developed later in theories about context, pragmatics and speech act theory).
- The speaker's attitude towards the referent is emotional (and may be expressed in choice of "polarity" as termed by subsequent works in linguistics).
- Other elements of emotion interfere in the situation of communication, such as the motives of communication, volition, factual circumstances (context) and pragmatic elements (tone, gesture, etc.). Moreover, the awareness of these elements is in itself a source of emotional content.

Stern's theoretical approach is grounded in the transition between theories based on symbols and theories based on cognition. The terminology hesitates between "symbolic" and "cognitive" as he creates a theoretical bridge between the two:

"The term 'symbolic' corresponds to what I have called 'cognitive'. It is true that any cognitive element of meaning must be, by definition, equivalent to an element of the language-user's apprehension of the referent, and in so far is representative or symbolic of the latter." (*Op.cit:* 47)

He also distinguishes meaning from signal, in the same way that word and speech are distinguished. To him, emotion is part of the signal, and therefore must be studied within the context of utterance.

"the signal system runs parallel with the symbols and is able to turn identical symbols into expressions for very different emotions." (*Op.cit:* 55)

Stern also offers a discussion on the role of mental images, and goes as far as discussing mental imagery as an "inner speech form." This theoretical position is evidently anchored in the era in which he is writing and has benefited from much discussion and refuting in more recent cognitive science research. Stern refers to an early cognitive theory, without necessarily supporting it, which purports that there are "imageless thoughts." This theory contends that mental imagery is a third kind of mental content, additional to images and thoughts. Stern concludes that "images seem to belong to mental context" (*ibid*: 53). The image is a fragment of mental activity. It is a symbol of it rather than an expression of it (see Delacroix 1924). However, cognitive elements are not limited to images, he contends, and include other perceptual aspects.

As regards the pragmatic dimension (that he does not call "pragmatic"), Stern points out the centrality of volition and quotes Delacroix (1924: 374):

"If a thing were quite indifferent to me I would not say it."

In the communication process, the source of emotional elements is outside the "triangle of subject-referent-word," in the subjective attitudes of speakers and hearers and can be additionally caused by external factors. Therefore "the import of the proposition is

communicated, but whether the hearer accepts the judgment depends on him" (Stern 1931: 57).

To depict the mechanisms of central and peripheral features in meaning and how they may or may not be convoked, depending on context and intention, Stern (1931: 61) gives the following example:

"If a builder speaks of *bricks* as a possible material for facing a building, he is probably thinking mostly of their colour and external aspect; if he is speaking of bricks as an alternative material for foundation, he is thinking of their durability and resistance to high pressure; if he is discussing the number of bricks likely to be required for a certain construction, he is turning his attention mostly to their size; and if he is asking about the number of bricks delivered last week, he will be thinking of them as entities, without paying attention, for the moment, to their characteristics."

Stern also highlights the role of vagueness in change processes. Vagueness can be based on objective uncertainty (related to one's knowledge of the world) or linguistic uncertainty (related to one's knowledge of language):

"Linguistic and objective vagueness involve a certain elasticity of the objective and semantic range, owing to lack of definite limits. If a border is vague, it may be easily over-stepped." (*ibid*: 68)

1.1.3.3. History, sociological interactions and sociolinguistic changes

Stern (1931: 59) also points out that "…some words, as *bourgeois* and *capitalist,* have a different emotional colouring for different classes in the speaking community." Nevertheless, this emotional colouring pertains to the field of sociology, studied by other authors, notably Meillet.

Meillet (1906) defends a sociological approach in which speaker interaction is central to linguistics. He treats language as a *fait social*[33] in the line of Durkheim's theory, placing the individual at the center of the analysis. Semantic change is grounded in the effects of the

[33] Literally a « social fact »

distribution of individuals into social groups. Each individual may belong to several social groups, through which they spread semantic changes, in particular via *borrowings*, here defined both as social borrowings and cross-linguistic ones. Meillet (1906) analyzes semantic change through linguistic, historical and social dimensions. At the psychological level, language is learnt in a discontinuous fashion, as children may attribute meanings to words that are not the meanings accepted and intended by adults. For Meillet, the discontinuous nature of transmission also allows linguistic changes to take place. He distinguishes three major mechanisms: The first type of change applies only to a few cases and is rooted in a grammatical process. Either the grammatical form of the word is the agent of change or conversely. One of the examples of this is the word *homo* in Latin that designated human beings, that gave *homme* in French with the same meaning, and then took on the meaning "human being of the masculine sex" since *homme*'s grammatical genre was the masculine. The second type of change is when the things referred to by words come to change.

" Les changements de ce genre atteignent constamment presque tous les mots ; mais on ne les remarque que lorsqu'ils présentent quelque chose de singulier et d'étrange : on dit du *papier* (latin *papyrum*) de chiffons ; la *plume* de fer s'est substituée à la *plume* d'oie sans que le nom ait varié ; et ainsi de suite : les changements des choses ne se traduisent que d'une manière restreinte par des changements des mots : car les mots étant associés à des représentations toujours très complexes s'associent facilement à des représentations qui ont avec celles d'une génération précédente quelques traits communs. Et c'est ainsi que la variation de sens de beaucoup de mots, c'est-à-dire au fond la variation des notions auxquelles est associé le nom donné, traduit des changements sociaux plus profonds[34]" (Meillet 1906 :10, e-book version).

The third and most important cause of semantic change is the repartition of people into social groups:

[34] "These types of changes apply constantly to almost all words; but they are noticed only when they show something unique and strange: we say "paper" (latin *papyrum*) of rags; the iron *feather* has substituted the wild goose *feather* without any variation in the name, and so forth: the changes in things result only in a limited way in changes in words: since the words are associated with representations that are always very complex, they associate easily with representations that have features in common with the previous generation. And it is so that the variation in meaning of many words, that is to say, in substance, the variation of notions to which the given name is associated, shows deeper social changes."

" Le fait fondamental est donc qu'un mot qui, dans la langue commune d'une société, a un sens étendu s'applique, dans un des groupes restreints qui existent à l'intérieur de cette société, à des objets plus étroitement déterminés, et inversement[35]" (ibid:12)

This view takes into account the associations a word has in a given human group, be it the distinction between men and women, or a professional, social or cultural group. There is a paradoxical dynamic in the fact that society at large tends to generate uniformity in language, whereas specific groups tend to generate differentiation, creating their own *jargons*. These jargons may then spread and enter language at large. They may also be rooted in etymological processes. Meillet cites the French word *arriver* for instance, which is etymologically related to the social group of sailors since its etymology is the Latin *ad-ripare*, "come to shore." The association with success at getting somewhere loses its strength as the word shifts from sailor vocabulary to the general language.

Hughes (1992) defines three types of sociolinguistic changes: symbiotic changes, mediated changes and "Orwellian changes." Symbiotic changes and mediated changes are defined in terms of historical period. Symbiotic change is defined as a process that could take place before the development of mass media; it is the result of broad social changes. For instance *noble* underwent a meaning shift, from referring uniquely to a social class to acquiring a moral meaning. Hughes attributes this shift to the breakdown of feudalism. In the same way, *profit* used to mean "beneficial to the whole social organism" and shifted to refer to "private profit", with the advent of capitalism (Hughes 1992: 113).

Mediated changes started to take place in the era of the printing press and mass diffusion of information. For instance, the meaning of *bourgeois* took on a pejorative meaning of "exploiter of the proletariat" after the publication and success of Marx' communist manifesto.

"Mediated changes occur as a result of interference in the semantic market by some powerful agency, and can be said to have the tacit or willing acceptance of a substantial portion of the speech community" (Hughes 1992: 118).

---

[35] "The fundamental fact is that a word in the common language of a society, which has an extended meaning, is applied to restricted groups existing inside this society to objects that are more closely determined, and vice-versa."

The third kind of sociolinguistic changes, "Orwellian" changes, involve manipulation by a powerful body:

"'Orwellian' changes involve semantic manipulation by an oligarchy in control of the media, usually for the purposes of ensuring ideological conformity via political propaganda" (Hughes 1992: 118).

For instance, in South Africa, *group* was substituted to *race*, and *townships* or *locations* to "non-white urban areas", under the pressure of Apartheid. For Hughes, the political use of *green* follows the same principle. Moreover, Hughes questions the arbitrariness of the sign, dear to Saussure, arguing that Saussure wrote before Communism, Nazism and Apartheid that the sign and the arbitrariness of the sign protected language from any attempt to modify it. However, in the light of these historical events, Hughes classifies this idea as naïve (Hughes 1992: 122).

1.1.3.4. Multiple paradigms

Keller's (1994) "invisible hand in language" theory also pays attention to the relationship people have with language, in that they create it, making it an artificial artifact. However, its evolution is not predictable or controllable and therefore language is also a natural phenomenon. The fact language is both artificial and natural, he contends, makes it a "phenomenon of the third kind," both a process and a finite object, molded by collective actions and intentions, but yielding an unintentional result.

"They [semantic changes] are, like those of the second kind [artificial phenomena], the results of human actions, and they are, like those of the first kind [natural phenomena], not the goal of human intentions" (Keller 1994: 56).

Language is therefore not a human-made artifact, neither is it a completely natural phenomenon, and thus it is a bit of both. The idea of the "invisible hand" is borrowed from Adam Smith (1982) who coined it to describe the self-regulating nature of markets in economics. The idea that language is a self-regulating process and artifact opens the door to the search for the mechanisms and rules underlying this self-regulation, therefore displacing the issue, and bringing us back to structural research, with a twist as to its functional nature.

Keller (1994) bridges the micro-level of the individual action with the macro-level of the institution, and posits that three steps are found in the "invisible-hand" theory. The first relates to individuals, their intention, goals and the context and conditions of their action. The second is the process based on individual action generating structure. The third is the structure, or result.

The holistic attempt in Keller's work is admirable, since contributions to language change issues often choose to address the issue under specific angles, and within extremely specific frameworks, consequently often missing the "big picture." Other authors have tried to bridge the different angles, bringing the individual, society, language and psychology together. Nerlich and Clarke (1988), for instance, delineate four paradigms: the individual, the collective, the linguistic and the psychological.

1.1.3.5. Beyond causal explanations

Extending semantic change understanding to the individual (psychology) and the collective (society) is an issue that must be addressed. However, the limit of a purely causal approach seems to be well summarized by Coseriu:

"En el fondo, la perplejidad frente al cambio lingüístico y la tendencia a considerarlo como fenómeno espurio, provocado por "factores externos", se deben al hecho de partir de la lengua abstracta – y, por lo tanto, estática -, separada del hablar y considerada como *cosa hecha*, como *ergon*, sin siquiera preguntarse qué son y cómo existen realmente las lenguas y que significa propiamente un "cambio" en una lengua. De aquí también el planteamiento del problema del cambio en términos causales…" [36] (Coseriu 1958: 17)

Coseriu refutes the classical distinction between innovation and result, rather referred to as a product (ergon) to insist on the idea that semantic change is a dynamic process (energeia). He further contends:

[36] "Basically, the perplexity before language change and the tendency to regard it as a spurious phenomenon, caused by" external factors ", are due to the fact of starting from abstract language - and, therefore, static - separated from speech and considered as *a given*, as *ergon*, without even asking what they really are and how languages actually exist and what in fact a "change" in a language really means. From this as well the approach of the problem of change in causal terms."

"Los cambios lingüísticos, […] sólo pueden explicarse (motivarse) en términos funcionales y culturales. Pero las explicaciones culturales y funcionales de los cambios no son de ningún modo "causales". La idea misma de 'causalidad' en la llamada "evolución" idiomática es un residuo de la vieja concepción de las lenguas como "organismos naturales", así como del sueño positivista de descubrir les supuestas "leyes" del lenguaje (o de las lenguas) y de transformar la lingüística en una "ciencia de leyes" análoga a la ciencias físicas."[37] (Coseriu 1958: 101)

The study of both language-internal mechanisms and external causes is necessary to encompass the complexity of the phenomena. Since several reference works in the area simply classify typologically these mechanisms and causes separately, not much has been said yet about the interaction of these two levels. However, some typologies do mix the two levels and some go beyond that by offering insight about how they interact.

1.1.4. Bridging internal and external causes

1.1.4.1. Classificatory attempts at bridging internal and external causes

1.1.4.1.1. *Classifying structural mechanisms of semantic change: Ullmann's approach*

Another theoretical landmark for classifying semantic change, according to Geeraerts' critique (1983) as well as Blank's (1999), is Ullmann's classification in terms of structural mechanisms, developed in his works (see Ullmann 1951; 1953; 1962; 1972). For Blank (1999:66), Ullmann's distinction between "causes," "nature" and "consequences" of semantic change "[…] has been for decades the most popular and important theory in this domain," and Ullmann's typology "until the 1990s was reputed to represent the "state-of-the-art in historical semantics" (Blank 1999: 70). Ullmann's work is based on the Saussurian distinctions between *signifiant* and *signifié*, transposed as sense transfer, and name transfer, and the paradigmatic and syntagmatic dimensions, transposed as semantic similarity and contiguity. It also strongly

[37] "Semantic changes, [..] can only be explained (motivated)… in functional and cultural terms. But the cultural and functional explanations of changes are not, in any way, "causal". The very idea of "causality" in what is termed idiomatic "evolution" is a vestige from the old conception of languages as "natural organisms" and from the positivistic dream of discovering the putative "laws" of language (or of the languages) and of turning linguistics into an "exact science" (science of laws) similar to the natural sciences."

draws from the central notion of semantic fields, transposed into network associations. The classification:

"[…] is based on the distinction between changes due to linguistic conservatism and those due to linguistic innovation. Within the last category, transfers of names, transfers of senses, and composite transfers have to be distinguished." (Geeraerts 1983: 217-218)

Linguistic conservatism can be defined as the mechanism of maintaining a word when its referent(s) change(s). The word undergoes semantic shift to adapt to the new referents. Ullmann provides the example of *pen,* whose meaning broadened to "any writing instrument with ink" as the quill pen was replaced by ballpoint, fountain pens and other types of pens. The linguistic referents change, but word forms stay and their meaning adapt. Linguistic innovation, however, is based on changes in the associative network of the word. Within this network, Ullmann distinguishes two types of associative links: contiguity (metonymical) and similarity (metaphorical). The idea that changes take place in the associative networks of words is used in this doctoral thesis, however, it is used in a broader sense, as the studied associative networks include linguistic context of use additionally to semantic fields.

Ullmann distinguishes between three types of linguistic innovation, transfer of names, transfer of senses and composite transfers. Transfer of names relies on the associative link between two senses. A form becomes used for the other by association. Transfer of senses relies on the associative link between two forms, causing one concept to be named by the other form. Composite transfers are both transfers of names and senses. The example of *a Rembrandt* which is an ellipsis of *a Rembrandt picture* is based both on name contiguity and sense continuity, in the metonymical transfer author-picture (example given by Geeraerts 1983: 218).

Table 2 sums up the major mechanisms at stake:

	Transfer of signifiant, based on associative link of signifiés	Transfer of signifié, based on associative link of signifiants
Paradigmatic links: similarity	METAPHOR	POPULAR ETYMOLOGY
Syntagmatic links: contiguity	METONYMY	ELLIPSIS

Table 2 Ullmann's classification of transfers, taken from (Geeraerts 1983: 219)

However, even if Geeraerts (1983) asserts that Ullmann's classification is sound, he formulates a few criticisms and offers his own classification based on them. Beyond terminological criticism, Geeraerts points out that Ullmann does not deal with the introduction of new words (neologisms) on the basis of morphological productivity and borrowings. Ullmann is rather concerned with transfers anchored in the existing system and does not take into account the creativity of the system and its permeability. His typology simply overlooks new referents. However, even if I agree with Geeraerts (1983) in that Ullmann's theory does not encompass all the phenomena of semantic change, the theoretical description of transfers via associative networks is nevertheless highly beneficial to study semantic neology. Geeraerts (1983) adds that Ullmann does not include any factor external to language, as social or emotional factors introduced by Meillet (1906) and Sperber (1923). In my opinion, this critique may be extended to all classifications and typologies that exclusively focus their efforts on language-internal mechanisms, as if language had no relationship whatsoever with its speakers.

However, Ullmann does include socio-cultural factors at some level, as he contends that semantic change can be due to the need for a new name (when a new referent has appeared), historical causes (from ideas and concepts influenced by social, cultural, political and technical factors), foreign influences (the influence of changes that have taken place in other languages), or finally linguistic causes (when collocations gives birth to transfer of senses).

Ullmann (1962: 193-197) lists six factors facilitating semantic change:

1. Discontinuity in transmission
2. Vagueness
3. Loss of motivation (severed etymological connection)
4. Polysemy
5. Ambiguous contexts
6. Structure of the vocabulary

The first factor, the notion of discontinuity in language, being handed down from generation to generation, is taken from Meillet (1906) who has a sociological approach of the issue. Moreover, while dealing with "causes", Ullmann warns semanticists looking for etymological explanations only, and states that the history of civilization is essential to understand semantic change, citing examples as the origins of the French word *croissant* having been coined in German (*hörnchen*), before it was then translated into French, to name the bread rolls with a crescent shape after a victory over the Turks, whose emblem was a crescent (Ullmann 1962 : 197). Ullmann's (1962) classification retains most of the central processes already cited, but organizes them differently, as follows:

<u>Nature of semantic change</u>

1. Similarity of senses (metaphor)
   Includes anthropomorphic and animal metaphors, metaphors translating concrete concepts into abstract ones, as well as synaesthetic metaphors
2. Contiguity (metonymy)
3. Similarity of names (Popular etymology)
4. Contiguity of names (Ellipsis)

<u>Causes of semantic change</u>

1. Linguistic causes (collocation based, equivalent to "contagion")
2. Historical causes
   a) Objects (new referents in discontinuity with the words, which stay unchanged, as in *car*)
   b) Institutions (new referents in discontinuity with the words, which stay unchanged)
   c) Ideas (idem)

3. Social causes
4. Psychological causes
    a) Emotive factors (see Sperber 1923)
    b) Taboo
5. Foreign influence
6. The need for a new name

Consequences of semantic change

1. Change in range: extension and restriction
2. Change in evaluation: pejoration and amelioration

This classification serves as a sound basis for further research in diachronic semantics. Among the offered theoretical approaches in that field, I have chosen to focus on Blank's and Geeraert's theories, which stand as strong representatives of the main currents in cognitive semantics.

1.1.4.1.2. *Cognitive typology of motivations and diachronic cognitive onomasiology: Blank's approach*

Blank (1999) analysed lexicalized innovations, since they show pragmatic acceptance by speakers. Change is perceived as result of social interaction in that it arises for efficiency matters, when speakers need to get something across and do not find any available way of expressing it. This idea is called *expressivity*. It strongly relies on the theoretical grounds of relevance theory (see Sperber and Wilson 1995), and Grice's cooperative principle (see Grice 1975), in that expressivity is defined as a communicative strategy that ensures that the hearer will get the point. Blank also insisted on the idea that this communication must take place with the "least possible expense," to communicate efficiently, echoing Grice's maxim of quantity. Expressivity and efficiency are the two basic motivations underlying other processes.

Blank's (1999) typology consists of six types of motivations:

1. New concept creating the need for a new name

2. Abstract concepts, distant and usually invisible referents, at the root of metaphorical and metonymical processes

3. Sociocultural changes that are at the root of conceptual shift of existing words

4. Close conceptual or factual relation, at the root of name transfers and polysemy. These include:

-frame relations, when a strong relationship between two concepts is at the root of expressing them with one word

-prototypical change (defined differently from Geeraerts [1983]), at the root of restriction, when a word is used to refer to the prototype of the category, rather than the whole category, or conversely, extension from the prototype to the whole category

-blurred concepts, arising when meaning transfers are made on the basis of a confusion, on the basis of concept resemblance

5. Complexity and irregularity in the lexicon, being reduced by speakers according to the efficiency principle, including:

-lexical complexity: the more frequent a word is, the less complex it becomes semantically, as speakers tend to reduce its *signifiant*.

-"orphaned word": contrary to lexical complexity, lexically isolated words tend to be integrated in a related class by speakers, via a process of popular etymology (previously referred to as folk etymology). For example, Fr. *forain* "non-resident" > "belonging to the fair" (example from Blank 1999: 78).

-"lexical gap": caused by asymmetric lexical structure, and existing polysemy with a large meaning gap between senses, therefore synonyms or co-hyponyms develop a similar sense to the source word. This is also coined "synonymic derivation". For instance Lat. *eques* meant "cavalryman" as well as "knight", therefore *eques*' co-hyponym *pedes*, "infantryman" gave birth to the meaning "plebeian" (example from Blank 1999: 79).

-“untypical meaning/ untypical argument structure”: when a word’s meaning is untypical of its word class, the tendency is that they shift towards a more prototypical meaning. The same process applies when a word has untypical argument structure, where the argument type shifts to a more normative argument structure.

6. Emotionally marked concepts:

Some conceptual domains are marked (or connoted in my terminology) with emotions, and may be furthermore marked with culturally-specific associations (taboo, for instance). Taboo concepts, like death, tend to be expressed by euphemism (for instance *to pass away* instead of *to die*, in Blank’s own example), or oppositely by dysphemism (an exaggerated expressive form). Euphemisms are based on metonymy, metaphor, semantic restriction, ellipsis or ironic antiphrasis. Emotionally marked conceptual domains may also give birth to hyperbole (based on metaphor, metonymy, extension of meaning, co-hyponymous transfer and contrast-based semantic change).

It is remarkable that Blank introduces a whole category based on the emotional connotations of concepts. Emotions are widely overlooked in numerous typologies, even if they are at the core of the need for expressive strategies, to either communicate or hide them, or make them seem different for social reasons.

Blank's (2003) posthumous publication “Words and concepts in time: Towards diachronic cognitive onomasiology” lays the theoretical framework for a modern cognitive onomasiology, in continuation with the aforementioned typology. The author insisted that both a taxonomical (similarity between concepts) and engynomical (contiguity between concepts) approach are necessary and complementary. He also insisted on the necessity to study several languages, whereas a lot of studies focus on one language only. He applied the idea of source and target concepts to pathways of diachronic change:

“Diachronic cognitive onomasiology […] looks for source concepts that seem to be universally recurrent, lays bare the associative relations between source and target concepts, and describes the lexical processes used by the speaker” (Blank 2003: 59).

1.1.4.1.3. *Cognitive typology and diachronic prototype semantics: Geeraerts' approach*

Geeraerts offers a typological approach in his work "Reclassifiying Semantic Change" (1983) which is further enriched with a much wider theoretical apparatus in his book "Diachronic Prototype Semantics" (1997).

Geeraerts's (1983) article is grounded in Ullmann's works, and expands on the aforementioned classification. Among the major points of criticism he raises, is that, according to him, Ullmann does not include new words (neologisms) on the basis of morphological productivity and borrowings, and he does not include psychological and emotional dimensions (though he admits that the approaches of Meillet and Sperber are complementary to his). Moreover Geeraerts judges the "distinction between name transfers and sense transfers […] terminologically spurious" (1983: 220).

Geeraerts further formulates a critique of Ullmann's functional and structuralist approach, which, in my opinion, is relevant to other structural approaches in the field:

"…restricting the functional processes under investigation to transfers within a given system diverts the attention from innovations transcending the limits of that system, and restricting the structures of investigation to cognitive ones diminishes the interest in the role of the non-cognitive (social or emotive) structure of language" (Geeraerts 1983: 224).

In his opinion, Ullmann also fails to distinguish between "ultimate causes and structural background" (*Op.Cit*: 237) of semantic change. He therefore offers a theory that accommodates this distinction.

Geeraerts distinguishes between expressive needs and efficiency principles, as two sides of the same coin. Expressivity corresponds to creative power. Expressive need in turn corresponds to a social and formal need in the community of speakers, while efficiency optimizes the creation obtained via expressivity. The principle of conceptual efficiency is a theoretical bridge with prototype theory, which gathers concepts into organized categories, according to a principle of economy/efficiency (see Rosch 1978; 1973; and Martinet 1964 for the original concept of economy in phonology).

Geeraerts therefore offers an alternative view, in which he includes the omitted elements of Ullmann's approach: borrowings, new concepts, synonyms and homonymic clash resolution:

	Expressive mechanisms	Principles of efficiency
On the conceptual level (with regard to *signifié*)	Borrowing lexical items that Introduce new concepts	Metaphor metonymy
On the formal level (with regard to *signifiant*)	Borrowing lexical items synonymous with existing words	Resolving homonymic tensions

Table 3 Geeraert's alternative classification to Ullmann's. Taken from Geeraerts (1983: 226)

Expanding on this preliminary table, formal and conceptual expressivity are defined as the main motives for creating or borrowing new words.

Formal efficiency is described as a pragmatic strategy, embedded in the situation of use, which accounts for the extra-linguistic motivation of ellipsis in addition to its formal nature. The same applies to popular etymology, being motivated by the need for transparency which is also a pragmatic strategy of efficiency.

Formal expressivity encompasses morphological production. Social and ideological factors (in particular, purism) influence the choice between a loan word and a morphologically-derived neologism.

Conceptual expressivity can be influenced by social factors, including cultural and historical ones. However, expressive needs are different for each individual, in terms of emotional attitude towards it. Therefore Geeraerts includes the emotional factor at the level of content (*signifié*) in addition to the cognitive level.

Conceptual efficiency is at stake in prototypical changes, which are rooted in metaphor and metonymy. Geeraerts, going against componential and logical semantics, argues that metaphors are prototypical phenomena, as is metonymy, to a lesser degree. He contends that "…prototype theory allows historical semantics to discern types of semantic change that go

beyond the limits set by the notions of metaphor and metonymy" (1983: 233) via family resemblance structure (*à la* Wittgenstein). Metaphorical and metonymical transfers, being category-based, "enhance the economy of the system" (*ibid:*233) according to a principle of the least effort, and reinforce the "structural stability of the system" (*ibid*:234).

	Expressive factors	Efficiency principles
On the level of (cognitive or emotive) content	Introducing new concepts expressing -objective changes in the world -subjective changes in one's knowledge of/attitude towards the world	Prototypical transfers, more particularly -metaphor (similarity relations) -metonymy (contiguity relations)
On the formal level	Introducing new word forms with social or stylistical expressive values	a. "One form, one meaning" (avoiding homonymic clashes) b. Economy of expression (ellipsis) c. Transparency (popular etymology)

Table 4 Table of semantic change factors taken from Geeraerts (1983: 234)

At the heart of Geeraerts' s (1983) argument, the concepts of conceptual economy and cognitive efficiency are the two principles underlying semantic change.

"Conceptual expressive need" (new referent, new attitude to an existing referent) "formal expressive need" (borrowings, morphological productivity) and "formal efficiency" (social and structural strategy) may be primary triggers of semantic change while "conceptual efficiency" (prototypical transfers) may not. It is rather a process. The cited needs and motives can and do combine most of the time, however Geeraerts admits it may be difficult to

decide precisely what these combinations are. To summarize these four factors Geeraerts (1983: 237) writes:

"Principles of formal efficiency guide the tendency towards an optimal relation between *signifiants* and *signifiés*. Principles of conceptual efficiency guide the optimalisation of conceptual organization as such. Formal expressivity involves the expressive value of the linguistic forms. Conceptual expressivity consists of the basic function of language, i.e. to convey ideas or feelings."

The attempt at considering all aspects of the issue of semantic change is remarkable, since Geeraerts combines an interactional perspective via pragmatics (formal efficiency), a structural approach via morphology and language contact issues (formal expressivity), a psycho-sociological approach giving space both to the individual and the collective (formal expressivity) and finally prototype theory, providing a framework of explanation for conceptual networks structures and transfer types (conceptual efficiency).

However the exact role of prototype theory in this framework and the defense of its suitability to deal with semantic change issues are not detailed depth in this article but in a subsequent work, *Diachronic Prototype Semantics* (1997). In this book and a digest of it (Geeraerts 1999), he contends that, in synchrony, a prototypical structure possesses four features: a degree of typicality, family resemblance structure, and its categories are blurred at the edges and cannot be defined by necessary and sufficient attributes. The degree of typicality is expressed in salience effects and structural weighting. The blurriness of the categories allows for flexibility.

Diverging from the (1983) article, Geeraerts modifies the roles he had attributed to expressivity and efficiency, contending that factors of prototypicality on the one hand and of isomorphism and iconicity on the other are rooted in the efficiency principle being either speaker-oriented or hearer-oriented. Expressivity is relegated to an essential background motivation.

I have mentioned earlier that the suitability of distinctions such as language and speech or synchrony and diachrony has been questioned by many theoreticians. In the same vein, Geeraerts questions the distinction between encyclopedic knowledge and the level of senses

established by structural semantics. To him, it is an unfruitful distinction in cognitive semantics since encyclopedic information is as important as the level of senses in the unfolding of semantic change:

"from a diachronic point of view, this means that semantic changes may take their starting-point on the extensional level just as well as on the intentional level, or in the domain of encyclopedic information just as well as in the realm of semantic information" (Geeraerts 1999 : 96).

Under a descriptive angle, semantic change can unfold via the modulation of core cases, in which change unfolds as the expansion of the prototypical core at the extensional level within a referential range. This corresponds to the process of extension seen in other typologies, under the light of a prototype theory approach. The saliency of the senses of a lexical meaning is redistributed. Semantic change may also manifest as a *polygenesis*. Polygenesis happens when idiosyncratic innovations emerge several times disconnectedly. These spontaneous innovations are rooted in the individual's semantic knowledge, their knowledge of etymology (even at an intuitive level) and the word's history of use, which leaves meaning traces through contexts of use and associative networks, that are stored in the individual's semantic memory. The blurred limits of categories make it possible for new meanings to emerge at the idiosyncratic level. Polygenesis also accounts for transient meanings. Semantic change can also be based on the evolution of the structural and functional elements observed in synchrony, namely the structure in radial sets, family resemblance, and the leveling into peripheral and central senses with varying structural weights.

The theoretical view of meaning as being organized into subsets is central to this work which relies on a model that works with subsets. However, these subsets are constituted differently. Working with subsets is a way to include the polysemous nature of words with a microscope view into word structure, how it is organized and how it interacts with other word structures and the environment. When new subsets enter existing structures, they tend to cluster with existing ones, connecting to each other via mechanisms, like metaphor, metonymy, and the like. The idea of clustering is also central here, since it is one of the implemented tools used in the case studies.

Subsets and clusters of subsets rendering the organizational complexity of polysemies, is what, according to Geeraerts (1997: 27-28), traditional and structuralist semantics lacks:

" ..traditional and structuralist semantics alike tended to neglect the internal, semasiological structure of lexical items (in the sense, that is, that they paid relatively little attention to phenomena such as the clustered nature of polysemy and the difference in salience of the various readings of an item)".

Defenders of structuralist semantics, however, raise an interesting point in criticizing the suitability of a prototype theory framework for the study of semantic change. Rastier (1999) raises the questions of how new prototypes emerge. By doing so, he points out that if the changes in the system only expand from the existing system, how are new concepts corresponding to new categories integrated? We seem to come across a problem similar to Saussure's. This seems to be a theoretical limit of the aforementioned approaches in prototype theory.

1.1.4.2. Metaphor theory and Historical pragmatics

In cognitive semantics, the speaker/hearer paradigm is central, as are subjectivity, efficiency and expressivity principles. Two main lines of analysis emerge: the importance of metaphor, being the most powerful process of conceptual mapping, and the importance of the pragmatic dimension, at the heart of innovation and diffusion.

1.1.4.2.1. *Metaphorical and metonymical mapping*

Most researchers admit that the most powerful processes among all the cited rhetorical figures in action in semantic change are metaphor and metonymy. Whole theories of semantic change are based on the centrality of metaphorical transfers, notably Traugott and Dasher's (2002) and Sweetser's (1990) diachronic metaphorical analysis.

On the theoretical grounds laid by Lakoff (1987;1993) and Lakoff and Johnson (1980), change rooted in metaphor is described as the "analogical mapping of a more concrete term from a "source" domain onto a more abstract term in the "target" domain" (Traugott & Dasher 2002: 75). The notion of "domain" is the major key to this approach, as it is slightly different from what is traditionally meant by a "semantic field." Lakoff's now famous example "ARGUMENT IS WAR" illustrates this process of source-target domain transfer. As

speakers use vocabulary from the WAR semantic field (such as *defend one's position*, *attack*, etc.) to describe ARGUMENT's processes, they rely on conceptual metaphor, otherwise referred to as analogy, in other theoretical frameworks. This process is not reduced to its synchronous aspect, as it generates innovations based on the same mechanism. This view is expanded upon in theories of conceptual blending (see Fauconnier and Turner 2003), which are based on the subconscious associations we make on this basis.

Sweetser (1990)'s diachronic metaphorical analysis relies on a threefold view: she gathers polysemy, lexical semantic change and pragmatic ambiguity phenomena, while looking for rules (regularity) in terms of mapping. In a study of metaphors of perception, she describes the interaction between the objective (intellectual) domain, the interpersonal communication one and the subjective (emotional) one. The objective, intellectual domain is itself divided into physical manipulation, mental manipulation, and sight, subdivided into knowledge and control. For instance, a word like *to grasp* is anchored in a physical domain (objective) and is mapped onto the mental domain in its sense "to understand." The following diagram illustrates this:

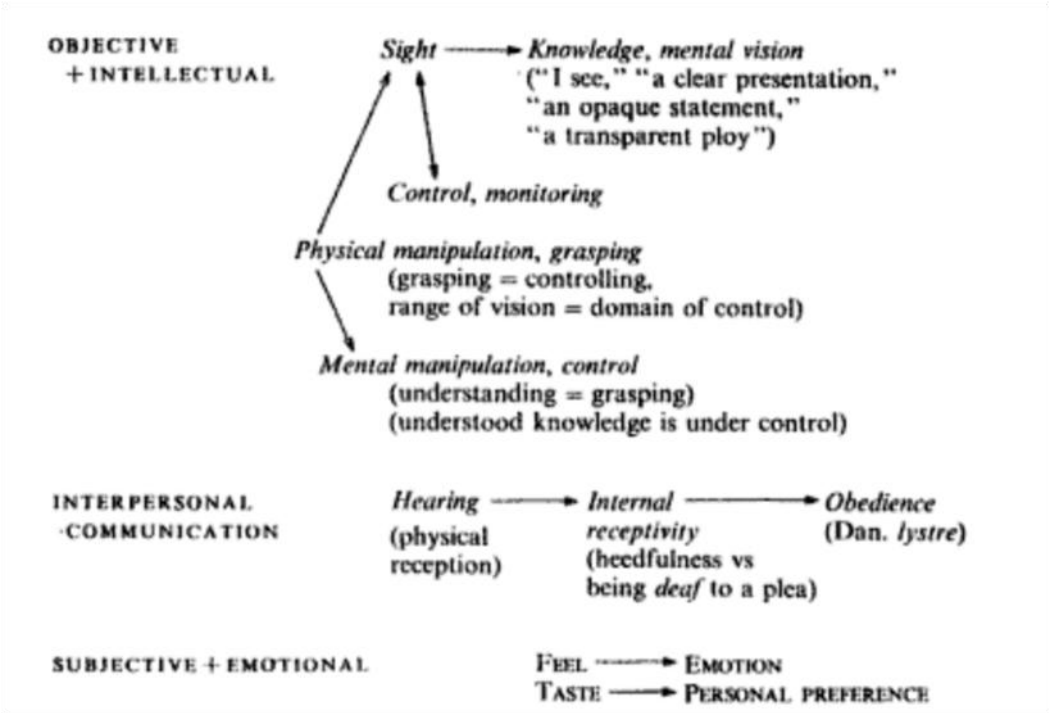


Figure 3 Diagram: the structure of our metaphors of perception, taken from Sweetser (1990: 38)

1.1.4.2.2. *Historical pragmatics & Discourse Analysis*

The physical to mental paradigm is also acknowledged by Traugott and Dasher (2002), giving examples such as OE *felan*[38] "touch" > "experience mentally" (2002: 95), grouping these types of items under the heading "tendency I" in the following classification:

> "Tendency I: Meanings based in the external described situation > meanings based in the internal (evaluative/perceptual/cognitive) described situation.
>
> Tendency II: Meanings based in the external or internal described situation > meanings based in the textual and metalinguistic situation.
>
> Tendency III : Meanings tend to become increasingly based in the speaker's subjective belief state/attitude toward the proposition " (Traugott and Dasher 2002: 95 citing Traugott 1989: 34-35).

Traugott and Dasher (2002) put forward an invited inferencing theory of semantic change in the framework of Historical pragmatics and Discourse analysis in which the main mechanism of change is claimed to be subjectification. They try and shed light on the regularity of semantic change patterns. "Invited inferences" or inferences generated within conversations or interactions can become "generalized invited inferences" when spreading. The "invited inference" takes place when a speaker/writer invites an addressee/reader to infer an implicature. Meaning is negotiated in that process and then may be conventionalized in a speech community via repetition and diffusion. The invited inference then becomes a generalized invited inference; it subsequently becomes conventionalized and finally semanticized. The main mechanisms within interactions that trigger semantic change are metaphor and conceptual metonymy. The authors apply this theoretical point of view to cases across Japanese, English and French mainly, in a cross-linguistic perspective. This approach unveils discourse mechanisms as well as linguistic typological ones, and therefore bridges social and linguistic dimensions. It contends that semantic change is rooted in the interaction between language in use and the linguistic system/structure, rather than maintaining two theoretical entities without true dynamic interaction as most structuralist theories do.

[38] Giving *to feel.*

However, the authors put a strong emphasis on grammatical phenomena, supporting the idea of unidirectionnality, whereas this work focuses on meaning change rather than meaning change via grammatical status change.

1.1.5. Discussion: Which theoretical framework is suitable for corpus exploration in short diachrony?

Semantic change has given birth to numerous classifications. Indeed, since linguistics has thrived to establish the discipline as a science, it has to abide by the rules of science, and "science is the systematic classification of experience" (Lewes 2004: 4). As Jespersen (1922: 388) puts it : "man is a classifying animal."

However, as we get caught up in the fastidious work of classifying, we sometimes tend to forget the original reason behind classifying and its consequences in the way we view knowledge. Moreover, language itself is a classification of the world by nature:

"Language is legislation, speech is its code. We do not see the power which is in speech because we forget that all speech is a classification, and that all classifications are oppressive" (Barthes 1979).

While trying to classify language phenomena, we thus come across an issue of circularity created by the use of metalanguage. In fact, we are classifying a classification.

In discussing the fine details of the nature of elements in a structure and the nature of their interactions, we sometimes forget to stand back and look at the "big picture." However, it is my contention that the activity of classifying, modeling, and ultimately understanding are complementary and need to be fed by each other. The interaction between these levels within a holistic paradigm is acknowledged by Coseriu (1958:16):

"La descripción, la historia y la teoría no son actividades antitéticas o contradictorias, sino complementarias, y constituyen una única ciencia[39]"

[39] "Description, history and theory are not antithetical or contradictory activities; they are rather complementary and are a single science."

One may also wonder whether there is a truly objective theoretical framework, among all the cited approaches, that prevails. I agree with Eagleton's (1983:201) claim that "objectivity is just whatsoever questionable interpretation of things has currently seized power." Even though many typological, structural and functional approaches bring precious insights as to the workings of semantic change, none of these approaches prevail enough for one to decide to work exclusively with one of them. Therefore, I have chosen an a-theoretical approach as much as possible, while keeping in mind the mechanisms described in the literature. Non-theoreticality is, of course, impossible to achieve by definition and is limited by the use of a model, itself a reflection of a theoretical view. The model I have chosen to work with makes a few assumptions about meaning: meaning is continuous, it is organized in subsets (rendering polysemy and variation in use), and these subsets may be analyzed as clusters which are groupings of them. However, in the model, both the constitution of subsets and clusters is adaptable in terms of sources as well as structural profile. Adaptability provides great flexibility on the one hand, but bears the risk of over-oriented user setting choices, and therefore enhanced subjectivity.

The main idea of this work is to observe semantic change directly as it unfolds. The capacity to do so has remained tightly linked with technological advances and it is only recently that we have become able to research such questions at a large scale. For linguists of the beginning of the century, such observation was unthinkable with their resources, as is stated by Bloomfield (1933):

"the process of linguistic change has never been directly observed;…such observation, with our present facilities is inconceivable."

To look at this process it is useful to characterize the major changes of states of a word undergoing changes in meaning and status in a temporal dimension. It is therefore useful to rely on a division of the process in three steps, as stated by Lüdtke (1999 : 50):

Step 1: OUTSET [innovation/creativity]

Step 2: INTERMEDIATE [diffusion/imitation]

Step 3: OUTCOME [result/difference]

The outset step corresponds to innovation in a speech act and can be a completely idiosyncratic innovation, the intermediate step corresponds to the repetition and diffusion of the innovation by a larger number of speakers, and the outcome step is the result, that is to say a neologism or meaning shift. What lexicologists call "the neological feeling" (fr. "sentiment de néologie") or speaker's feeling of "newness" takes place in the second step, while in the third step the innovation has completely entered "ordinary language" in such a way that new generations learning and using the new term will not experience any such feeling of newness. I have chosen to focus more specifically on the second step, namely diffusion, as it unfolds, and as it enters textual production. The choice of working with corpora corresponds to the need of working with real data, even if the whole of reality may never be represented properly. I assume here that working with a sample of language that has really been produced may provide more realistic analyses than working with a limited representation such as a dictionary, a thesaurus or an ontology (or even a theoretical model about those). Text allows for accessing the syntagmatic, the conceptual, the historical, and the cross-linguistic dimensions that form the essential framework to study language and meaning. The syntagmatic level corresponds to the linguistic context, itself connected with the semantic context, in the conceptual dimension. In turn this semantic context is connected with the factual context of history, which is in turn connected to the other relative historical contexts of other cultures and languages. To round the circle, this culturally relative context is also different at the linguistic level, as similar concepts across languages do not have the same semantic networks and do not have the same metaphorical, compositional and dependency constraints. In my view, to try and talk about meaning at the collective level, we have to understand its holistic grounding in language, concepts, history and culture. Those vast domains are different contexts in which meaning is grounded and through which meaning exists. At the individual level, however, other factors come into play. Intuition is a fundamental key to how we perceive language, and quite often researchers working on learned languages need to rely on native speakers' intuitions to validate theirs. Our mother tongue(s) do shape our understanding of language, since it is the first structure and system we learn, and the one we have the most anchored contextual information available in. In addition to the intuitive level, words and combinations of words acquire an emotional connotation for each speaker. Some of the emotional content may be collective (as in *World War II*, for instance), and another share of it may be restricted to communities of speakers or smaller social groups. However, there remains an emotional "note" or "touch", in synaesthetic terms

(embodied) that is strictly individual, and associates the emotions of the plethora of experienced situations attached to a word progressively to one's semantic knowledge. At the level of imagination, the associative networks we develop individually are all different in essence even though they have some common features and overall comparable structure. Concept and idea association, via analogy, emotional echo, or personal fantasy is a phenomenon comprising immense variation. Semantic (including encyclopedic) knowledge, taken here at its widest definition, is therefore as varied as the number of human beings in existence, and cannot be reduced to a set of features.

Although it is useful to group phenomena in classifications and try and grasp the organizational structure within a classification, browsing text for semantic change *at large* with no established structure beforehand, may shed light on the processes that several phenomena have in common and/or how they interact. Since no process may operate in total isolation from its environment, detecting different processes together may help show how they are related and connected. With an exploratory approach based on empirical data, one may observe mechanisms of change at different levels of granularity. For instance, if a neology based on a morphological derivation is detected, it is intrinsically related with the source word of which the meaning may be impacted, and with the other existing derivations, as well as with all the other words containing the same derivational morpheme, etc… It is moreover connected with the words it collocates with, as new collocations may emerge or existing ones undergo connotational and/or semantic drift. A metaphor for this is the image of "navigating" these networks to try and grasp the geography of meaning and all its routes. While doing so, we have at our disposal all the existing theoretical distinctions between the different types of meaning change and their structure, and may check whether these theories apply to the data at hand or not. Nevertheless, there are some cited theoretical positions that must be questioned since they turn out to limit the investigation:

"Deux principes négatifs semblent déjà acquis : dépasser la théorie du signe et son postulat d'une correspondance simpliste entre signifiant et signifié, et reculer au-delà du mot graphique les frontières de pertinence de l'unité lexicale. Le néologisme de sens n'est rien sans

ses règles d'insertion lexicale dans la phrase et/ou le syntagme ; il n'est rien non plus sans le discours — ou l'interdiscours — où il prend son sens. [40]" (Bastuji 1974:7)

Bastuji (1974) insists that semantic neology is detectable and understandable in context, be it the direct context of the sentence or paragraph, or the larger context of utterance. I share her understanding of semantic neology as a specific type of polysemy. Looking at the dynamics of polysemy in time and context is precisely the approach I have chosen to pursue.

[40] "Two negative principles seem to have already been assimilated: going beyond sign theory and the assumption associated to it, that there is a simplistic correspondence between signifier and signified, and pushing the borders of relevance of the lexical unit beyond the graphical word. Sense neologism is nothing without its rules of lexical insertion in the sentence and/or the phrase, it is also nothing without the discourse –nor the interdiscourse- from which it takes on its meaning."

Chapter I.2 From theoretical to computational models:

Statistical semantics in the context of social, technological and scientific paradigm shifts

In the past decades, we have witnessed a major shift in linguistics from theory to data-based investigation. Data has become central to the discipline. The growing constitution of corpora and databases in linguistics witnesses this shift. The theoretical approaches described in the first chapter have brought useful frameworks to analyse semantic change but these frameworks have most of the time been applied to long diachrony issues with a theoretical perspective. Most of these works rely on case studies with an attempt at generalization limited by manual analysis. Case studies give birth to generalizing hypotheses, since mechanisms observed for a word may apply to large groups of words. Types of mechanisms as well as their structure and relationships can be defined. This is what typologies try to do by classifying a series of mechanisms under the same umbrella. However, to browse data at a large scale, manual analysis soon becomes overwhelming. With the advent of new technologies, and its impact on statistical and computational linguistics, purely theoretical approaches have been enriched and modified by the integration of statistical and modeling tools. Two aspects have changed: on the one hand, new technologies make language a different object of study and this object itself is at the center of new needs, and on the other hand the methods to study it have also evolved with new scientific paradigms. Therefore, the object of study has changed, and the methods to study it have changed. Moreover the relationship between academic research and the general public is also changing, as bridges between the two are created within that new paradigm.

I first look at the changing context and methods, encompassing the issues that a changing technological context implies. In the second part, I look at applied statistical works that deal with semantic change and neology, both in the academic realm and outside of it, since the said context change created new bridges between the two. This aspect includes semantic change in a wider frame of text and discourse analysis. In the third part, I look at the first semantic change modeling attempts, by pioneers in the field, who anticipated the current rise of models that will be described in Chapter I.3.

As we are faced with a major paradigm shift both regarding the nature of language and regarding the role of technologies in our lives, at the sociological and scientific levels, the need for computational tools and models arises. Models provide organized representations of data and the possibility to manipulate substantial amounts of it. The need for a representation of language has been acknowledged by thinkers long before that paradigm shift. Indeed, the question of language representation is vast and has been raised by philosophy, philosophy of mind, psychology, linguistics, semiotics and cognitive science, as well as all visual disciplines in graphical design, digital technologies and arts. It is intertwined with the question of conceptual representation.

For Aristotle, it is through representation that we are able to organize the world to make sense of it. For Plato, however, representation brings the risk of illusion. For the philosopher Locke (1959) conceptual knowledge is visually based by nature. However, that assertion is far from being taken for granted and has generated a great amount of research notably in cognitive science.

In cognitive science and neuropsychology the perceptual and motor anchorage of conceptual knowledge in brain processes has been thoroughly studied. Brain imaging studies claim that the processing of word-meanings is distributed in the brain (see Binder et al. 2009 for a critical review). Hence, in studies using words as stimulus material, visual, gustatory, auditory, olfactory, and action related words trigger activities in the brain regions involved in processing these sensations. Consequently, the need for a language representation may also arise from the fact that conceptual knowledge processes are tightly intertwined with perception and action.

The artificial intelligence field of knowledge representation (KR) and formal semantics, among others, cover the issue of how to represent or "code" conceptual knowledge. In a way, the need for the representation of language is deeply linked with the need for a representation of concepts that one may work with. However, in these approaches we keep creating more languages, or meta-languages, to represent language itself, and fall into the retroactive trap of representation.

Models are a possible answer to the need for representation. However, one must distinguish between two types of models: The "models" developed before the advent of technology in

linguistics are theoretical and are rather "frameworks" of analysis, classification, categorization and interaction. The second type of models corresponds to the tools used in sciences that rely on mathematical simulation, such as climate studies. These "models[41]", are mathematical paradigms of abstract representation. Meanwhile the underlying conceptual foundations of these two types of models are similar.

Modeling language implies treating it as a collection of signs, features or numbers, and getting away from its abstract nature. But language cannot be treated only as a countable whole, as notes Milner, citing Plato:

" La langue, même si on l'imagine totalité dénombrable, est aussi nécessairement marquée d'hétérogénéité et de non-superposable [42]" (Milner, 1978:29 quotes Plato, Sophist 262 a).

Milner refers here to the need for categorization at a grammatical level. The need for a distinction between the different parts of speech is confronted with the heterogeneous aspect of language. For language to be dissected into a finite set of elements it has to be perfect, however it is not. But later, Milner also explains that reality is conceived as something that can be represented and that language pertains to reality. According to him, our opinions about language are intrinsically linked with our perception of reality:

"Pourtant ce sont bien des thèses touchant le statut de ce réel qui sont en jeu dans les divers discours tenus sur la langue ; la partition majeure se laisse résumer ainsi : *le réel est conçu comme représentable ou non*[43]." (Milner, 1978:30)

Milner states that in our need for representation, we rely on its structure in networks, formed by repetition:

---

[41] I use the term "model" to refer to the second type hereinafter.

[42] "Language, even if we imagine it as a countable whole, is necessarily marked heterogeneity and non-superimposableness"

[43] "However, it is clearly theories regarding the status of reality which are at stake in different discourses concerning language; the main score can be summarized as follows: reality is understood as representable or not."

"Ce que le sujet demande au réel (…) c'est qu'en quelque manière une représentation soit possible : à ce prix seulement, par quoi l'imaginaire le rançonne, le sujet pourra supporter ce qui, de soi, lui échappe. A cela, deux conditions : que pour le sujet il y ait du répétable, et que ce répétable fasse réseau.[44] " (*ibid*: 30)

This second view encompasses more than language in that it questions the representation of reality, but it is also applied to language as a part of reality. Those two quotes are representative of two aspects of language that are both intrinsic to it but are difficult to deal with conjointly within a unique approach. On the one hand, language is ungraspable due to its heterogeneity. On the other hand, language may be reduced to a mathematical-like system where what is observed is the structure created by repetitions, patterns and forming networks. This second aspect is attached to the foundations of modern linguistics, since structuralism, Chomskyan theory (as developed in (Chomsky 1965), for instance) and all computational approaches are rooted into it. The first aspect, however, is the building ground of literature and poetry. (However, literary and poetical analysis includes some of the second aspect in that they use structural methods). This dichotomy has been observed by many linguists, notably by Sweetser (1991: 12):

"work in linguistics has tended to view semantics in one of two divergent ways: either meaning is a potentially mathematizable or formalizable domain ..., or meaning is a morass of culturally and historically idiosyncratic facts from which one can salvage occasional linguistic regularities"

Bridging those two aspects is central to this work, as semantic change is deeply rooted in the second but may highly benefit from the first.

When Milner talks about repetition and the necessity for this repetition to create a network, he says so at an epistemological level and not in the same way that a computational linguist

[44] "What the individual asks of reality (...) is that a representation be possible in a way: and only at this price, through which the imaginary captures it [reality], the individual will bear what part of themselves is foreign to them. For this to happen, there are two conditions: there must be something repeatable for the individual, and this repeatable thing must become a network."

would do. Nevertheless, this assertion is valid beyond its epistemological rooting and may be understood at the level of computational linguistics. If Milner evokes the need for reality and language to form the abovementioned network in order for them to be bearable through the level of representation, at a philosophical and psychological level; this assertion may be transferred to the discipline of computational linguistics and more largely to systemics. In turn, this will make us reflect on the philosophical status and nature of the items and frameworks we use to analyse language in this field. What is the meaning of repetition? What is a network? At what stage can we say there is a pattern, and at what stage can we say that a pattern has changed? This need for representation and the idea of using frequency patterns and networks are at the center of a shift in the treatment of semantics. This shift is rooted in changing methods but also in a changing social and technological context, giving birth to new needs and new methods.

1.2.1 Changing context and changing methods

1.2.1.1. Changing context and immediacy issues: Adapting to language paradigm shifts

1.2.1.1.1. *New paradigms of diffusion*

As stated in the introduction of this chapter there is more than one paradigm shift in the advent of technology in the present context. For many sociologists, we are witnessing the transition from an industrial to an information society. At the heart of this transition, new modes of communication, such as the Internet and mobile phones, play a central role.

"Today's agenda concerns the Internet especially, the 'information superhighway' and cybersociety brought about now by information and communication technologies (ICTs). Hot topics now are electronic democracy, virtual relations, interactivity, personalization, cyborgs and online communities. Much comment now seizes on the speed and versatility of new media to evoke the prospect of radical transformations in what we may do." (Webster 2006:3)

Beyond the introduction of new tools and methods, language itself is deeply impacted by these changes. With the Internet, the idea of a user-created language took on a strong rooting. Crystal (2005) talks about the advent of new Internet linguistics but the shift is even stronger, modifying people's relationship to language and its rules. While Crystal outlines the differences between forms of language on the Internet and written and spoken language, as

well as stylistic innovations to adapt to the medium, he does not question how the relationship we have with language is modified. Rather, he states that, as a new form of language emerges, a new branch of linguistics should arise as well, called "Applied Internet Linguistics". The Web is also a source of corpora for academics and provides multilingual resources to build corpora quickly and cheaply:

"For many languages there are no large, general-language corpora available. Until the Web, all but the richest institutions could do little but shake their heads in dismay as corpus-building was long, slow and expensive. But with the advent of the Web it can be highly automated and thereby fast and inexpensive." (Kilgarriff et al. 2010)

The language of the Internet and mobile communication is only in the process of being characterized. This process gives birth to coinage attempts like "Netspeak", for instance, found in (Altmann, Pierrehumbert and Motter 2010).

The most explored corpus in this work *Le Monde* covers the period from 1997 to 2007. I observed that across various word frequency graphs that there is a peak of use and creativity in the period covering 1999 to 2002. This period corresponds to mass access to the Internet in France correlated with the introduction of broadband technologies and the decreasing price of access. This context may explain these sudden peaks since the multiplication and acceleration of international exchanges play a role in creativity.

1997	1998	1999	2000	2001	2002	2003
2 479 409	3 696 826	5 361 268	8 448 514	15 635 144	18 039 953	21 750 456
2004	2005	2006	2007	2008	2009	2010
23 723 605	26 149 202	28 767 899	40 807 003	43 891 158	44 697 966	50 292 729

Table 5 Number of Internet users in France from 1997 to 2012. Years in red are the years covered by the corpus *Le Monde*. Source : http://donnees.banquemondiale.org/[45]

45 Cited Source : « Rapport sur le développement des télécommunications/TIC dans le monde et base de données de l'Union internationale des télécommunications. »

The introduction of mass access to the Internet, high speed communication technologies, followed by mobile technologies, at wide scale in Western societies, seem to be one of the keys to understanding recent language change. These technologies placed immediacy and international contact at the heart of the communication paradigm in the years 2000.

As a consequence of immediate access to information, the paper dictionary is now considered an almost obsolete object, as people will search the Internet to check a word's meaning. For many (non-linguist) people, if a word is found by their search engine, it does exist. The fact the dictionaries do not attest it seems to have lost authority.

Beyond electronic versions of established references like dictionaries and the mainstream press, Internet and mobile technologies offer blogs, forums, RSS feeds, social networks, and instantaneous tools like Twitter. Twitter may be the utmost expression of communicational immediacy, impacting the diffusion of information at a deep level, by reducing diffusion time to the instantaneous. Information and language are in a strong relationship, even though it cannot be asserted that they equal each other. Nevertheless, the newly acquired status of language through immediate diffusion makes it the most powerful informational tool at our disposal. TV programs are now integrating Twitter-based information to quickly evaluate people's opinions on a given subject (of course, the fact that people who do not use Twitter are not taken into account make these opinion evaluations biaised). In this context, new meanings, when they are put through the Internet, have the ability to reach millions of people in a second. Therefore, the usual time frames set out by long diachrony studies may not apply any longer. The impacts of this paradigm shift have not yet been analyzed in full as they raise new questions for everyone, including sociologists, philosophers, historians, politicians, and linguists. In semantics, one of the impacts that clearly stand out is the need to work with smaller time-frames and therefore study short diachrony, complementarily to long diachrony. We are now faced with immediacy issues and user-based issues, calling for new adapted approaches. It is also worth noting that the linguistic policies emerging from governmental bodies[46], trying to regulate the use of English borrowings and new words are overtaken by the dazzling authority of the Web. What is the value of a word meaning used all over the internet

[46] See the *Commission générale de terminologie et de Néologie* in France : http://www.dglf.culture.gouv.fr/dispositif-enrichissement.htm

but not appearing in any official dictionary? How long does a word have to be used online to consider it has entered usage? What is the status of ephemeral meaning changes?

The idea of a semantic Web has been around for about a decade, aiming at making Internet search more grounded in the complexities of meaning, beyond basic keyword match. Recently, the leading internet company Google has started to implement semantic functions in its search engines, mostly integrating disambiguation algorithms which allow the user to distinguish between *Apple*, the brand, and *apple*, the fruit, for instance (see the numerous press articles on Google's semantic search systems, such as Newman 2012 and Anon 2012).

On intuitive grounds, it is generally accepted that language change accelerates with the evolution of the Internet due to the new type of diffusion language undergoes. It is also intuitively accepted that language change accelerates. Nevertheless, it is worth pointing out that this feeling has been acknowledged by several generations. For instance, Bréal states in 1899 :

> "Dans nos sociétés modernes, le sens des mots se modifie plus vite qu'il n'avait coutume dans l'antiquité et même chez les générations qui nous ont immédiatement précédées.[47]" (Bréal 1899: 116)

However, the mechanisms of change within that new diffusion paradigm are a largely unexplored topic. Therefore, I will attempt to lay tentative foundations to deal with this essential topic. If the topic has not received enough interest in the scientific community yet, the large number of tools to explore it, however, has rocketed, with the advent of statistical methods for text analysis, boosted by rising computer power.

1.2.1.2. Changing methods: statistical methods for text analysis.

As a consequence of the paradigm shift described above, both regarding the nature of language and its diffusion, the methods to analyse semantic change have to include these new aspects to be in tune with the object of study, and the society which generates and is impacted by it. Consequently there is a growing use of statistical techniques based on co-occurrence

[47]In our modern societies, the meaning of words changes faster than it used to in antiquity and even in the generations that directly preceded us

and modeling techniques. The term *statistical semantics* has gained usage since the publication of Furnas et al. (1983), one of the earliest articles coining this subfield's name. A growing number of conferences in the area attest to its growth, and the emergence of *computational semantics* (see the conferences and workshops such as GEMS, DISCO, IWSC and the ESSLLI workshops[48] among others) and corresponding publications in the area, in journals such as *Computational Linguistics*).

I first look at statistical methods not using a mathematical model. These methods have mostly flourished to conduct corpus studies, and are based on word frequencies, word co-occurrence frequencies, and distribution issues.

1.2.1.2.1. *A renewed interest in corpus*

Computers brought the possibility to store and browse greater amounts of data as well as the possibility to implement statistics on large sets of data. This naturally led to a renewed interest in working with corpora. The growing introduction of statistical methods into linguistics is tightly linked to the convergence of linguistics and Natural Language Processing (NLP). NLP combines Linguistics, Computer Science and Artificial Intelligence. One of the elements that bridged the two was corpus. The interest in corpus rose again much before statistical methods became widely adopted in linguistics. NLP was more concerned with rule-based methods, but a major debate opposing the advocates of rule systems against those of 'real' data analysis brought corpora to the forefront again. There has been corpora at the scientific community's disposal since the 1960's, however these corpora now get bigger and bigger and may contain more information, of syntactic, semantic, morphological and stylistic nature. For Habert, Nazarenko and Salem (1997), the renewed interest in corpus is also due to the support of the NLP community which needed language resources to go beyond rule-based systems:

"La tradition des linguistiques de corpus a reçu ces dernières années un appui vigoureux et inattendu de la communauté du TALN qui a donné un nouvel essor à la

[48] GEMS (2009, 2010, 2011) : Geometrical Models of Natural Language Semantics, workshop in conjunction with the Annual Meeting of the Association for Computational Linguistics (ACL). DISCO (2009): Distributional Semantic Beyond Concrete Concepts, workshop in conjunction with the Annual Meeting of the Cognitive Science Society (CogSci). IWSC: International Conference on Computational Semantics (ACL). ESSLLI: European Summer School on Logic, Language and Information. Workshops: Compositionality and Distributional Semantic Models (2010), Workshop on Distributional Lexical Semantics (2008) … among others.

constitution et à l'utilisation de corpus annotés. (…) Il s'agit en fait d'un changement profond de paradigme. Jusque-là, l'objectif des recherches en TALN et en Intelligence Artificielle était avant tout de « modéliser », de formaliser le savoir humain, de dégager les règles sous-jacentes." (Habert, Nazarenko, & Salem 1997: 10)[49]

Therefore disciplines which were working with separate paradigms joined in on the use of corpora, thereby confronting the findings of historical linguistics, philology and semantics with NLP. This joint adventure, however, is still in its infancy, demanding the creation of common grounds, as scientists from these disciplines may have different points of view on the nature of language.

1.2.1.2.2. ***Word frequency and distribution***

While looking at a corpus, we are looking at series of words organized in a system that have varying frequencies and co-occurrence frequencies. The distribution of these frequencies is influenced by a few factors: First the available number of words in the given language, second, the frequency profile of these words: are they used often or not?, third, the patterns that most texts tend to show in this distribution and may be considered as a type of norm, fourth, the specific distribution pattern of the corpus owing to its genre and style.

To detect what words vary beyond a "normal" threshold from a text, we first have to characterize the text itself and the frequency profile of a word in the given language at stake.

Some words are more frequent than others in language. Among the available vocabulary in French, the "Académie Française" considers that people use effectively between 1 500 to 3 000 words on average, in daily spoken language, but we find about 60 000 words in a dictionary like *Le Petit Robert*, or *Le Larousse* and about 100 000 in *Le Trésor de la langue française*. These figures are roughly equivalent in English and Spanish, with *The Oxford English Dictionary* containing 171 476 words in current use, and 47 156 obsolete words[50] and

[49] "In recent years, the tradition of corpus linguistics has received vigorous and unexpected support from the NLP community which gave new impetus to the creation and use of annotated corpora. (...) This is a profound change in paradigm. Until then, AI's and NLP's research goal was primarily to "model," to formalize human knowledge and to identify the underlying rules."

[50] Source: http://oxforddictionaries.com/words/how-many-words-are-there-in-the-english-language

the *Dictionario Real Academia Española* containing approximately 90 000 words. Within that available source of vocabulary some words tend to enter use as others stay in the shadow, and are only known by specialists in various fields.

If frequency is a good indicator of word use, it is also influenced by invisible rules according to which some words are more frequent than others. The distribution of higher and lower frequencies in language constantly evolves as some words become more fashionable and others obsolete. However, there are some strong tendencies that may be grasped. Some recent dictionaries (such as the French dictionary Antidote[51]) and databases (such as Lexique[52] and Brulex[53] in French and the MRC[54] in English) include a frequency rating for each word. Frequency ratings have also been taken into account in language-based games like Scrabble. The MRC, a recognized psycholinguistics database gives average frequency ratings for English words, on the basis of a 1 50 837 words corpus, as can be seen in Table 6 , an excerpt from the publicly available MRC database:

69991	the
36472	of
28910	and
26235	to
…	…
861	state
850	those
849	people
840	too
839	mr
837	how
834	little
832	good
816	world

Table 6 Frequency ratings in hierarchical order of a few English words, taken from the MRC database

[51] http://www.druide.com/a_dictionnaires.html

[52] http://www.lexique.org/

[53] http://lcld.ulb.ac.be/outils/brulex

[54] (Wilson, 1988)

Function words occupy the top rows of the table, with plain words showing much lower frequencies. Some computational semantic studies simply get rid of function words in their data via the use of stop-word lists, while others choose to keep them, since they may play a role in the studied phenomenon. In the course of this work, I have decided to keep them as will be detailed in Part III.

Other similar tools are available including genre statistics. Davies[55] established comparative tables of frequencies across sub-corpora of different genres in the Corpus of Contemporary American English (COCA, 425 million words, 1990-2011)[56] and shows that the distribution substantially differs across genres. For instance the word "all" shows higher frequencies in spoken language, but "teacher" has a higher profile in academic texts, or "win" in newspapers, as shown in Table 7.

This may seem obvious but must be kept in mind as we explore corpora, since numerical and statistical data are influenced by these factors.

rank	lemma	PoS	DISPERSION	TOT FREQ	SPOKEN	FICTION	POP MAG	NEWSPAPER	ACADEMIC
				402,377	81,69	78,752	83,275	79,368	79,292
221	all	r	0,93	177317	78464	46018	21778	17690	11880
361	win	v	0,93	111478	25724	7987	22158	48277	7092
371	teacher	n	0,88	116100	7578	7560	10163	14733	75912

Table 7 Excerpt from the publicly available frequency distribution of the COCA

Retrieving the frequencies of words and co-occurrent words using statistical methods has been conducted on corpora in corpus linguistics and computational linguistics for data mining tasks, information retrieval, summarization, sentiment analysis, concept extraction and clustering as well as other NLP tasks. There are numerous available software solutions to do so, among which is the outstanding tool R (see for instance Baayen 2008).

[55] Professor of Linguistics at Brigham Young University, Provo, Utah, USA

[56] Genres: spoken, fiction, popular magazines, newspapers, and academic. Source : http://www.wordfrequency.info

1.2.1.2.3. *Frequency and change*

Pagel, Atkinson and Meade (2007) compared two hundred meanings in four corpora in English, Spanish Russian and Greek, in an Indo-European long diachrony perspective and found that:

"…frequently used words evolve at slower rates and infrequently used words evolve more rapidly." (Pagel, Atkinson, & Meade 2007)

To achieve this conclusion, they studied the rate of replacement of cognates (words of similar meaning and etymological root across languages within a language family, here Indo-European). They claim that the frequency of word use is a basis for predicting this rate of replacement. The frequency distribution of cognates across the four languages is almost identical, allowing for a theory that encompasses Indo-European languages rather than a language-specific theory. The method is summed up as follows:

"We estimated the rates of lexical evolution for 200 fundamental vocabulary meanings in 87 Indo-European languages. Rates were estimated using a statistical likelihood model of word evolution applied to phylogenetic trees of the 87 languages. The number of cognates observed per meaning varied from one to forty-six. For each of the 200 meanings, we calculated the mean of the posterior distribution of rates as derived from a Bayesian MCMC model that simultaneously accounts for uncertainty in the parameters of the model of cognate replacement and in the phylogenetic tree of the languages" (*ibid:* 717*)*

This makes sense since rarely employed words, when they come to be used, have a vague meaning for most people and thus are more likely to undergo substantial modifications. Frequently used words, on the contrary, benefit from a stronger shared meaning and therefore may not undergo drastic meaning changes easily.

1.2.1.2.4. ***Co-text & Context***

Various expressions are found in the literature to talk about the context of an occurrence: "context", "linguistic context" and "co-text" (sometimes spelled "cotext"). There can be confusion about what these terms refer to precisely, since "context", at the widest definition, can refer to context of utterance, factual circumstances or linguistic context. Moreover, there are contextual levels, starting with the neighboring string of words of a target word, all the

way to compositional, syntactic, utterance, text and corpus contexts, as well as the extra-linguistic context. The expressions of "linguistic context" and "co-text" seem to include these levels in different ways, according to the chosen theoretical perspective.

Working with the linguistic context or the co-text is a way to add a pragmatic dimension since it gives a picture of the word *in use*. The general idea behind using context to analyse a target word is that "we know a word by the company it keeps"[57], a motto relayed in the distributional semantics community to justify the choice of working with context as will be detailed in Chapter I.3.

The co-text, or linguistic context, is the direct environment a word appears in; it is the contiguous text to it. It is not definitional of a word, however, it shows:

-the type of use the word undergoes with its pragmatic implications,

-the polysemy of the word depending on the number of clearly different meanings it can endorse,

-compositional aspects, and

-the idiomatic uses of a word.

Reteunauer's (2012) doctoral thesis distinguishes between the local and global co-text. This distinction is sometimes made by other authors using the term "linguistic context" for the local aspect and "context" for the global aspect. Even within the idea of "local" context, definitions vary:

"L'environnement local correspond au voisinage proche de l'unité lexicale ciblée. La définition précise de ce qu'on entend par "voisinage proche" varie selon la perspective adoptée. Le voisinage proche peut se délimiter sur les plans lexicographique, syntaxique,

[57] Saying attributed to Firth (1957: 11)

énonciatif ou typographique, dont chaque unité propre est susceptible de correspondre à une unité sémantique.[58]" (Reteunauer 2012: 11)

However, linguists generally accept the fact that local and global context and cotext are interrelated:

"Le cotexte global et le cotexte local ne sont pas indépendants. L'approche dominante dans la lignée logico-grammaticale privilégie la compositionnalité du sens, autrement dit, une détermination du global par le local. À l'inverse, la tradition rhétorique herméneutique a renversé la perspective et elle a mis en avant le principe de détermination du local par le global.[59]"(ibid. 2012: 15)

This applies at the level of words taken in the context of text. At the level of words, Barsalou (1982) suggests that meanings have context-dependent and context-independent relations, as he analyses the nature of information in concepts:

"Context independent properties form the core meaning of words. […] Context dependent properties are a source of semantic encoding variability" (Barsalou 1982:1)

Meaning change is influenced by both, and the study of context patterns is a way to understand its mechanics. Changes (or variability) in context-dependent properties can lead to changes in context-independent properties:

"Context-independent properties are activated by the word for a concept on all occasions. The activation of these properties is unaffected by contextual relevance. Context-dependent properties are not activated by the respective word independent of context. Rather,

---

[58] "The local environment corresponds to the direct vicinity of the target lexical unit. The precise definition of what constitutes "near neighbors" varies depending on the chosen perspective. Close vicinity can be defined at the lexicographic, syntactic, utterance or typographical planes, of which each unit is likely to correspond to a semantic unit."

[59] "Global cotext the global and local co-text are not independent. The dominant approach in the logico-grammatical tradition emphasizes meaning compositionality, i.e., a determination of the local by the global. Conversely, the hermeneutic rhetorical tradition has reversed the perspective and has put forward the principle of the local being determined by the global."

these properties form the core meaning of words, whereas context-dependent properties are a source of semantic encoding variability." (Barsalou 1982:1)

I have chosen to use the expression "linguistic context" since it seems to be most widely used. By "linguistic context" I mean all aspects and levels of it, at the exception of the extra-linguistic context, which is studied too, at a sociolinguistic level. Statistical analysis techniques offer the possibility to extract information from the linguistic context. This analysis can be conducted at different levels, depending on the statistical method that is chosen. This flexibility is necessary in semantics, since case studies of target words show different degrees of complexities. Therfore statistical techniques have to be adapted to the precise phenomena at stake. As stated in the introduction, there is no single measure or mathematical formula that can deal with the different types of semantic changes altogether (yet). The next section deals with the tools and applications of statistical analysis in corpora.

1.2.2 Applied Statistical analysis in corpus

Statistical tools to assess frequency variation and context variation have been easily available for the past thirty years. Not only do researchers apply statistical measures to word frequencies, but they also do so with concepts (see for instance Nazar's 2011 doctoral thesis for a quantitative approach to concept analysis). The methods described in Chapter I.1 have therefore been applied with the help of statistical tools. For instance, morphological studies highly benefited from statistics and computers. In this perspective, Baayen (1991) offers a quantitative analysis of morphological productivity, based on the word frequency distributions of morphological classes in corpora in English and Dutch, suggesting that productivity is correlated with token frequency. But beyond morphology and productivity issues, researchers have offered tools dealing with creativity, context, features, argument structure, dialectal variation and specialized uses among others, and can deal with the press, the Web, dictionaries, thesauri, or more complex databases. Statistical analysis is used in texts, in literary and discourse analysis, to determine the vocabulary and style of an author, evaluate the evolution of their style over time, or compare their style with the ones of other authors. It is a tool in linguistics, discourse analysis, diachronic studies and data mining at large.

1.2.2.1. Semi-automatic comparison of databases with press or Web based corpora

1.2.2.1.1. *Early contributions*

One of the most widespread and early techniques in terminology to detect neologisms in corpora is to compare candidates extracted in a target corpus to a reference corpus, which can be another text corpus or a dictionary or thesaurus. The reference corpus is generally representative of a "norm", if specialized press is compared to mainstream press, or representative of a period of time. This method has been widely used in terminology, using various theoretical perspectives and languages.

Cabré and de Yzaguirre (1995) describe an early semi-automatic method for the detection of neologisms, based on the comparison of Catalan and Castilian Spanish databases. The method relies on filtering candidates that have no match in the reference dictionary corpora. The authors sort candidates step by step, with the help of disambiguation, part of speech tagging, stop words (function words) morphological analysis and error filters. This system has been developed since by the authors and colleagues working in the same framework, for instance in (Cabré et al. 2003) which describes *Sextan*, a program retrieving all the instances that are not recognised by an electronic dictionary in Spanish and Catalan press corpora. Such methods became widespread two decades ago. An exhaustive overview of them in the French Language is given in the 20th edition of the journal *Terminologies Nouvelles* (Garsou 1999) gathering works of the "Réseau International de Néologie et de Terminologie" ("International Neology and Terminology Network") now called the "Réseau International Francophone d'Aménagement Linguistique[60]" ("International French Speaking Network for Linguistic Development"). The journal articles deal with Internet search for neology, semi-automatic search for neologisms in corpora, and general questions of neology and terminology. Within this journal, L'Homme, Bodson, and Valente (1999) describe a semi-automatic method in which automatic filtering alternates with human selection and analysis. They use the *Banque de terminologie du Québec*[61] (BTQ) as a reference, to detect neologisms from medical

[60] http://www.rifal.org/ (last accessed 06/06 /2012). The Rint (Réseau international de néologie et de terminologie) and Riofil (Réseau international des observatoires francophones de l'inforoute et du traitement informatique des langues, "French Speaking Observatories International Network for inforoute and NLP") merged to create the "Réseau International Francophone d'Aménagement Linguistique" in 2000.

[61] Terminology Bank of Quebec

corpora. The authors underline the difference between the semi-automatic method they use and the manual method. Roche and Bowker (1999), in the same volume, describe *Cenit* (a *Corpus-based English Neologism Identifer Tool*). They also describe other tools for neology available at the time:

"Il existe déjà quelques systèmes créés dans le but de détecter des néologismes. Le système Aviator (Blackwell 1993 ; Collier 1993; Renouf 1993a) identifie les néologismes dans la presse anglaise pour fournir aux enseignants des ressources lexicales. De même, le projet Obneb (Cabré and de Yzaguirre 1995) a pour but la détection des néologismes dans la presse, mais cette fois en espagnol et en catalan. L'outil NeoloSearch (Janicijevic and Walker 1997) identifie les néologismes français qui se trouvent sur l'Internet, tandis que le système Cordon (W3Cordon[62]) utilise des techniques statistiques et donc n'est pas lié à une seule langue." [63] (Roche and Bowker 1999: 12)

Cenit is based on a filtering method to extract neologisms in specialized terminology. However, it is limited to the detection of formal neology:

" *Cenit* n'identifie que les néologismes de forme. Il ne peut pas détecter les néologismes de sens, créés, par exemple, par extension sémantique ou par changement de catégorie grammaticale. [64]" (*ibid* : 13)

This limitation has partly survived time and progress. We are becoming more and more effective at detecting sense neologisms. However, they still strongly resist a purely automatic treatment even if several attempts are getting close to this aim.

---

[62] obsolete

[63] "Systems to detect neologisms already exist. The system Aviator (Blackwell 1993; Collier 1993; Renouf 1993a) identifies neologisms in the English press to provide teachers with lexical resources. Similarly, the project Obneb (Cabré and Yzaguirre 1995) is aimed at the detection of neologisms in the press, but this time in Spanish and Catalan. The tool NeoloSearch (Janicijevic and Walker 1997) identifies the French neologisms that are on the Internet, while the Cordon system (W3Cordon) uses statistical techniques and thus is not bound to a single language."

[64] "*Cenit* identifies formal neologisms only. It cannot detect sense neologisms that are created, for instance, by semantic extension or change in grammatical category."

1.2.2.1.2. *Neology networks and observatories*

The aforementioned networks of neology observatories are far from being unique in the academic world. A series of academic neology observatories are active in the world, with the aim to detect, model, list, and understand all forms of neology. These observatories are language-specific. In France, the Bornéo[65] database has been created by the ATILF[66]. In Spain, the The Observatori de Neologia at the Universitat Pompeu Fabra is leading similar research, in association with observatories of Latin America and Spain, working on Catalan and Spanish of Spain and Latin America as well as with observatories of various romance languages[67] via the Neorom project:

" Il consiste en la création d'un Réseau d'Observatoires de la Néologie des langues Romanes (NEOROM) qui, au moyen d'une même méthodologie de travail, produira des résultats dont la comparaison aura quelque validité pour analyser le renouvellement lexical des langues romanes à travers l'étude des néologismes spontanés ou planifiés qui apparaissent dans la presse, et dans d'autres médias oraux, écrits et audiovisuels, ainsi que dans d'autres situations de communication." [68] (Cabré, 2006 : 116)

The idea behind these types of networks is to federate a system of neology watch, echoing the watch provided by private companies or authorized governmental bodies. They try to set the lines for a common theoretical and methodological ground to obtain results that may be shared across language communities.

---

[65] http://www.atilf.fr/borneo/

[66] Analyse et Traitement Automatique de la Langue Française (French langage NLP and analysis). http://atilf.atilf.fr/

[67] For Spanish, Catalan, Galician, Italian, French, Canadian and Belgian French, Portuguese, Brazilian Portuguese, and Romanian

[68] "It consists in the creation of a Network of Neology Observatories for Romance Languages (NEOROM), which, with a similar methodology, will produce results which will be valid in order to analyse the lexical renewal of Romance languages via the study of spontaneous or planned neologisms that appear in the press, and in other spoken, written and audiovisual mediums, as well as in other communication circumstances."

1.2.2.2. Including features and/or argument structure

As we create tools to research language, the view we have of language structure, if encoded in the tool, is reflected in the results and analysis. Choosing a specific theoretical framework, however, is very hard to avoid, and has benefits, in that it provides a coherence to the analysis.

For instance, Reteunauer (2012), in her doctoral thesis, implemented statistical tools with respects to the interpretative semantics theoretical framework of Rastier (1987), developed in Rastier and Valette (2009) and several other works. Within this theoretical framework, meaning is codified according to a series of features, within a lexicographic reference corpus, and phenomena from text corpus are compared to the database. Here the reference (or "exclusion") corpus is a lexical one, in the form of an enriched dictionary that encodes semantic features to which innovations in textual corpora are compared. If the lexical corpus does not contain the innovation detected in the textual corpus, the item is selected as a neology candidate. The POMPAMO[69] system, developed at the CNRTL[70] also works this way, relying on a morpho-syntactically annotated corpus.

Mejri (2006) also offers a strategy for the automatic detection of neologisms based on an enriched dictionary type database that includes argument structure. He argues that the stratified methods which are widely applied do not fit well the detection of sense neologisms:

"Elles sont fondées sur un découpage faisant de chaque niveau de l'analyse linguistique une discipline autonome : phonologie, morphologie, syntaxe, sémantique, lexique, etc. Appliquée à la néologie, cette démarche rencontre beaucoup de difficultés dans la reconnaissance automatique.[71]"(Mejri 2006 : 2)

Moreover, the author rightfully argues that it is necessary to go beyond word level and include sentence context as well as domain, register and the context generated by

[69] http://www.cnrtl.fr/outils/pompamo/requetes.php

[70] Centre National de Ressources Textuelles et Lexicales (Lexical and textual resources national center)

[71] "They are based on a segmentation which makes each level of linguistic analysis an autonomous subject: phonology, morphology, syntax, semantics, lexicon, etc. When applied to the study of neology, this approach encounters many difficulties to perform automatic recognition."

circumstances. To overcome stratification and the multiplicity of contexts and meanings, Mejri suggests gathering all this information in a single enhanced dictionary. This dictionary lists all the possible meanings of a word in relation to its context of use and relies on features and argument structure. For instance, the feature "human" is enriched with a list of all the descriptions it can stand for. Verbs are encoded with argument structure. As an example, the verb *adhérer* would be depicted as follows:

> "adhérer[72] :
>
> adhérer/N0 : inc/N1 : à inc/N2 :/Ad :/Nm : adhérence/Sy : coller/D : tec/
>
> adhérer/N0 : hum/N1 : à ina < opinion >/N2 : de hum/Ad :/Nm : adhésion/Sy : être d'accord avec/D : psy/"(Mejri 2006: 6-7)
>
> adhérer/N0 : hum/N1 : à hum < collectif >/N2 :/Ad : adhérent/Nm : adhésion/Sy : être adhérent/D : pol/ " (Mejri 2006: 6-7)

In this example, the first line corresponds to the meaning "glue, stick", the second to "agree with" and the third to "become a member of", three lexical units of the French verb *adhérer*. The method uses domain by including "tec" for technology in the first definition, "psy" for psychology in the second, and "pol" for "politics" in the third.

The method requires building complex databases and basically re-encoding all the occurrences in the lists of descriptions. Ideally, a generic system could avoid that.

Beyond the search for a valid method, corpus study also teaches us about the culture and society text is produced in.

1.2.2.3. Text statistics for socio-political analysis and the media and statistical detection of neologisms in press corpora

The French disciplines of "Sémiométrie" (Lebart, Piron, and Steiner 2003), "Textométrie", "Logométrie" and "Lexicométrie"[73] which are branches of lexicology, text statistics,

[72] Join, agree with or adhere

[73] Translatable as semiometrics, textometrics, logmetrics and lexicometrics

discourse analysis and sociology, all provide analyses of current political and social issues. Socio-political analysis can be conducted with statistical co-occurrence and frequency variation measures with the help of various computer programs. For instance, the speeches of several French presidents have been compared by Leblanc and Fiala (2004) relying on statistical co-occurrence software. These works include semantic change issues within a wider perspective of textual and discourse analysis and are not centered on semantic change *per se*. However they do partially deal with it and related issues, in particular the deployment of connotation in text.

As mentioned earlier, there are numerous works offering detection and analysis methods for neologisms in press corpora. Among them, the outstanding work of Renouf (2007) exploits a UK press corpus from 1989 to the close of 2005, with additional data from the Web.

"Hypotheses about the relationship between meaning and surface textual patterning have proved to be sound, allowing us to cumulatively develop algorithms for automated systems capable of identifying a number of lexical and lexico-semantic phenomena in text across time. These phenomena include neologisms (Renouf 1993; (Baayen and Renouf 1996), new word senses (Renouf 1993 b, c, d), sense relations and changing sense relations within text (Renouf 1996; Collier, Pacey, and Renouf 1998) and the kinds of productivity and creativity of new words entering newstext […] The methodology for each of these procedures varies, but basically involves "feeding" a specific time chunk of chronologically sequenced, fresh textual data through a set of software filters which detect novel words as well as new collocational environments of existing words" (Renouf 2007: 62-63)

The author focuses on (morphological) productivity, creativity, word life cycles and collocational profiles. She notes that creativity in the press is limited compared to literary creativity, but is generally used to create stylistic effects. Moreover, "vogue words", as she calls them, are tightly linked with real-world events. In terms of life-cycle, these vogue words tend to show high frequency peaks before their frequency diminishes drastically and eventually stabilizes. Renouf also analyzes words separately within a collocation. For instance, to fully understand the expression *weapons of mass destruction*, she looks for other collocations based on *weapons of* without *mass* or without *destruction*. Examples are *weapons of mass distraction*, or *weapons of class destruction*. She points out that these substitutions seem to follow rules. *Distraction* and *class* sound very similar to *destructio*n and *mass*, for instance.

In case studies, Renouf shows how a vogue word has generally been around for a while. For instance the word *chav* is found in UK corpora in the nineties but shows a peak in use in 2004. This peak is accompanied by semantic shift, as *chav* now refers to a person of low class and low education with a poor style, whereas in the nineties it only meant a young person. The case studies in Part III. of this work are congruent with this observation.

"Probably the central revelation is that words have a life-cycle consisting, in the most general terms, of birth or re-birth, followed by gentle or steeper upward trajectories in frequency of use and leading to brief or lengthier moments at the zenith of popularity, after which they take faster or slower downward paths, until they reach a stable level of use. During this life-cycle, words which make a sufficient impression on the public imagination will also spawn a number of productive and creative variants. […] A final observation is that our diachronic, empirical study of the data reveals and traces parallels between language and contemporary world events, and thus acts as a window on contemporary culture; in particular on recent world events and on aspects of youth culture as reflected in the media." (Renouf 2007: 87-88)

1.2.2.4. The media and statistical semantics

There is nothing new in stating that academic works on political language and the language of the media provide a way to decipher the political and cultural content behind discourse and to shed light on all the implicit contents that are conveyed in this particular type of language. What is relatively new, however, is the growing interest of non-academics for these works, and the way they are used in the media and society.

These specialized works become more and more available and demanded outside of the academic realm. The general public now includes analytical tools in their everyday understanding of the media, via the press and the Internet. The mainstream press integrates data from semantic analysis. For instance, *Le Monde* gives the floor to Jean Veronis (and his blog[74]), their official linguist partner to decode political news through a semantic lens. Veronis deciphers the news with frequency measures of key words. For instance, he compared the vocabulary of Martine Aubry and François Hollande before the French primary election

[74] http://blog.veronis.fr/

opposing them at the head of the left wing party[75]. Veronis publishes frequency tables of the major keywords used by the candidates to clarify their political profile through their use of connoted vocabulary. Veronis is also associated to Linkfluence[76] a social Web watching company, offering trend watching, brand awareness watching, social Web notoriety watching and the like. These types of services are offered by a growing number of companies, using techniques from computational semantics to watch language and opinions online and offer a picture of them to several clients. These services existed long before the Internet boom; however, they are now based on the Internet as a resource and are progressively exploiting methods developed by academics. There is a shift as regards the type of population in touch with semantics and discourse analysis. A few years ago, companies were in touch with them through other companies offering surveys, opinions polls, and trend watching. Now, these tools are known by the general public, outside of the executive realm, and it is almost the norm for TV programs and the press to include semantic analysis to the array of tools they convoke. However, the meaning of *semantics* is undergoing a process of semantic connotational drift, as popular etymology now tends to associate it with the idea of *semantic Web*, therefore greatly narrowing its original meaning. Moreover, the semantics of *semantics* seem to have blurred in this process.

Political analysis is one of the major areas that push the integration of statistical semantics into mainstream media. Key words in politics may also have a different connotation depending on who is employing them and these connotations may give birth to semantic drifts. For instance, when Arnaud Montebourg (a left-wing French politician) mentions the idea of *démondialisation* ("deglobalisation"), a recent neologism serving the ideas of anti-capitalist movements, he does so with the seriousness of a political program, while right-wing politicians colour the word with great irony, thereby discrediting their adversary's program.

The rising integration of semantic tools and methods into mainstream media is clearly anchored in the assumption that language and information is the same thing. This idea is also developed by Picton (2009) who centers her analysis on the relationship between language

[75] http://politicosphere.blog.lemonde.fr/2011/10/13/les-mots-de-aubry-et-hollande-des-contenus-presque-identiques-des-styles-differents/#xtor=RSS-3208

[76] http://fr.linkfluence.net/

and knowledge, putting them in parallel rather than assimilating them completely. She calls this:

> "… la pertinence d'un parallèle entre langue et connaissances[77]" (Picton 2009: 253)

We are witnessing the integration of statistical semantics methods and discourse analysis into mainstream media channels and the Internet, which creates a new interest, with a wider scope than ever, for methods emerging from linguistics. Meanwhile, the variety of possible approaches is rocketing, with various sub-disciplines offering insights to analyse semantic change, in terms of computational tools or theoretical approach, or both.

1.2.2.5. Insights from specialized terminology and lexical variation

1.2.2.5.1. *Specialized terminology, language varieties, and sub-communities: petri dishes for neologisms*

New meanings and new words are often created in specialized domains, in which the apparition of new realities and objects to name is quicker, and where new concepts emerge more rapidly. Leading edge areas such as spatial research[78], advanced physics, neurosciences, or artificial intelligence produce their own share of new vocabulary through borrowings, neologisms and semantic changes. This makes specialized terminology a petri dish for neologisms and innovations. Beyond specialized terminology, some language sub-communities at the sociological level demonstrate a high level of creativity, to define their identity and stand apart from mainstream concepts and their associated language. Here, language has a highly conceptual, stylistic and sociological value. Sub-communities may be identified through their musical taste (such as the hip-hop, electro, punk, and rock communities, among others) or their interest (for instance "geeks" interested in new technologies) hobbies (football, role plays, etc.) in relationship with certain clothing habits, and language styles. These sub-communities are a treasure of semantic shifts and neologies, however, most of them remain contained within those communities. Moreover a word's niche, that is to say its relationships to speakers communities and its association with specific topics, plays an important role in its possibility of evolution, as demonstrated by Altmann,

[77] "… the relevance of a parallel between language and knowledge".

[78] Studied by Picton (2009)

Pierrehumbert and Motter (2010) who claim that niche is more determinant than frequency in short language dynamics, on the basis of work conducted with the online USENET corpus. Their finding is that in modern online corpora niche is the most efficient criteria to assess the chances of survival for lexical innovations. The notion of niche is borrowed from biology:

"A new word, like a new species, must establish itself in a niche to survive in the language." (Altmann, Pierrehumbert and Motter 2010)

The corpora these authors rely on are composed of the contents of two USENET discussion groups, one about music, and one about Linux operating systems. Case studies focus on slang such as *lol* ("laughing out loud"), product words such as the software *gnome* in Linux jargon, or personality names such as *Eminem*, a rapper. The authors measure the dissemination of each word across users (D^u), or its "indexicality", and across threads (D^t), or its "topicality". They show that D^u and D^t are predictors of word fate, in terms of how they disseminate, rather than in terms of pure frequency. Indeed, some words may have a high frequency but be used only by a few users, whereas other words may show lower frequencies but disseminate across users effectively, thereby having more chances to reach a wider community of speakers. What is interesting for lexicologists is when specific words and concepts spread out of these restricted communities and reach a wider array of speakers.

1.2.2.5.2. *Statistical analysis of dialectal variations and lexical innovation in linguistic sub-communities*

Studies using word co-occurrence patterns show how new word meanings transfer from sub-communities to a more "mainstream" language. For instance Chesley (2011) questions how people learn vocabulary from listening to hip-hop music, by asking:

"AAE [African American English] words used in Rapper's Delight, such as fly (''cool/attractive'') and bad (''cool''), subsequently enjoyed some prominence in Mainstream American English (MAE; also called ''Standard'' American English) throughout the 1980s. Could it be that non-African-American speakers learned these words through listening to hip-hop songs such as Rapper's Delight?"

Here the cultural aspect of the hip-hop sub-community merges with dialectal variation; African American English being dealt with as a variation of Mainstream American English.

The use of the word *bad* meaning *cool* is typical of the reversals found in semantic change processes, exactly as *wicked* came to mean *cool*. The main context of use of *wicked* shifted from the Bible, in which it meant "evil" with a connotation of power, to the general language and took on the meaning "bad", until it entered American and British underground music and in particular Hip-hop, Jungle and Drum'n'Bass music in which it means "cool", probably a borrowing of the American slang listed in 3b:

"I. adj.

1. Bad in moral character, disposition, or conduct; inclined or addicted to wilful wrong-doing; practising or disposed to practise evil; morally depraved. (A term of wide application, but always of strong reprobation, implying a high degree of evil quality.)
a. of a person (or a community of persons). The Wicked One , the Devil, Satan.
[…]
3. b. Excellent, splendid; remarkable. *slang* (orig. *U.S.*).
[…]

II. *absol.* or as n.

4. In sense 1a : chiefly in biblical and religious use; often opp. to righteous n. 2a."[79]

Lexical variation across language varieties is also studied on its own, and provides synchronic tools that may be adapted in diachrony, since the mechanisms at stake show great similarity, aside from the temporal aspect. For instance, Peirsman, Geeraerts and Speelman (2010) look at lexical variation between Netherlandic Dutch and Belgian Dutch, retrieving cross-lectal (or *cross-dialectal*) synonyms from corpora, using a distributional model and showing the suitability of co-occurrence based models with the issue.

1.2.2.5.3. ***Specialized terminology***

Specialized terminology is a great source of innovations as areas such as space exploration research need to adapt to new concepts and names new objects rapidly as shown in Picton (2009) who offers a full-fledged analysis of short diachrony phenomena in specialized corpora. Her standing point is to extract the evolution of knowledge through the detection of language innovations. In a similar vein, Dury and Drouin (2009) use a corpus from ecology to retrieve a series of useful indices to detect neology and what they call *necrology* (or

[79] Source: OED

obsolescence). Dury and Drouin (2009) rely on ecology, as it is a domain that has been confronted with radical changes recently and became one of the hottest themes in society and politics. These changes may be observed in the temporal window they chose: from 1950 to 2005. They show how a word, judged obsolete, is replaced by a synonym, or how one synonym takes over a series of competing words, such as *ecosystem* taking over *microcosm, holocoen, biosystem* or *bioinert body*, or how a word like *superorganism* disappears from specialized language. Moreover the authors offer a series of indices to detect semantic change in corpora, strikingly similar to mine, which are detailed in Part III.

Drouin, Paquin and Ménard (2006) offer a semi-automatic detection method for neologisms in French. Their method relies on the comparison of corpora from different periods and with different degrees of specialization. The authors use the software *TermoStat* (see Drouin 2003) to extract neology candidates in two types of corpora. They compare a specialized corpus dealing with terrorism, covering a ten year time span from 1995 to 2005, subdivided into two sub corpora before and after September 11th 2001, to a non-specialized corpus, composed of all the articles published in the newspapers *Le Monde* in 2002. The specialized corpus mainly contains military and political journals, internet sites and academic material. The authors study frequency variations, and rely on a statistical measure of specificity[80], in line with previous work by Lafon (1980) and Lebart and Salem (1994) to extract a list of candidates. The candidates are terms related to terrorism as well as terms that have a relationship with the latter, including words in actantial, morphological and paradigmatic relationships. These candidates are then ranked. The co-occurrent words of these selected candidates is subsequently analyzed with the software *SATO* (Duchastel, Daoust, and Della Faille 2004) to study their context. Results include prefix and suffix production (such as *hyperterrorisme* and *djihadisme*) and notably the creation of an abbreviation: *ADM* (standing for *armes de destruction massive*, "weapons of mass destruction"). The authors therefore give a socio-political picture of the language of terrorism after the world shacking event of September 11th 2001. In the corpus study described in Part III of this work, terrorism and terrorism related vocabulary show extremely high frequencies and frequency variation, with the spectrums of *Iraq, Afghanistan* and *Georges Bush* reaching the highest frequencies.

[80] In French : « mesure de la spécificité »

1.2.3 Pioneer computational works in modeling semantic change

At the beginning of this chapter I mentioned the need for representation. This need does not only apply at an abstract level but also at a practical, graphical level. Graphs are more and more seen to represent language, starting with simple mind-maps, available to the lay person, all the way to complex models.

"L'utilisation de graphes en sémantique lexicale s'appuie sur un regain d'intérêt plus général pour la théorie des graphes en sciences sociales[81]" (Loiseau et al. 2011)

Before looking at graphs and models for semantic change, it is worth taking a detour via pioneer works in modeling semantic change.

Before the burst of models in semantics, pioneer works by Nerlich and Clarke (Nerlich and Clarke 1988; Nerlich and Clarke 1999b; Clarke and Nerlich 1991) set out an attempt at modeling semantic change with the help of a computer.

Nerlich and Clarke (1988) and Clarke and Nerlich (1991) suggest retaining three major factors in order to model meaning: frequency, memory accessibility, and expressivity. They created a computerized model which simulates word change dynamics and obtained wave-like patterns similar to the pattern of change described by Berrendonner, Le Guern and Puech (1983). They assert that:

"Language changes according to a regular pattern" (Clarke & Nerlich, 1991)

Berrendonner, Le Guern and Puech (1983) describe a dynamic in which the frequency dominance of a lect is overtaken in time by the frequency dominance of a second lect "in competition", and then by a third. In this pattern, only frequency is taken into account. Nerlich and Clarke (1988) and Clarke and Nerlich (1991)'s model, however, relies on the idea of synonymic competition (such as *baby*, *toddler*, or *infant*), implementing an onomasiological view, and establishes a feedback loop between frequency and accessibility, subsequently

[81] The use of graphs in lexical semantics relies on a more general renewed interest for graph theory in social sciences"

enriched with expressivity as a random variable. They state that expressivity is the major factor in the replacement of a word by another in terms of frequency dominance. Expressivity is defined as a type of *motivation* of the sign. Expressivity wears off as a word loses its novel feel:

"A frequently used word wears out its expressivity and novelty and is - in the long run- absorbed into the stock of the words of normal usage" (Nerlich et Clarke 1988: 78)

Although the results seem to show a clear wave pattern, the authors question whether this might be due to the model itself, and acknowledge that:

"Although the emergence of the word-waves is interesting in itself, their origin still seems to be a mystery. They would arise in any system where the relative values of two variables describing a set of objects were linked by positive feedback in this way together with similar effects of random variation. The properties of the model arise from its mathematics, not the linguistic facts we used it to stand for." (*ibid*: 235)

By doing so, they set the path for further research in semantic modeling. The onomasiological approach looks at how meanings transfer between words rather than how words change meaning (the semasiological approach). It is inspiring and will be taken into account in this research. In effect, it seems that the way the question is asked first and foremost does impact the research methods:

"To take some specific examples, we did not ask how a word such as *nice* started out by meaning 'foolish' and came to mean 'pleasant'; or *aftermath* went from meaning 'second crop of grass' to meaning 'consequence' or 'resulting state'; but by contrast, how with the passage of time the word of choice for a senior member of a company has come more and more commonly to be *executive* rather than *manager;* or why the word of choice for a small ensemble of pop musicians has switched from *group* to *band."*(*ibid:*228)

It is essential to deal with synonymic competition while looking at meaning transfer, as word meanings evolve in a network and not on their own. A word and its evolution may not be analyzed as a single unit with a life of its own, but has to be included in the semantic networks it is part of. This point of view is defended by Buchi (2000) :

"D'une manière générale, il nous semble que la lexicologie historique ne remplit qu'imparfaitement son rôle si elle ne prend pas en considération la dimension onomasiologique, en particulier le phénomène de la synonymie." [82]

Theoretically, Nerlich and Clarke (1999b) give a central role to blending theory in the spirit of Oakley (1998) and building on Lakoff and Johnson (1980), viewing meaning as follows:

"Meanings are not mental objects bounded in conceptual places but rather complex operations of projection, binding, linking, blending, and integration over multiple spaces." (Nerlich and Clarke 1999b) citing Turner (1996 : 57)

By applying this view to meaning and meaning change, they imply that metaphor and metonymy are central processes in the nature of meaning and meaning change, and may not be treated as a peripheral phenomenon. Meanings map onto each other, and may transfer part of their content from a lexical unit to another, thereby modifying other connected lexical units. Connected lexical units are defined here as all words and word's sub-meanings pertaining to the same semantic domain, in a paradigmatic axis, or being in a syntagmatic relationship to them. These works are not only pioneer in that they bring modeling, but they also bridge theoretical advances so as to adapt to computer models. They also offer bridging results with those obtained in synchronic investigation of rhetorical figures and with results in language acquisition in children. Young children (up to age 2,5) tend to over-exploit metonymical and metaphorical extensions while older children use metonymy and metaphor in creative ways as part of semantic communicative strategies.

"There seems to be a developmental sequence going from compelled metonymically or, metaphorically and even synecdochically based overextensions to more creative pretend-naming, to the use of similes, to the production and then understanding of metaphors. These can be regarded as overlapping stages in a child's semantic development."(Nerlich and Clarke 1999b: 8)

[82] "In general, it seems that historical lexicology fulfills its role imperfectly if it does not take into consideration the onomasiological dimension and in particular, the phenomenon of synonymy."

By adopting an onomasiological view, integrative of meaning transfers and of the teachings of historical linguistics and philology regarding the role of rhetorical figures, to adapt to a computational paradigm, the works of Nerlich and Clarke open the path to adaptations that have become necessary in the current scientific and sociological context. Their attempt at modeling has now been enriched by a plethora of more advanced models that may be relied on to understand semantic change.

Chapter I.3: Semantic Change with context models

1.3.1 Studying semantics with context models

Among the available approaches to tackle issues in semantics, the growth of multidisciplinary methods combining semantics and modeling, cognitive sciences and psychology also gained momentum with the growing introduction of statistics and technology as well as the influence of cognitive sciences. One of the primary goals of cognitive sciences was to bridge several disciplines concerned with knowledge and the brain to come to a unified framework and a more holistic paradigm. At the heart of these approaches, and through the interaction between disciplines, a series of questions arose. For instance the suitability of models in terms of their fitting human psychology was questioned, as well as the issue of dealing with and representing properly large amounts of textual data. These issues also led to the integration of other emerging fields: information technology and visual analytics, with the search for adequate interfaces and the study of their interaction with data. Visual analytics can be defined as "the science of analytical reasoning facilitated by interactive visual interfaces" (Thomas and Cook 2006: 10). These highly multidisciplinary approaches are still in their infancy. They now benefit from a great deal of contributions from the scientific community at large.

In semantics, a diversity of vector models is born out of that growth. A few studies of semantic change *per se* use these models. I first look at what these models are and how they apply to semantics, and then at semantic change studies that are conducted with them. A substantial introduction to the models themselves is necessary first. Moreover, the model used in this study is also a vector model.

The choice of the model depends on the investigation that is conducted, the type of data being represented, and the types of variables coming into play. Indeed, models may take several variables and thus allow for combining different types of information, such as syntactic and semantic information. The development of models that may deal with these two aspects and ultimately represent sentence meaning is one of the major challenges in the field (see Mitchell and Lapata 2008).

1.3.1.1. Vector Space Models

The idea of using geometrical spaces and vector space models to represent and retrieve relationships between items or elements of language dates back to the 1960s[83]. After the upsurge in use of these models in the 1960s, enthusiasm progressively faded. In the past 20 years, however, there was a revival in their use in linguistics and semantics, notably due to their evaluation as suitable models to reflect the cognitive organization of knowledge in cognitive sciences and psycholinguistics.

Some research claim that vector models are a suitable reflection of the organization of the human mental lexicon, see for instance Landauer et al. (2007) and Ji et al. (2008).

This revival also impacted corpora studies since these models possess the advantage of retrieving data automatically from corpus. They turned out to be more efficient in dealing with raw and annotated text as there is no need to create and gather data to create lexicons (although it can be done too).

Vector space models (coined by Schütze 1993) are mathematical models that represent words, documents, or other items, by context vectors in a multidimensional space. Several types of vector space models allow for filtering, indexing and retrieving large sets of information. An exhaustive state of the art and comparison of the main word vector space models may be found in Sahlgren (2006). The use of these models in semantics relies on the *distributional hypothesis*, anchored in the idea that the linguistic context of a word, retrieved via the study of co-occurrence patterns, can tell us about its meaning. Moreover, measures of *semantic similarity* between context vectors may be used to detect words with similar meanings.

1.3.1.2. The distributional hypothesis

As stated earlier, the general idea behind distributional semantic is that "you shall know a word by the company it keeps" (Firth 1957: 11). This idea emerged from reference works by Wittgenstein (1953) and Harris (1954) before it was made popular by Firth (1957). Its centrality in distributional semantics is acknowledged by Turney and Pantel (2010: 142,143):

[83] The SMART information retrieval system (System for the Mechanical Analysis and Retrieval of Text) developed in the 1960s at Cornell University, New York, is considered to be the first instance of a vector space model used to retrieve textual data (see Salton, Wong, and Yang 1975).

"The distributional hypothesis is that words that occur in similar contexts tend to have similar meanings (Wittgenstein 1953; Harris 1954; Weaver 1955; Firth 1957; Deerwester et al. 1990). Efforts to apply this abstract hypothesis to concrete algorithms for measuring the similarity of meaning often lead to vectors, matrices, and higher-order tensors."

This idea is the foundation for word space models using context vectors:

"The general idea behind word space models is to use distributional statistics to generate high-dimensional vector spaces, in which words are represented by context vectors whose relative directions are assumed to indicate semantic similarity. This assumption is motivated by the distributional hypothesis, which states that words with similar meanings tend to occur in similar contexts. According to this hypothesis, if we observe two words that constantly occur with the same contexts, we are justified in assuming that they mean similar things. " (Sahlgren 2005)

According to the latter " the distributional hypothesis is usually motivated by referring to the distributional methodology developed by Zellig Harris" (Sahlgren 2006: 22)

Looking at the direct neighbours of a word is therefore a means to extract its meaning. However, a word's context does not technically equal its meaning. More specifically, it gives insights as to its use, including sociolinguistic and linguistic use.

To generate the space, a matrix is built from the corpus, either raw or tagged. Tagging includes corpus tokenization, part-of-speech tagging- and/or parsing. Turney and Pantel (2010) distinguish three classes of vector space models: term–document, word–context, and co-occurrence pair–pattern matrices models, and note that each type of model serves specific types of data. Indeed, vector space models may handle different types of data in different theoretical frameworks. For instance, in psychology and psychometrics, the matrices correspond to subject –item entries. The three types of models are based on chunking data in different ways. However, there are always two levels to compare.

"If we generalize the idea of documents to chunks of text of arbitrary size (phrases, sentences, paragraphs, chapters, books, collections), the result is the word–context matrix, which includes the term–document matrix as a special case." (Turney and Pantel 2010:146)

There are several ways to handle words, either by relating them to the document they belong to, or to the *chunk* of text they belong to (sequence of words, sentence or paragraph). In document-based models, a tf*idf weighting measure is generally used to deal with word frequency. "Tf*idf" stands for "term frequency-inverse document frequency". This measure shows how important is a word in a corpus. The figure rises with frequency but is offset by it; therefore balancing out figures for the most common words. Tf*idf measure is thus interesting as regards the relevance of a word in a document.

The general hypothesis is that "if units of text have similar vectors in a text frequency matrix, then they tend to have similar meanings" (Turney and Pantel 2010 :153). The "bag of word hypothesis"[84] applies to word-document matrices. When column vectors are similar, it indicates that meanings are similar. This hypothesis can be extended to patterns (see Lin and Pantel 2001), therefore patterns that co-occur with similar word pairs can be said to indicate similar semantic relations. Conversely word pairs that co-occur in similar patterns indicate similar semantic relations (see Turney et al. 2003 for the latent relation hypothesis).

Moreover, two measures are used to estimate the efficiency of the system: precision (or positive predictive value) that evaluates how relevant is a document to a query, and recall rate (also called sensitivity) that evaluates the probability of that document being relevant to the query.

A few basic notions are necessary to understand word space vector models: types and tokens, stemming, co-occurrence and context windows.

[84] "In mathematics, a bag (also called a multiset) is like a set, except that duplicates are allowed. For example, {a, a, b, c, c, c} is a bag containing a, b, and c. Order does not matter in bags and sets; the bags {a, a, b, c, c, c} and {c, a, c, b, a, c} are equivalent." (*ibid.:*147)

1.3.1.3. Types, tokens, stemming, co-occurrence orders and context windows

1.3.1.3.1. *Types and Tokens*

In a corpus, words may be treated in terms of *types* or *tokens*. A *token* is the single instance of a word, for example if the word *house* appears 60 times in a given corpus, it has 60 tokens. The type encompasses all instances of the same word. The *type* of *house* corresponds to its class, or all its occurrences grouped under one heading, and therefore the type of house is 1.

1.3.1.3.2. *Stemming*

How the word is defined as similar or not, depends on whether part-of-speech tagging includes *houses* and *house* under the same type, or, if raw unstemmed text is dealt with, then *house* and *houses* are treated as two different types. The choice of that treatment may show extremely relevant when looking at semantic changes that are anchored in the use of inflected forms. As will be explained in Part III, I have chosen to investigate both stemmed and unstemmed text.

1.3.1.3.3. *Co-occurrence*

Co-occurrence may also be dealt with at different levels. Co-occurrence within a sentence for instance, gives highly syntagmatic results. For instance, while studying the word *bird*, syntagmatic word co-occurrence gives output such as *tree*, in a fictitious sentence of the type "the bird is in the tree". However, in semantics we are also interested in paradigmatic co-occurrence, which would give an output of the type *robin*, for the query word *bird*. To include these paradigmatic results, one method is to implement second-order co-occurrence, in which words that are both co-occurrent with *tree* are retrieved. Therefore, if there are two sentences "the bird is in the tree" and "the robin is in the tree" then second-order co-occurrence will also extract *robin* as a second order co-occurrent word of *bird*.

1.3.1.3.4. *Context window*

As explained above, models may be based on word/document or word/chunk matrices. In word/chunk methods, the chosen *chunk* is also called a *context window*. Its size does matter, as underlined by Peirsman, Heylen, and Geeraerts (2008) and may be sensitive or not to word order. We may decide to look at n words before and/or after the target word, to choose the sentence as a chunk, or the paragraph, or any other window size.

1.3.1.3.5. *Semantic similarity in vector space models*

Not only can we measure document similarity, but also word similarity, by comparing the row vectors in the term-document matrix instead of the column vectors. The measurement of *semantic similarity*, sometimes coined *semantic relatedness* (although the latter encompasses a greater number of relations) is based on the distances between the items in the generated space. The underlying idea behind semantic similarity measures is that words that are closely related semantically will tend to group in space and there will be a shorter distance between them than with the other ones. In spaces, this distance may be calculated between pairs (two words) or groups (a cloud of coordinates). There are numerous measures of semantic similarity within vector space models, as the idea seemed to emerge in the 1990s and is still in the course of development and testing. These measures can rely on probability (weighting) or on geometrical observations.

1.3.1.4. Widespread vector space models in linguistics

Among the most widespread models in linguistics, Latent Semantic Analysis (LSA)[85] (see Landauer and Dumais 1997) and derived models such as Probabilistic LSA (PLSA) or closely related models such as Latent Dirichlet Allocation (LDA) (see Blei, Ng, and Jordan 2003) have attracted numerous researchers and benefit from extensive updates from these research communities.

LSA and LDA are distributional models, also known as topic models, built around the assumption that related words will tend to co-occur within a single context with higher frequency than unrelated words. Topic models were first designed to extract topics in collections of documents. They both rely on word and frequency matrixes, to generate high dimensional spaces of representation and measurement:

"The input to LSA is a matrix consisting of rows representing unitary event types by columns representing contexts in which instances of the event types appear. One example is a matrix of unique word types by many individual paragraphs in which the words are encountered, where a cell contains the number of times that a particular word type, say model,

[85] http://lsa.colorado.edu/

appears in a particular paragraph, say this one. After an initial transformation of the cell entries, this matrix is analyzed by a statistical technique called singular value decomposition (SVD) closely akin to factor analysis, which allows event types and individual contexts to be re-represented as points or vectors in a high dimensional abstract space (Golub, Luk, and Overton 1981). The final output is a representation from which one can calculate similarity measures between all pairs consisting of either event types or contexts (e.g., word-word, word-paragraph, or paragraph-paragraph similarities)." (Landauer and Dumais 1997)

The Hyperspace Analog to Language model (HAL; (Lund and Burgess 1996) relies on very similar methods, however it uses word based co-occurrence contrarily to LSA and LSA-like models that use document based co-occurrence. Term-term co-occurrence is also developed in LSA-like models like Infomap[86]. It is the type of model used in this work.

1.3.1.5. Choice of a model

Within models, there are several mathematical and statistical techniques to obtain a space and to measure distance within it. For instance, computing phases such as SVD or Dimensionality Reduction which gets rid of negative values, as discussed in Van de Cruys (2010), is sometimes avoided by implementing Random Indexing (RI) that cumulates vectors into index vectors (Sahlgren 2005; Rosell 2009; Rosell, Hassel, and Kann 2009). The detailed mathematics behind these computations is complex and pertains to modeling issues beyond the scope of this work.

It is evident that the choice of the model impacts the representation of data, and there is huge variation across possible procedures as stated by Patel, Bullinaria, and Levy (1998) as each model shows itself finer than another on specific tasks. However, when using several models for the same task, results can also show a large degree of similarity. This was the case when crossing results from the model used in this doctoral thesis, the Semantic Atlas, with results produced by Infomap (Schütze 1996), a model in the same vein as LSA, on data mixing semantic, historical and phonological aspects (see Boussidan, Sagi, and Ploux 2009).

In computational semantics, factor analysis models are used to represent and analyze

[86] http://infomap-nlp.sourceforge.net/ Stanford, CA.

semantic relations like synonymy, coordination, superordination, and collocation as shown by Utsumi (2010). Vector space models are also used in semantics to compare dialectal variation in synchrony by Peirsman, Geeraerts, and Speelman (2010) for instance. These two approaches are very close to the topic at stake, since the detection of synchronic variation may be adapted to diachronic variation and the analysis of semantic relations intervenes in deciphering the mechanisms of change that may rely on them.

1.3.1.6. Limitations of vector space models

Vector space models, like any model, have limitations and biases. For the ones that are document-based, the size of documents has on impact on the output, and therefore results may be corpora-biased. Moreover, the lack of agreement between researchers on standards results in a plethora of models and makes it difficult to replicate or compare results across models. Nevertheless there is a tendency to try and establish standards in NLP by launching shared tasks and comparing results. In the computational semantics community, examples of that can be found in calls for shared tasks, as launched by the International Conference on Computational Semantics as in 2008 or by the ESSLI (Lenci 2008)[87] where the calls invited the comparison of semantic representation of texts as produced by several systems and models.

The main limitation of many word space models in semantics is that they generally associate a single vector per word, and thus fail to take into account the polysemous nature of words at the first step of modeling. One of the main advantages of the model that I use in this work, the Semantic Atlas, compared to the aforementioned models is that it builds spaces starting from polysemy upwards. This way, all meanings are kept. This method involves heavy databases, but provides a flexible and precise enough picture of meaning to conduct full-scale semantic studies.

1.3.2. Semantic change studies with context models: a state of the art

Very few studies of semantic change relying on vector models have been produced. This may be because the topic itself seemed to pertain more to non-computational linguistics or because

[87] http://www.sigsem.org/wiki/STEP_2008_Conference_Report

the models themselves needed validation in the scientific community as to their reliability and adaptability in semantics. Therefore the precise topic of conducting semantic change studies with context models is new and relatively unexplored.

The few following contributions, however, open paths in the area, providing rich results in terms of measurement and representation. They ask the question of whether change is measurable, and whether change may be detected and analyzed automatically. Those questions are similar to those asked in this work. Although the growing use of frequency measures across diachronic corpora in historical corpus linguistics is assessed by Hilpert and Gries (2009), the fact that frequency patterns alone are not sufficient to address the topic is acknowledged by all the authors quoted in this chapter.

Cook and Stevenson (2010) and Sagi, Kaufmann and Clark (2009) assess acknowledged cases of semantic change drawn from the historical linguistics literature to find grounds for detection, while Holz and Teresniak (2010), Heyer, Holz, and Teresniak 2009 and Rohrdantz et al. (2011) focus on semantic change detection and visual representation in topic models and LDA models.

1.3.2.1. Measuring polarity

Cook and Stevenson (2010) rely on the notion of *polarity*. Polarity is the tendency of words to possess a negative or positive connotation. Work on polarity, such as Kloumann et al. (2012), focuses on the emotional aspect of language by studying the neutral, positive or negative biases of words. This bias is called *semantic orientation* by the authors. Cook and Stevenson (2010) therefore deal with the detection of changes in semantic orientation, corresponding to the categories of amelioration and pejoration established by historical linguistics. To detect these phenomena, they rely on three English corpora covering a four century time span. To establish the polarity of words, they rely on a variant of the Pointwise Mutual Information method as described by Turney and Littman (2003), in which they compare target words to words of which the polarity is known, previously gathered in a manually compiled set of seeds. An excerpt of the result is shown in table 8 below:

Expression	Change identified from resources	Change in polarity	Polarity in corpora	
			Lampeter	CLMETEV
ambition	amelioration	0.52	-0.76	-0.24
eager	amelioration	0.97	-1.09	-0.12
fond	amelioration	0.07	0.14	0.21
luxury	amelioration	1.49	-0.93	0.55
nice	amelioration	2.84	-2.48	0.36
*succeed	amelioration	-0.75	0.81	0.06
artful	pejoration	-1.71	1.33	-0.38
plainness	pejoration	-0.61	1.65	1.04

Table 3: The change in polarity, as well as polarity in each corpus, for each historical example of amelioration and pejoration. Note that *succeed* does not exhibit the expected change in polarity.

Table 8 Table of polarity changes taken from Cook and Stevenson (2010)

This approach brings fine elements to the study of semantic change since it gives insights as to how the meanings drift in terms of connotation, rather than limiting itself to the detection of new words or new meanings. It therefore provides a picture of the process as it unfolds, in a computational and diachronic perspective. The authors also validate their results by having them judged by subjects in a task.

1.3.2.2. Measuring density and variability

The work of Sagi, Kaufmann, and Clark (2009), etxtended in (Sagi, Kaufmann, and Clark 2011) relies on a word space model called *Infomap* [88] in the vein of LSA. Acknowledged cases of semantic change drawn from the literature in historical linguistics are tested in a corpus created on the basis of Helsinki corpus (Rissanen et al. 1994). Outstanding cases are the words *dog* shifting from referring to a specific bread of dogs to referring to the species, in a process of broadening, and *deer* undergoing the opposite shift, from referring to all animals to the specific deer, in a process of narrowing. *Silly*, drifts from meaning "happy" to "stupid", in a process of pejoration.

To test cases, they measure the *semantic density* of these words in a semantic space, and look at the evolution of that density across time, or the *variability* of that density. Density is calculated as the average angle between vectors in a semantic space:

[88] Infomap [Computer Software]. (2007). http://infomap-nlp.sourceforge.net/ Stanford, CA.

"Our method allows us to assess the semantic variation within the set of individual occurrences of a given word type. This variation is inversely related to a property of types that we call *density* – intuitively, a tendency to occur in highly similar contexts. In terms of our LSA-based spatial semantic model, we calculate vectors representing the context of each occurrence of a given term, and estimate the term's cohesiveness as the density with which these token context vectors are "packed" in space." (Sagi, Kaufmann and Clark 2009)

The results show a decrease in density over time for *dog* as its meaning broadens, and an increase in density for *deer* as its meaning narrows. The method is used both on these acknowledged cases and on detection. The following multidimensional scaling plots (Figures 4 and 5) illustrate this:

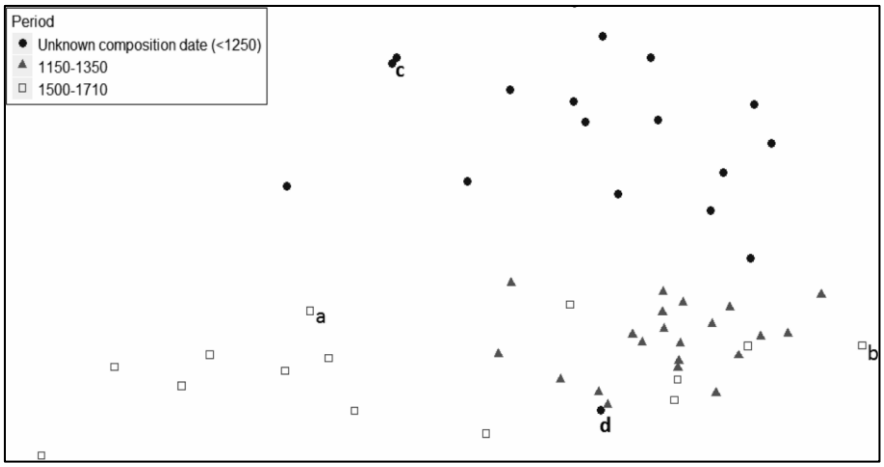


Figure 4 Multidimensional scaling of the context vectors for the word *deer*, taken from Sagi, Kaufmann, and Clark (2011: 177)

In this plot, recent occurrences tend to gather close to each other in space, as the meaning of *deer* narrows. The opposite phenomenon is seen for *dog*, where occurrences tend to disseminate in space as its meaning broadens:

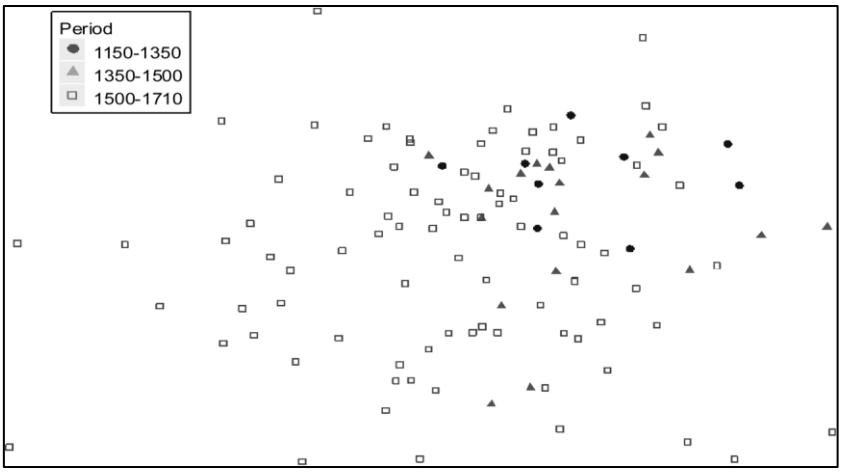


Figure 5 Multidimensional scaling of the context vectors for the word *dog*, taken from Sagi, Kaufmann, and Clark (2011: 176)

1.3.2.3. Topic change and Visual analytics

The most precise accounts of semantic change detection with a context model using graphical representation seem to be those of Heyer, Holz, and Teresniak (2009), Holz and Teresniak (2010) and Rohrdantz et al. (2011).

These works focus on the automatic detection of topic change, since they rely on topic models which are a type of vector model organized by documents. Heyer, Holz, and Teresniak (2009) and Holz and Teresniak (2010) create a measure of *volatility* inspired from econometrics, a discipline applying mathematical and statistical tools to evaluate economic data. The time sliced corpus is composed of daily news taken from the *Wortschatz*, for German sources, and the N*ew York Times* for English sources. To detect topic change, they rely on meaning change, so to obtain a picture of society's most discussed 'hot' topics, leading to a better understanding of the former.

"The assessment of change of meaning of a term is done by comparing the term's global contexts of the different time slice corpora. The measure of the change of meaning is *volatility*. It is derived from the widely used risk measure in econometrics and finance, and based on the sequence of the significant co-occurrences in the global context sorted according to their significance values and measures the change of the sequences over different time slices." (Holz and Teresniak 2010)

Volatility is computed as follows:

1. Built a corpus where all time slices are joined together.
2. Compute for this overall corpus all significant co-occurrences $C(t)$ for every term t.
3. Compute all significant co-occurrences $C_i(t)$ for every time slice i for every term t.
4. For every co-occurrence term $c_{t,j} \in C(t)$ compute the series of ranks $\text{rank}_{c_{t,j}}(i)$ over all time slices i. This represents the ranks of $c_{t,j}$ in the different global contexts of t for every time slice i.
5. Compute the coefficient of variation of the rank series $\text{CV}(\text{rank}_{c_{t,j}}(i))$ for every co-occurrence term in $c_{t,j} \in C(t)$.
6. Compute the average of the coefficients of variation of all co-occurences terms $C(t)$ to obtain the volatility of term t

$$\begin{aligned} \text{Vol}(t) &= \underset{j}{\text{avg}} \left(\underset{i}{\text{CV}} \left(\text{rank}_{c_{t,j}}(i) \right) \right) \\ &= \frac{1}{|C(t)|} \sum_{j} \underset{i}{\text{CV}} \left(\text{rank}_{c_{t,j}}(i) \right) \ . \end{aligned}$$

Figure 6 Computing volatilty, taken from Holz and Teresniak (2010)

This measure is independent from word frequency and is efficient in large scale detection. However it does not dig into the unfolding of semantic change processes in detail. Using the rank series of every co-occurrent for a term, referred to as its "global context" in the article, allows for an extraction of topics that change even if they are not "new" to the corpus. Indeed rank change shows change in semantic importance in a word's co-occurrent list, independently of its frequency. Rank is related to the hierarchical order of appearance of co-occurrent words, and therefore rank variation has more to say about words context change than frequency alone. Volatility and frequency variation provide two measures connected with the media and the events it covers, as well as with the meaning changes that are induced, temporary or not. For highly connoted words such as *Irak*, frequency and volatility are not correlated, as shows the following graph:

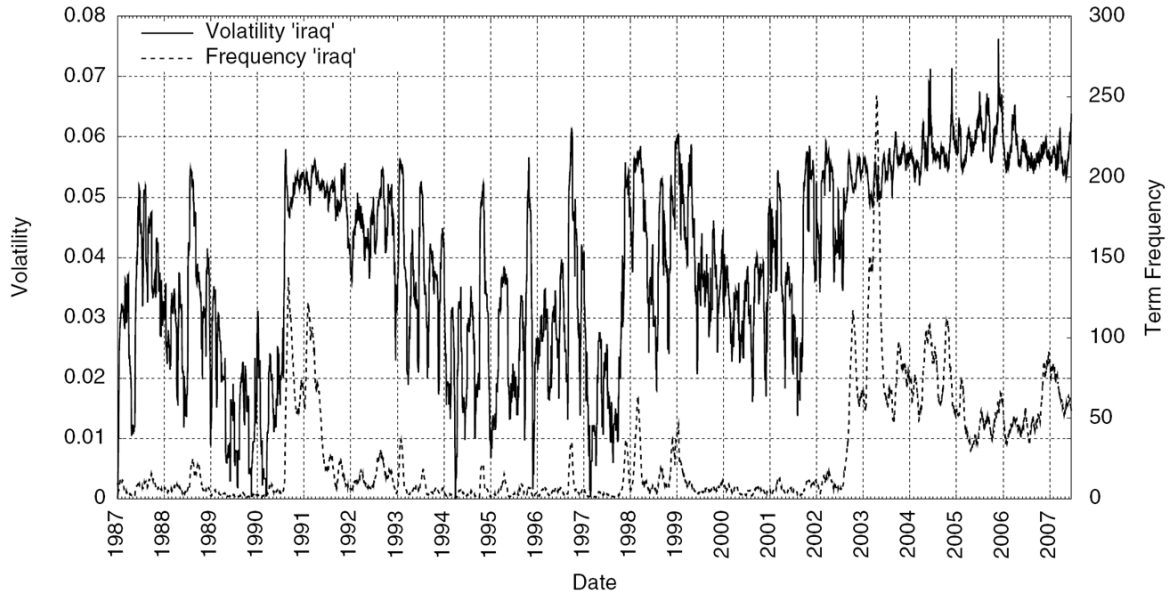


Figure 7 Thirty day volatility and frequency of the word Irak, from 1987 to 2006, based on the *NYT*, taken from Holz and Teresniak (2010: 335)

The authors also offer a graphical interface for the representation of results, called the "volalyzer", a volatility analyzer[89]. The interface is flexible, allowing for time-span selection, and the visualization of frequency and volatility[90]. Visual analytics is a powerful tool to represent large amounts of complex data, and I have tried to integrate some of it in this work, through the collaboration with a student specialized visual analytics research. However, Holz and Teresniak (2010) offer the most advanced freely available framework to date for the representation of topic and meaning change with visual analytics.

1.3.2.4. LDA and visual analytics

Rohrdantz et al. (2011) also combines visual analytics and NLP to address semantic change issues in diachrony. The context model they use is LDA (Latent Dirichlet Allocation). They also use the *New York Times* press corpus. They focus specifically on English words that gain a new sense by extending their context, in the area of technology. The method is based on word context rather than word/document representations as it is in topic models. Indeed, this

[89] http://aspra23.informatik.uni-leipzig.de:8400/blazeds/volalyzer.swf

[90] However the interface does not offer the calculation of volatility for each query, as at the time of access it only provided already stored results.

approach allows them to study semantic change in more detail, as argued by the authors. The corpus is trained with LDA, and contexts (with a window of 25 words before and after the target word) are extracted and associated with a time stamp:

"For the set of all contexts of a key word, a global LDA model was trained using the MALLET toolkit2 (McCallum 2002). Each context is assigned to its most probable topic/sense, complemented by a specific point on the time scale according to its time stamp from the corpus."(Rohrdantz et al. 2011: 2)

By establishing a more detailed list of contexts for each word (with a window of 25 words before and after target word as well as time stamps), they can provide a picture of which sub-meanings may take over others. To do so, they also rely on graphical representation, as in the following visual:

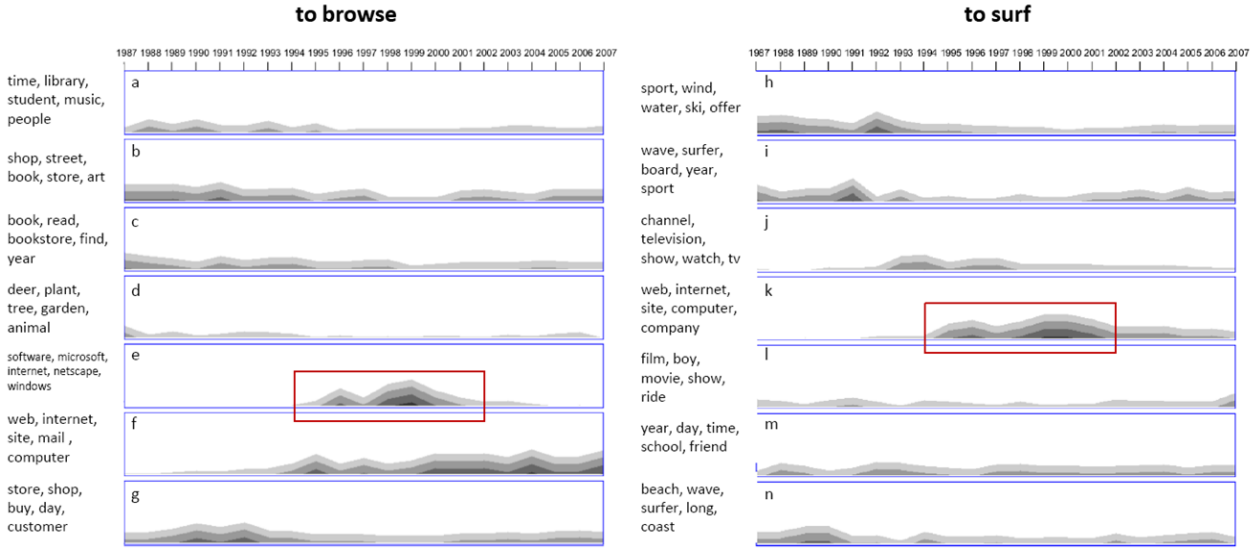


Figure 1: Temporal development of different senses concerning the verbs *to browse* (left) and *to surf* (right)

Figure 8 Visualization of *surf* vs. *browse* taken from Rohrdantz et al. (2011)

On Figure 8, we can clearly see that the senses of *browse* related to the set "software, microsoft, internet, netscape, windows" in e) and the senses of *surf* related to the set "web, internet, site, computer, company" in k) both flourish from 1994 to 2002. The y axis corresponds to percentages of word contexts for each sense. Again, these senses are highly related with real world events, here the introduction of the Netscape Navigator in 1994. I have

also mentioned the relationship between the meaning of *navigation* in French and the Netscape Navigator in Boussidan and Ploux (2011). Indeed, meaning shifts related to the internet have had a strong cross-linguistic impact.

The evaluation of results is conducted with regards to dictionary definitions. The authors claim that LDA performs better than LSA at detecting semantic change.

1.3.2.5. Distributional semantics with a Web corpus

Gulordava and Baroni's (2011) article entitled "A distributional similarity approach to the detection of semantic change in the Google Books Ngram[91] corpus" applies co-occurrence methods to detect semantic change in the American English 2-grams Google corpus, composed of books, divided in two time slices extracted from the 1960s and the 1990s. The authors rely on the logarithmic ratio between frequencies of word co-occurrence in the two time slices. They apply a measure of Local Mutual Information (LMI) on the co-occurrence matrices to detect the similarities between words in the two time slices. Therefore, a target word has a vector for the 1960s and one for the 1990s, which can be compared within the same space, by computing the cosine of their angle (similarly to what Sagi, Kaufmann, and Clark 2009 described). Similarity is rated on 1, the highest scores indicating that the word is stable and the lowest scores that it may have undergone semantic change or at least a change in the use of the word. The authors chose to conduct the experiments on a sample of mid-frequency words. By doing so, they are likely to detect changes in commonly used words; however, as stated earlier, low frequency words are also of a particular interest and it is regrettable that they are not dealt with. The low similarity scored words are then compared to a database containing human ratings. The correlation between similarity values and figures obtained via human judgments is then computed to assess results. Nevertheless, some of the low similarity words were judged unchanged by raters, as they rather showed expansion of context of use with no visible impact on meaning. The authors give the examples of the words *sleep* and *parent*:

[91] A n-gram is a sequence of n-words.

	'parent'	'sleep'
60s	p. company 2643	deep s. 3803
	p. education 1905	s. well 1403
	p. corporation 1617	cannot s. 1124
	p. material 1337	long s. 1102
	p. body 1082	sound s. 1101
	p. compound 818	dreamless s. 844
	common p. 816	much s. 770
90s	p. families 17710	REM s. 20150
	single p. 10724	s. apnea 14768
	p. company 8367	deep s. 8482
	p. education 5884	s. disorders 8427
	p. training 5847	s. deprivation 6108
	p. involvement 5591	s. disturbances 5973
	p. family 5042	s. disturbance 5251

Figure 9 Examples of the top weighted 2-grams containing 'sleep' and 'parent'. Taken from Gulordava and Baroni (2011)

These results, which the authors consider erroneous, show a shift in perception of the concepts *sleep* and *parent*. The change in direct linguistic context of use reflects sociological changes. In the case of *sleep*, the shift is from a personal perception to a more medical and scientific perception. In the case of *parent*, the shift is from a stronger work context to a stronger family context, with the notable apparition of *single*, reflecting sociological change. This type of sociological change is for some linguists at the heart of semantic change over long periods of time, typically Centuries:

"*Les modifications de notre vie et de nos usages atteignent notre vocabulaire* : madame, mademoiselle, famille, domestique, dîner *n'ont plus le même sens qu'au XVI*^{e} *siècle*."(Huguet 1934: v)

These result point to the limit of using context change to detect semantic change, but they also show that tools for semantic change can be used to detect perception trends.

Indeed, whatever the mathematics behind detecting semantic change with context change markers, this problem seems to always arise. At this point, human analysis is required. However, one of the supplementary tools that can guide this analysis is visual representation.

PART II THE ACOM MODEL AND EXTENSIONS FOR DIACHRONY

Chapter II.1: What is ACOM?

In the same perspective as the studies just described, this work relies on a vector model providing geometrical spaces and representations of data. This model is called the Automatic Organizing Contexonym Model (ACOM)(see Ji 2005; Ji, Ploux, & Wehrli 2003; Ploux, Boussidan & 2010), a model derived from the Semantic Atlas (SA)(see Ploux 1997; Ploux & Victorri 1998; Ploux, Boussidan & Ji 2010). The SA is a vector model based on correspondence factor analysis. It is related to the vector models described in Part I but factor analysis models have their own history.

2.1.1 Factor analysis models

2.1.1.1. Factor analysis in sciences, social sciences and psychology

Factor Analysis, and its variants Correspondence Factor analysis, Principal component analysis and Multiple Component Analysis are statistical methods that create multidimensional spaces on the basis of matrices, and are useful to group items which have a similar profile and isolate the ones that differ. Factor analysis has been used since the beginning of the XXth century to explore and represent data, detect the degrees of variances in data sets, and make predictive models in various areas. Its widespread applicability across disciplines is due to the technique that is essentially designed to compare large data tables and matrices on the basis of a common scale. Factor analysis hierarchically sorts all the relationships between the lines and the columns of the matrix and creates a space to represent these relationships.

One of the earliest uses of factor analysis was in psychology when psychologist Charles Spearman implemented it to research the existence of a measurable mental ability, a discipline later known as psychometrics. However, beyond psychology, it has been a tool for sciences at large. It has been broadly used in physical sciences to study populations in terms of variables, for instance in agronomy, ecology, or biometry, as in (Ferrand et al. 2009) who evaluate water nitrate evolution. It has also been widely applied in sociology as in (Bourdieu 1979) who determines the relationships of different groups within a social space in terms of variables such as the economic capital or the cultural capital. This aspect has been widely used in marketing, surveys and opinion polls, to map populations, customs, habits, brand consumption, etc. Indeed, factor analysis gives a relatively precise map of these data:

"L'analyse factorielle permet de faire surgir la structure des données, la façon dont chaque variable se situe par rapport aux autres, de manière différentielle et relationnelle. La sociologie structurale de type bourdieusien, et Bourdieu lui-même, en ont fait un outil de représentation puissant, au service de leurs thèses : l'outil permettait de mettre au jour la structure multidimensionnelle et relationnelle du champ étudié. [92]" (Dozo 2008)

2.1.1.2. Factor analysis in Linguistics

Factor analysis has been applied to text, in literary analysis and discourse analysis to determine the vocabulary and style of an author, evaluate the evolution of their style over time, or compare their style with the one of other authors. It is a tool in linguistics, discourse analysis, diachronic studies and data mining at large.

In Europe and France in particular owing to the influence of Benzécri (especially 1980) in social sciences, factor analysis methods have been integrated to statistical co-occurrence software to classify data and add a graphical layer of representation to discourse analysis.

The use of factor analysis and related statistical methods is at the root of the French disciplines of "Sémiométrie" (Lebart, Piron, and Steiner 2003), "Textométrie", "Logométrie" and "Lexicométrie" [93] which are branches of lexicology, text statistics, discourse analysis and sociology. These works rely on statistical software including factor analysis such as LEXICO[94], WEBLEX[95], HYPERBASE[96], ALCESTE[97], SATO[98], XAIRA[99] or DtmVic[100].

---

[92] "Factor analysis makes data structure appear and the way each variable is situated in relationship to the other ones, in a differential and relational way. Structural sociology of a Bourdieusian type, and Bourdieu himself, made it become a powerful representation tool, supporting their arguments: the tool enabled them to reveal the multidimensional and relational structure of the field under study."

[93] These disciplines are already mentioned in Part I, Chapter 2, section II.2. They are translatable as semiometrics, textometrics, logmetrics and lexicometrics

[94] http://www.tal.univ-paris3.fr/lexico/

[95] https://weblex.ens-lsh.fr/

[96] http://ancilla.unice.fr/~brunet/pub/hyperbase.html

[97] http://www.image-zafar.com/index_alceste.htm

[98] http://www.ling.uqam.ca/sato/

Research using statistical methods and factor analysis or factor analysis for discourse analysis purposes may be found in the International conference on statistical analysis of textual data.[101]

In psycholinguistics, a factor analysis model similar to the SA[102], has been applied to study the mental lexical organization, on the basis of *proxemy* (proximity) links, which are equivalent to semantic similarity links (see Gaume et al. 2008).

In semantics, the main idea is that we can group similarities in a space, and therefore classify words in terms of semantic features or similarities around axes.

2.1.1.3. Semantic distance and clustering

Once a factor analysis generated space is obtained, the distances between the points (which may be words, concepts, or other items) can be measured. Points close to each other tend to denote similarity. Points that group together may then be classified in sets or "clusters" on the basis of their mathematical relationship. Clustering allows for grouping and organizing data of the same kind. There are different algorithms to cluster data, either hierarchical or using K-means. At this point, clusters may be compared, relying on their centroids, or on their elements.

2.1.2. The Semantic Atlas (SA) and the Automatic Contexonym Organizing Model (ACOM)

2.1.2.1. How it works

The earliest article describing the Semantic Atlas model dates back from 1997 (see Ploux 1997). In this article, the author describes how a mathematical model may handle semantic data, using topological distance as the main indicator of semantic similarity relationships, in the line of distributional semantics. The space that is generated with Correspondence Factor Analysis is composed of coordinates (a point for a clique) organized around axes. In these

---

[99] http://www.oucs.ox.ac.uk/rts/xaira/

[100] http://www.dtmvic.com/

[101] Les journées internationales d'analyse statistique des données textuelles (JADT)

[102] And relying on the same database as concerns synonymy relationships.

spaces the first axis represents the most important dimension, and the second, third, and subsequent axes are decreasingly important. There are three levels of granularity: the clique, the envelope and the cluster levels. The cliques are the coordinates which are grouped into envelopes that represent words, and clusters which are thematic sets of either cliques or envelopes.

The original SA model therefore provides maps to navigate within very closely related meanings in a continuous paradigm. Additionally, it offers the possibility to combine requests and thus to cross results within the same map and create a space combining completely different meanings. The model has evolved over the years, and has been substantially enriched with extensions for new applications as well as new databases. It has benefited from the contributions of numerous students. First, it has been applied to translation, by extending its capacity to combine requests, and second, it has been applied to contextual analysis by replacing the handling of the synonymy relationship by the one of co-occurrence.

What the original SA offered was a treatment of synonymy and polysemy, extended to a bilingual version. H.Ji who carried out his PhD at the L2c2, focused his work on adapting the SA to corpus analysis, by replacing the relationship of synonymy in the model by the one of co-occurrence, giving birth to ACOM (Ji 2005). Ji called co-occurrent words in context "contexonyms" and implemented software that deals with second order co-occurrence while offering offers some amount of flexibility in settings, so the user may adjust it to the type of data they are dealing with. ACOM's basic process is summed up in (Boussidan and Ploux 2011) as follows:

> "For each slice[103] t, a word-association table is constructed using all headwords (see Ploux, Boussidan, and Ji 2010 for a complete methodological description). Each headword Wi t (1_i_N, where N is the total number of types in the corpus slice) has children (cjs) that are arranged in descending order of co-occurrence with W^{i}_{t}[104]:

[103] Here a « slice » refers to a chunk of text.

[104] 1Children with co-occurrences under a 10,000th of the global frequency of the headword Wi t are removed to reduce noise.

W^i_t : c1; c2; : : : ; cn

We apply two factors to filter this table: α where $0 \leq \alpha \leq 1$ to eliminate the rarely co-occurring children of W^i_n:

W^i_t: c1; c2; : : : ; ck

where k = n α and n is the original number of children of W^i_t , and β where β ($0 \leq \beta \leq 1$) to cut off rarely co-occurring of children of c_j:

(c^m_j : g1; g2; : : : ; gl($1 \leq j \leq k$; l = m β))

On the basis of that table, cliques are calculated."

(α) allows for selecting the number of search words, (β) for selecting their children for second order co-occurrence and (γ) for selecting the number of cliques.

This way, ACOM provides results very similar to statistical software like Lexico, but is the only model doing so with a clique implementation. However, the method used by Rohrdantz et al. (2011) implements a clique-like technique using domains rather than contexonyms. By combining the two aspects, ACOM offers a highly flexible and detailed interface to represent textual meaning and conduct statistical exploration not only on words but on cliques as well.

2.1.2.2. Cliques

Indeed, the originality of the SA relies on the concept of cliques. Conceptually, a clique is a minimal unit of meaning, at very fine grained level. In the original SA, cliques are lists of words which are related to each other by synonymy (in the broad sense), in ACOM they are lists of co-occurrent words. Mathematically, a clique is an object that designates a maximal, complete and connected sub-graph. The graph here is a set of words (the nodes) and relations (the arcs) that link up words. For instance, a request on the online synonymy model on the word *house* provides 47 associated words and organizes them in 27 cliques. The cliques are subsets that may be shared across words. The following associated words are organized in the corresponding set of cliques:

Associated words[105]:

house : abode, accommodate, accommodation, accommodations, address, audience, bed, bed down, billet, billet on, board, building, business, business firm, clan, company, concern, cover, domicile, dwelling, establishment, family, firm, habitation, harbour, home, household, legislature, line, lodge, mansion, menage, parliament, partnership, place, planetary house, put up, quarter, quarter on, residence, shelter, sign, sign of the zodiac, tenement, theater, theatre, tribe

Cliques:

1 : abode, address, domicile, dwelling, house, residence
2 : abode, domicile, dwelling, habitation, home, house, residence
3 : abode, domicile, dwelling, home, house, place, residence
4 : accommodate, bed, bed down, house, lodge, put up
5 : accommodate, bed, board, house, lodge, put up
6 : accommodate, bed, board, house, lodge, quarter
7 : accommodate, billet, billet on, house, lodge, quarter, quarter on
8 : accommodate, house, lodge, shelter
9 : accommodation, accommodations, dwelling, habitation, house, residence, tenement
10 : address, cover, house
11 : audience, house
12 : building, house
13 : business, company, concern, establishment, firm, house
14 : business, house, line
15 : business firm, firm, house
16 : clan, family, house, tribe
17 : company, firm, house, partnership
18 : cover, house, shelter
19 : dwelling, establishment, house
20 : family, home, house, household, menage
21 : family, house, line
22 : habitation, house, mansion, residence
23 : harbour, house, lodge, shelter
24 : house, legislature, parliament
25 : house, lodge, place, quarter
26 : house, mansion, planetary house, sign, sign of the zodiac
27 : house, theater, theatre

Cliques provide a representation of meaning going beyond the lexical unit and thus also allow for sense navigation inside and outside the word boundary. In effect, a meaning subset represented by one or many cliques may be shared by several words and form the link

[105] Source : http://dico.isc.cnrs.fr/dico/en/aclique?r=12010820_1326050157_714227015

between them. In this respect, cliques represent a conceptual level of meaning almost independent from words. They reflect the internal complexity and organization of the semantic structure of a word as well as the structural relationship of the word with others words (either within a lexicon, across lexicons or within a corpus). Therefore they are useful to analyse semantic networks at the level of the lexical unit and at the level of sets of lexical units. In the example for the word *house,* clique n° 20 (family, home, household, menage) is part of the networks of 4 words. It is the unique clique for *menage* and for *household*, and is part of a set of 3 cliques for *family,* and another set of 3 cliques for *home*.

For a given word there is an underlying topology to the set of associated cliques that allows for navigation from a semantic value to another in a continuous paradigm. Cliques may organize meanings into value types, such as physical, emotional or perceptual aspects. Since each clique is connected to the next one by one synonym in common, a progressive transition from a meaning to another at subtle semantic levels is made possible. Therefore cliques provide an extremely detailed tool in the treatment of polysemy. Since the reorganization of a word's polysemy in terms of hierarchical value of meaning subsets is one of the major mechanisms of semantic change, this model provides an adequate framework for the modeling of these mechanisms as will be detailed later on.

The closest existing tool to cliques in computational semantics is the concept of 'synsets' developed in Wordnet[106] (Fellbaum et.al 1998) a hierarchical model for meaning representation. Wordnet is an electronically available lexical database, originally developed at Princeton University, which organizes meanings into hierarchical sets in terms of semantic and lexical relations. The synsets, which are tree-like conceptual structures of words, rely on synonymy, hyponymy, hyperonymy and domain. The number of synsets is an indicator of the word's degree of polysemy. However, this organization into units is subjective and imposed by the linguists who work on that system. The English version of Wordnet is a substantial database, while versions in other languages are still in the course of development.[107] But

[106] http://wordnet.princeton.edu/

[107] Wordnet has a high lexical coverage in English, however most of its equivalents in other languages are translations of the English version and not language specifically built models. The Eurowordnet project encompasses Dutch, Italian, Spanish, German, French, Czech and Estonian, and many developments are in progress. These versions have a much poorer lexical coverage than the English one does.

cliques differ from synsets as there is no human decision in the organization of concepts, and all concepts are treated equally with the same mathematical and statistical implementation. Cliques are an infra-linguistic tool, at the level of representation, in which no linguistic judgment intervenes. Therefore, cliques are suitable for an exploratory approach of language rather than a theoretically lead approach.

ACOM cliques show how the word's contexts overlap or not and allow for discrimination of contextual sets rather than meaning sets. If we submit the same request for *house* while running ACOM on a sample press corpus, we get the following output:

Related words :

house: big, brother, built, business, city, close, commons, corner, court, dark, daughter, entered, family, front, garden, gentleman, live, lived, master, met, person, public, road, rooms, servants, street, summer, walked, window, windows

Cliques :

1 : big, front, house,
2 : brother, daughter, family, house,
3 : built, business, city, house,
4 : built, city, house, road, street,
5 : built, family, house,
6 : business, city, house, public,
7 : city, court, house, public,
8 : city, entered, house, street,
9 : city, house, lived, street,
10 : close, dark, house,
11 : commons, house, public,
12 : corner, dark, house,
13 : corner, front, garden, house, street, window,
14 : corner, front, house, road, street,
15 : court, house, person,
16 : family, house, person,
17 : front, garden, house, street, walked, window,
18 : front, garden, house, street, windows,
19 : front, house, road, street, walked,
20 : garden, house, summer,
21 : gentleman, house,
22 : house, live, lived,
23 : house, master,
24 : house, met, person,
25 : house, met, road, street,
26 : house, rooms, street, windows,
27 : house, servants,

With contexonyms, the cliques contain words directly related to the context of use of *house*: the place in which the house is (street, city), the people related to the house, the things the house is composed of.

2.1.2.2.1. ***Building the space***

On the basis of the list of cliques, a matrix is created where the cliques are lines and the words are columns. Then a semantic space is generated from the matrix via Correspondence Factor Analysis. In contrast to other word space models that work with word/document matrices or word/paragraph matrices the SA does not encounter frequent problems of size limitation. The Semantic Atlas has no size limits regarding corpora. It encounters size limitations regarding the quantity of data that may be represented in a space. However, it is rarely the case that one may want to represent enormous amounts of data, since the maps soon become saturated. On the basis of the matrix, the coordinates of cliques are calculated with Correspondence Factor Analysis. Unlike models that assign a vector or a node to a word, geometrical modeling associates a delimited space with a word in a continuous fashion. $\chi 2$ distance is used between words rather than Euclidean distance as it renders geometrical organization more accurately (see Ploux 1997). Words are represented by the envelopes containing cliques. This way, word envelopes may overlap or not, thereby providing a very instantaneous visual tool for definition or translation purposes. Therefore, points stand for cliques and envelopes for words, but there is an additional degree of representation and classification: clusters.

2.1.2.3. Clusters

With clusters, the model allows for categorization: a hierarchical classification algorithm generates clusters that organize senses into thematic sets. The number of clusters (2 to 6) may be defined by the user to shed light on the specific aspects they wish to study. In the online version, the number of clusters is set to 3 by default. Users may cluster either cliques or words, on the basis of the center of gravity of envelopes, directly from the output.

> "Clusters:
>
> -Separate values within a monolingual representation on scales such as perception, action, emotion, etc.
>
> -Detect overlapping and non-overlapping values across two languages.

-Separate values of a word's meaning according to its context of use.

Furthermore, as clusters allow for overlap detection, they may be used to analyze other types of linguistic data." (Ploux, Boussidan, and Ji 2010)

2.1.2.4. Maps

Cliques, words and clusters are represented in a flexible map generated in Java. It is flexible at different levels. First, the user may set the percentages of co-occurrent words to retain with the parameters (α), (β) and (γ)[108]. Once the map is generated the user may change the clustering type, either based on cliques or words, and change the number of clusters. They may also change axes to visualize the other dimensions of the shape.

2.1.2.4.1. *Example of a Semantic Atlas map for the representation of synonymy and polysemy*

Figure 10 represents the word *bright*. The representation was obtained with the Semantic Atlas in its original version based on synonymy relationships. We have shed light on 4 clusters which group sub-meanings that compose the polysemy of *bright*: *clear* (in yellow), *clever* (in red), *happy* (in blue) and *luminous* (in green). The envelopes with gradations of colours between clusters show how the transition may be done between these sub-meanings. The transition between *clear* and *clever* is performed via *blazing*, the one between *clever* and *happy* via *apt*, the one between *happy* and *luminous* via *beaming* and the one between *luminous* and *clear* via *glowing*.

[108] As stated earlier (α) allows for the selection of the number of search words, (β) for the selection of their children for second order co-occurrence and (γ) for the selection of the number of cliques

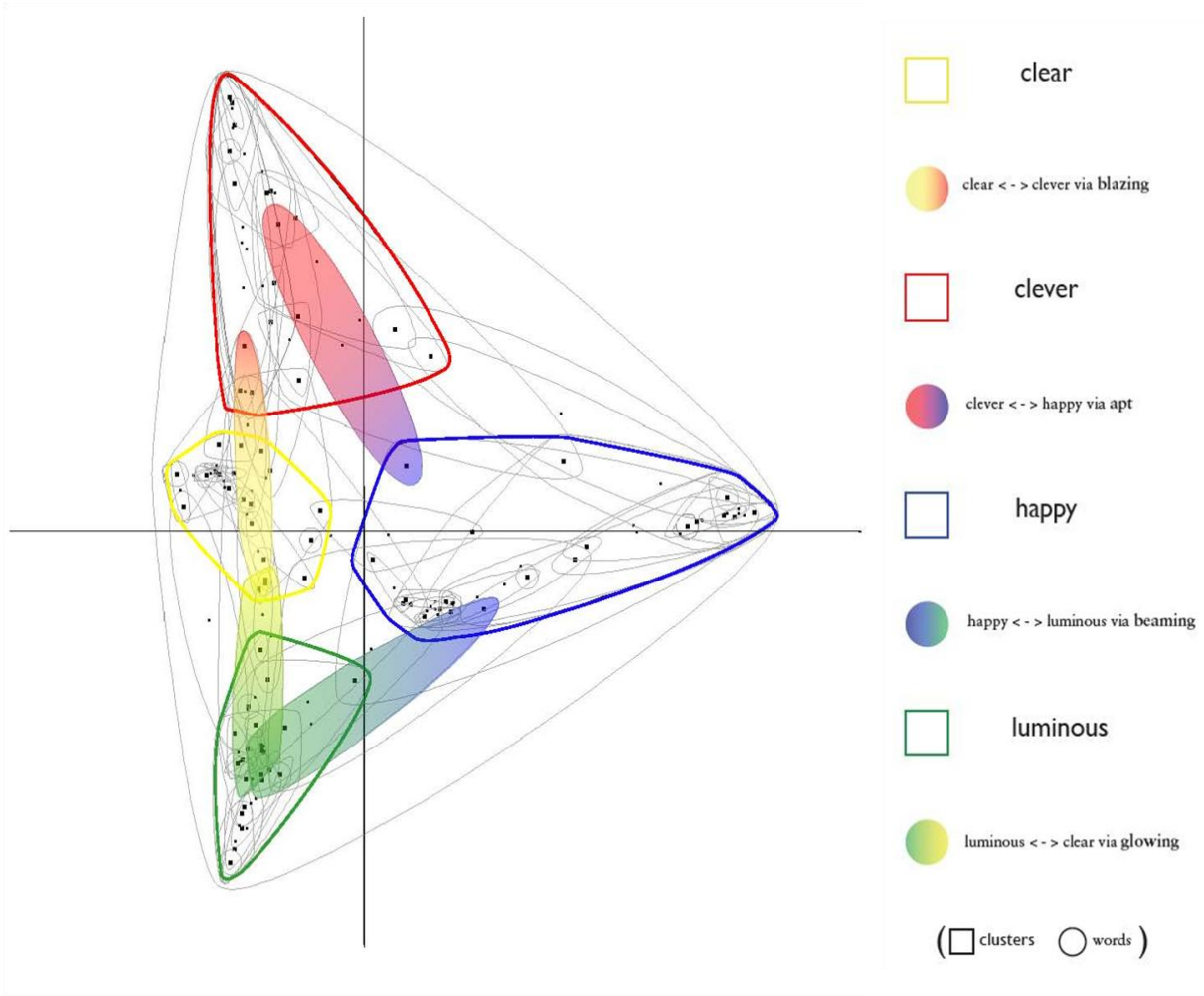


Figure 10 Representation of the polysemy of the word *bright* with the Semantic Atlas[109]

2.1.2.4.2. *Example of an ACOM map*

In the same way that the map for *bright* separates sub-meanings and sheds light on the organization of polysemy, an ACOM map separates contexts and sheds light on the organization of use and connotation. In Figure 11, a representation of the word *conductor* in a sample press corpus, clusters show distinct contexts of use for the word: the conductor of an orchestra, of a bus, or an electric conductor. The concepts of station and tickets act as

[109] Presented on a poster at the Language Resources and Evaluation Conference in Malta, 2010. Collaboration with Charlotte Franco (computer programming) and Anne-Lyse Renon (Graphic Design, Data visualization).

connectors. The model therefore extracts the different contexts of use from the corpus, and organizes them.

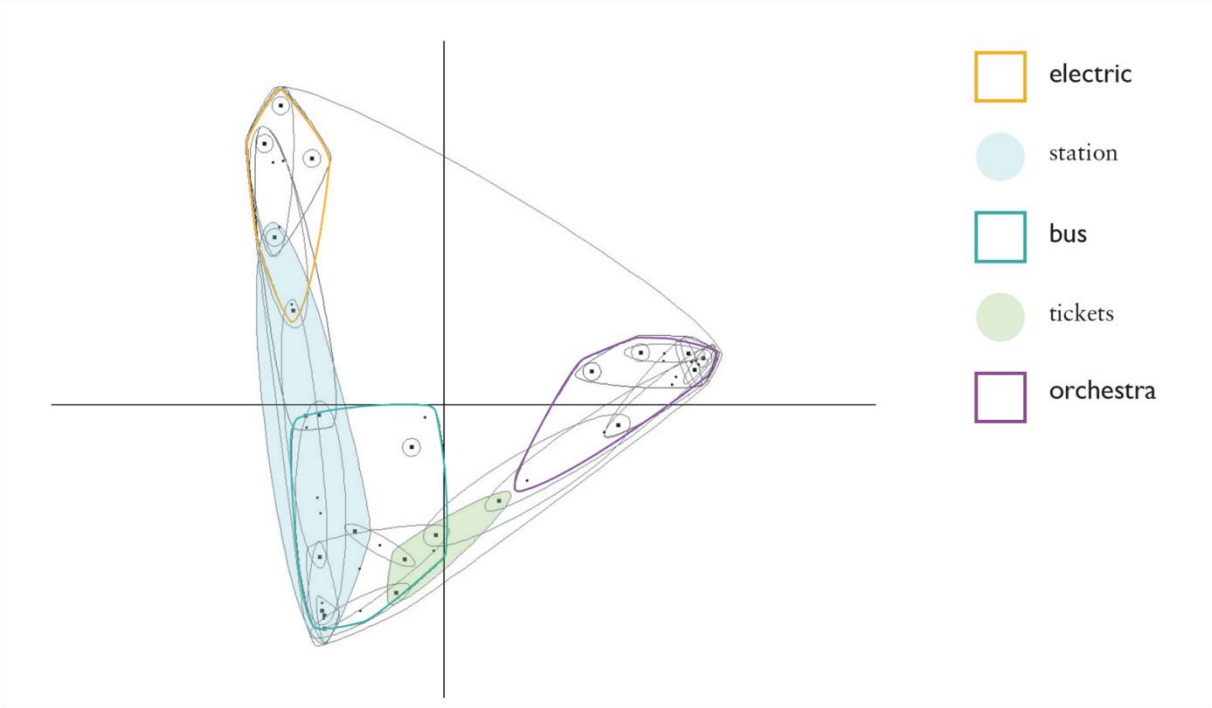


Figure 11 Representation of the word *conductor* and its contexonyms with ACOM[110]

2.1.2.5. Limitations of the SA and Factor analysis models

With factor analysis based models, a series of issues related to representation enter into play, since the spaces which are obtained and their organization largely depends on the type of factor analysis being used, on the settings for data treatment, and on the mathematical choices made to obtain the final space. For instance, Ploux and Victorri (1998) note that using Euclidean distance may create optical illusions of proximity or distance between items even if they are not close or far to one another. The authors replace the Euclidean distance by a $\chi 2$ distance to fix that problem. Similarly to vector models, there is no current standard in the field allowing for replication or comparison of results across models. Moreover,

[110] Presented at the Language Resources and Evaluation Conference in Malta, 2010. Collaboration with Charlotte Franco (computer programming) and Anne-Lyse Renon (Graphic Design, Data visualization). Based on an example put forth by S.Ploux and H. Ji.

multidimensional models organize data into several dimensions bringing to the front the most important elements. In some instances the elements contained in further away dimensions may contain interesting elements, especially in a diachronic perspective, if said elements change status and come to the front. This makes the diachronic uses of multidimensionality difficult as it involves very large sets of data. In the same way, computing becomes heavy when working at multi-entry word level not to mention sentence level. Finally, clustering depends on separate algorithms that have a great deal of impact on the way data is sorted.

Overall, ACOM offers a framework that is rich in that it keeps a large amount of detail in its representations, and is flexible enough to adapt to different styles of text and across several languages. Since ACOM provides an excellent treatment of polysemy in synchrony, it is adequate to think that it can provide such a treatment for the dynamics of polysemy in diachrony that underlies semantic change.

What the variation of co-occurrence patterns may show us is the dynamics of use in a given corpus. It is clear to many that there is no such thing as a purely stable pattern of use. However a “regular” degree of variation may be characterized, from which unexpected variations may be deduced.

Chapter II.2 ACOM extensions for diachrony

2.2.1 On using co-occurrence patterns and ACOM to study diachronic changes

2.2.1.1. Levels of analysis

The idea of working with co-occurrence patterns to detect and analyze diachronic evolutions relies on a view of meaning based on the observation of semantic networks and their dynamics rather than the observation of lexical units separately. The networks give hints at two levels: first, at the level of language, or what a word's lexical semantic contexts are and how they are organized, and second, at the level of language in use, that is to say the particular context of use observed in society. Therefore there are two types of context being dealt with: the linguistic context, at the level of language, and the language use context at the level of pragmatics and sociology. The underlying hypothesis is that those two levels work together and are interdependent. A modification in a word's context of use has an impact on its semantic network in language, and that impact may be either temporary or persistent, and may profoundly modify the semantics of a word in time.

ACOM combines tools of statistical analysis that are traditionally used in computational linguistics with graphical representations traditionally provided by Correspondence Factor Analysis based space models. Both tools are adequate to deal with diachrony phenomena. At the conceptual level however, one may ask in what way the changes in linguistic context of words across chunks of corpora may be an indicator of semantic change. This is a sensitive topic since these changes, taken broadly, are not in themselves indicators of semantic change. In fact, they can become indicators above certain thresholds of change. However, these thresholds have never been set by previous research and therefore this work steps into an unexplored domain with regards to thresholds. But what are these thresholds? Change in meaning cannot be defined by a unique mathematical or statistical threshold. It is anchored in several mechanisms: the first is the substantial modification of a word's semantic network, the second is the measure of the spreading of that modified network to a large enough community of users, and the third is the stabilization in time of that modified network. These three levels

may interact, or may be impacted differently and the relationship between them differs on a case by case basis. Consequently, this research has to delve into case analysis – which will be hereinafter referred to as the *microscopic level* - and cannot be dealt with uniquely at large scale – hereinafter referred to as the *macroscopic level*. This coining is inspired from Nerlich and Clarke (1988) who define micro- and macro- dynamics, as the level of the single innovation and the level of the form and structure of language. Here, the microscopic level is the one of detailed case analysis; the macroscopic one corresponds to trends over whole corpora. The microscopic level of detailed analysis must be taken into account at the macroscopic level and vice-versa. The method I have chosen for this work relies on moving back and forth between the macroscopic and microscopic levels, and test whether specific examples reveal what may apply to large sets of word changes, and vice-versa. Indeed, manual analysis of detailed examples in the literature have provided typologies and frameworks of analysis that demand testing at a larger scale, while computational studies at a large scale have provided an understanding of patterns and mechanisms that demand testing at smaller scales.

2.2.1.2. On co-text

Co-text is the most accessible and easy to use data to access word meaning change in its linguistic and pragmatic context. It is anchored in the here and now, and this is both a perk and a pitfall. Indeed, the co-text in a press corpus is directly in touch with extra-linguistic realities; these realities deeply colour results and therefore these results are only valid within that specific corpus/context. However, phenomena that may be observed with a specific co-text can be researched in other corpora to check whether the results go beyond corpus-specific manifestations. Moreover, one can choose what context window they are working with. ACOM uses the sentence by default, as the minimal context window. As defended by Mejri and Stern, among others, the sentence is the base level to analyse word meaning in text:

"Nous croyons que la solution réside dans une approche qui abandonne le postulat selon lequel l'unité minimale de signification est le mot. Même si le mot peut être un point de départ à l'analyse, sa signification ne peut pas être conçue indépendamment du contexte dans lequel il est employé. Ce contexte peut avoir plusieurs configurations. Il est d'abord phrastique, et c'est le niveau de base qui détermine la valeur réelle de toute unité lexicale. S'y

ajoutent les domaines, les niveaux de langue et l'environnement contextuel.[111]" (Mejri 2006:3)

"It is doubtful whether such elements of context are normally, connected with single words in speech; it seems more probable that they rest upon entire utterances, or perhaps sentences. The experiments with isolated words cannot always show what actually happens in connected speech." (Stern 1931: 63)

In accordance with these critiques, case studies that target a specific word need to be conducted at sentence level and they also need to take into consideration the movements observed in the networks of co-occurring words in these sentences.

The degree of variation of the co-text partly relies on the natural richness of use of the target word, which differs for each word. This richness is related to how polysemous the words are, how idiomatic they are (that is, if they are employed in several idiomatic expressions with radically different meanings) and on the variety of contexts they can be used in, as well as their flexibility, including metaphorical expressions and transfers. Words have different combinatorial profiles, grammatically and semantically.

The ACOM model captures these combinatorial structures automatically in text, thanks to the system of cliques. It organizes combinations in continuous meaning types. Since the source for the model is produced text, the obsolete combinations given by dictionaries do not appear in the output. This facilitates the diachronic approach, at the stage of human analysis. However, word order is lost in the representation, but can be retrieved in the corpus.

2.2.1.2.1. ***Precautions with co-text***

When using a model, there are a few precautions to take with combinatorial structures. The combinatorial possibilities of words are not equal. In the same way that transitivity restrictions and rules apply to verbs, combinatorial restrictions apply to nouns. For instance, a

[111]"We believe the solution lies in an approach that abandons the assumption that the minimal unit of meaning is the word. Even if the word can be a starting point for analysis, its meaning cannot be conceived independently of the context in which it is used. This context may have several configurations. It is first sentential, and this is the basic level that determines the actual value of any lexical item. Added to this are domains, registers and contextual environment"

word like *body* has an extended range of use since it can be used to refer to animate or inanimate bodies, to one or several persons or things as well as in a wide range of thematic contexts (anatomy, politics, text, etc.). However, a word like *guava* only refers to a distinct fruit, and a word like *looter* only refers to a person who is looting. Those two words have few chances of undergoing a multiplication of meanings, unless they are consciously used in a novel construction for stylistic effect. If so, the change is brutal, and therefore easy to detect. Gradual change, however, often happens unseen and is more challenging to detect.

A word like Fr. *monde* ("world"), for instance, is mainly combined with other adjectives or verbs, coming before or after it in expressions, under three distinct meanings: a) *terre, b) univers* and c) *personnes* ("earth, universe and people"). The quantity of possible contexts we may find the word *monde* in is therefore high. Examples of combinations are: *un vaste monde* (a. adj. + n.) *conquérir Le Monde* (a. vb+n.), *le monde animal* (b. n+adj), *un monde peuplé de …* (b. n.+vb.), *entrer dans un monde* (b. vb.+n.), *un monde fou* (c. n. adj.), *connaître du monde* (c. vb.+n.)[112]. These combinations are rich both in terms of polysemy and in terms of possible positions the verbs and adjectives may assume. Each of these types of combinations, or pattern, appears in numerous expressions. For a given word, the possible types of combinations differ from a lexical unit to another. For instance, *monde* can be combined according to the pattern n+vb only in the lexical unit b).

Concrete referential expressions, which have not produced figurative and metaphorical meaning extensions, are the poorest type. The more precise the referent is, the poorer its range of contexts. Most technical and highly specialized referential expressions also have poor combinatorial ranges.

---

[112] Source : « Dictionnaire des combinaisons de mots. » Le Robert (Le Fur 2008: 647-648). The expressions can be translated as follows :
- un vaste monde, "the wide world"
- conquérir le monde, "to conquer the world"
- le monde animal, "the animal kingdom"
- un monde peuplé de … , "a world inhabited with…"
- entrer dans un monde, "enter a world"
- un monde fou, "lots of people"
- connaître du monde, "to know people"

This shows the importance of the evaluation of the number of contexts for a word, as well as the importance of word order. The natural variation of a word may be evaluated, so that variations going beyond this threshold be considered and studied as possible traces of change in use.

2.2.1.3. Geometric modeling

The geometrical metaphor the model is based on poses the question of the relationship between shapes and meaning. Indeed, all vector space models and Correspondence Factor Analysis models assume that it is intuitively and psychologically acceptable to work with geometrical and spatial representations of meaning. However, there is no major scientific work yet that shows the value of that relationship beyond assumption. Research in the field of data visualization is questioning this relationship and offering a plethora of possibilities (see for instance Rendgen 2012). Ongoing work on the conceptual anchorage of the relationship between form and meaning and the relationship of graphic design with science can be found in A. L Renon, (see Renon, to be published). In effect, computational semantics vector models and data visualization and information graphics methods can be combined to better exploit their common grounds. In collaboration with Anne Lyse Renon, then a student at the L2C2, we have tried to explore the relationship between computational semantics and data visualization, by applying a graphic design approach to studies of semantic change conducted with ACOM.

2.2.1.4. Thresholds

As evidenced above, scale, diffusion and time issues come into play and cannot be disentangled or treated separately. In addition, the reasons why some changes will show rocketing diffusion across time and scales do themselves rely on a series of factors. These factors involve linguistic, sociological, psychological and sometimes political dimensions. The complexity of the interaction of levels and factors makes it clear that no single measure of change in corpora can sum up the phenomena at stake. It is therefore necessary to work with several measures and tools for analysis and to allow some amount of flexibility in combining them on a case by case basis. Finding mathematical thresholds, even approximate ones, which can be said to have a linguistic value, may only be done after replicating these analyses on several corpora and averaging out results. For this reason, the resulting thresholds

that come out of analysis are only valid for the corpus at stake, and are to be taken as a small contribution to refining threshold settings for several corpora in the future.

2.2.2 Implementation

2.2.2.1. Choosing the corpus

The first corpus I worked on was the French newspapers *Le Monde* (1997-2007). Most of the data analysis in this doctoral thesis has been based on this corpus. A lot of French corpus studies are conducted on *Le Monde*, simply because it is easily available, and because the amount of work already produced on it allows for comparison of results. The reason why a lot of research is conducted on press corpora (beyond copyright and availability considerations) is that they provide a homogenous enough style to avoid biases due to idiosyncratic uses and offer a "window" on world events. The English and Spanish language corpora have been chosen to match the French one as closely as possible. Therefore, I tried to obtain press corpora in these languages with the same amount of stylistic homogeneity and with the closest cultural status in their countries of edition. Cultural status is by definition an approximation, but the political orientation of the newspapers, the number and types of readers they attract and the types of sections they offer provide a few tools for comparison. In English, the corresponding corpus used here is *The New York Times*. In Spanish, the access granted to *El País* and *Lavanguardia* allowed for simple data extraction, but this corpus does not match the French and English ones in terms of size and homogeneity.

2.2.2.2. Chunking

To adapt ACOM for diachronic studies, the corpus is first split into time slices. For commodity, the unit used here is the month as the press corpora that are used cover a decade or two. This unit is small enough to provide fine details on meaning evolution. Using a smaller unit would not enhance precision at this scale.

2.2.2.3. Stemming

The chunked corpus is then transformed into an ACOM database according to two different formats: stemmed and unstemmed. Stemming is the process of grouping the different inflected forms of a word under one unique heading: the lemma. In stemmed version, inflections (singular, plural, feminine and masculine for nouns and adjectives, and

conjugations for verbs) are gathered under a single heading with a POS (part of speech). This is done via Tree Tagger[113]. For instance, if the words *house* and *houses* appear, stemming groups them under the heading "NOUN house". The use of stemming is highly discussed in computational linguistics. On the one hand, it makes processing easier and clearer, but on the other hand subtle differences in meaning may be contained in uses of feminine or plural forms and are lost in the stemming process. In his study of the loss of meaning in the stemming process, Lemaire (2008) notes:

> "Par exemple, le contexte du mot *soleil* n'est pas le même que celui du mot *soleils*, notamment parce que les prédicats associés à ces deux formes ne sont pas systématiquement les mêmes. Ainsi, le mot *brille* apparaît dans un de nos corpus généraux 39 fois avec la forme *soleil* et jamais avec la forme *soleils*. De même, le mot *rayon* apparaît 311 fois avec *soleil* et seulement 2 fois avec *soleils*. A l'inverse, le mot *étoiles* y apparaît 68 fois avec *soleils* et 10 fois avec *soleil*. Ce sont donc les différences de contextes entre les diverses formes d'un lemme qui défavoriseraient les corpus lemmatisés, pour des algorithmes qui utilisent justement ce contexte. " [114]

In agreement with the idea that systematic stemming induces a loss of semantic information, corpora are explored under two forms: it is easier to work on general tendencies in stemmed version, but possible to access a finer level of detail in unstemmed version when needed. Moreover, stemming is a problem with neologisms and rare words, since Tree Tagger is trained beforehand and cannot find them. Therefore, all neologisms show the highest rate of tagging errors.

---

[113] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[114] "For instance the context of the word *sun* is not the same as t *suns'*, notably because the predicates associated to these two forms are not systematically the s ame. Thus, the word *shines* appears in one of our generic corpora 39 times with the form *sun* and never with the form *suns*. In the same manner, the word *ray* appears 311 times with *sun* and only twice with *suns*. Conversely, the word *stars* appears 68 times with *suns* and 10 times with *sun*. Therefore the context differences between the various forms of a lemma put stemmed corpora at a disadvantage for algorithms that are precisely using this context."

The ACOM format of the output for a word request in tagged version is:

> -part of speech/ word/ word frequency / co-frequency of the 1st co-occurrent / co-frequency of the 2nd co-occurrent… co-frequency of the n co-occurrent.

The co-occurrence output in diachrony is sorted hierarchically so to provide tables. To show the type of semantic differences that characterize inflected forms here are samples of the outputs for an ACOM query drawing on Lemaire's (2008) example. The following tables are excerpts of the hierarchically sorted raw co-occurrences for *soleil* ("sun") (Table 9) and *soleils* ("suns") (Table 10) respectively in unstemmed format and soleil (n.) in stemmed format (Table 11).

date	Nb	…	été	…	coucher	lumière	rayons	vers	Terre
199701	89	…	5	…	3	4	4	2	8
199702	117	…	7	…	5	3	3	3	5
199703	110	…	3	…	2	2	7	3	4
199704	130	…	6	…	2	2	6	1	3
…	…	…	…	…	…	…	…	…	…
200709	40	…	4	…	2	0	1	1	1
200710	54	…	1	…	1	2	3	1	0
200711	53	…	1	…	2	3	5	1	5
200712	40	…	2	…	1	1	1	1	1

Table 9 Excerpt of the co-occurrences for *soleil* in unstemmed version, on the corpus *Le Monde* (1997-2007), retaining the first plain (content) words

date	Nb	…	étoiles	trois	…	deux	…	Faudel	…	planètes
199701	2	…	0	2	…	0	…	0	…	0
199702	4	…	0	2	…	0	…	0	…	0
199703	2	…	0	0	…	0	…	0	…	0
199704	3	…	2	0	…	0	…	0	…	1
…		…			…		…		…	
200704	1	…	0	0	…	0	…	0	…	0
200705	3	…	0	1	…	1	…	0	…	0
200709	1	…	0	0	…	0	…	0	…	0
200710	2	…	0	0	…	0	…	0	…	0

Table 10 Excerpt of the co-occurrences for *soleils* in unstemmed version, on the corpus *Le Monde* (1997-2007), retaining the first plain (content) words

There are very few occurrences of the plural compared to the singular. The tables confirm Lemaire's statements, since *étoile* does not appear in the major co-occurrent words for the request in the singular, but does so for the request in the plural and the request in stemmed

version. The two forms are used differently as *Faudel* testifies, among other artist names, who can use the plural form for poetical effect. *Soleils* is, as Lemaire states, secondly associated to astronomy issues while *soleil* is much more generic, including both the poetic and the astronomical aspects. Loosing this information about word use would drastically weaken semantic analysis. In stemmed version, both aspects are merged.

date	nb(NOM/NAM)	NOM jour	ADJ grand	NOM/NAM terre	NOM rayon	(…)	NOM/NAM étoile	NOM/NAM lune	NOM planète
199701	89	4	2	9	5	(…)	3	2	3
199702	121	6	8	7	5	(…)	1	1	2
199703	112	5	2	4	8	(…)	9	3	4
199704	132	7	6	4	6	(…)	6	0	3
(..)	(..)	(..)	(..)	(..)	(..)	(..)	(..)	(..)	(..)
200710	56	2	5	1	3	(…)	0	1	1
200711	53	3	2	5	7	(…)	2	1	5
200712	40	4	0	2	2	(…)	0	2	1

Table 11 Excerpt of the co-occurrences in stemmed version for the noun *soleil* in the corpus *Le Monde* (1997-2007), retaining the first plain (content) words. NOM stands for "noun" and NAM for "proper noun".

2.2.2.4. Databases

On the basis of the ACOM formatted corpus a database of frequency evolution is created for each word, giving its frequency per month as well as its normalized mean frequency. Normalized frequency is obtained by dividing the raw frequency of the target word in each time chunk by the total number of words in the chunk and multiplying it by the mean total number of words per time chunk. Using normalized frequencies rather than raw frequencies prevents the chunk's size from generating a bias in the results. As regards the corpus *Le Monde*, there tends to be lesser and lesser verbal content over the years, being replaced by more space for images. Using raw frequencies on it would result in decreasing slopes of frequencies for a lot of words that are in fact stable or increasing, and it would also result in drastic exaggerations for the slopes of words with decreasing frequency. With this frequency database, simple graphs showing word frequency evolution can be generated and case studies may be conducted as well as studies of overall corpus trends. Another database is then generated containing the evolution of the frequency of each pair of co-occurrent words.

2.2.2.5. Filtering

On the basis of these databases, words can be filtered and classified according to their statistical profile, at different levels of granularity. The variation of frequency or normalized frequency may be extracted to detect a change in frequency of use, and classify the slopes in a hierarchical order to denote the tendencies of use over the whole corpus.

The slope here is the regression coefficient of frequencies. Then the variation of co-occurrent words with a target word may be detected, and once detected this variation may show changes in the use of a single word and the reorganization of its semantic network. To do so the window that is used is the sentence. Then the coefficient of variation is calculated for each word; it is the ratio of the standard deviation to the mean. This measure shows the variability of the word and is also used efficiently by Holz et Teresniak (2010) to detect topic change in press corpus. The overall method for filtering is summed up in Figure 12.

2.2.2.6. Hypotheses

Figure 12 shows that two main paths of analysis come out in the selection phase. These two paths are the basis for the following hypotheses: In the first one, frequency variation as well as co-occurrence frequency variation is detected, showing a change in status of a word in its use and semantic networks. In this case the hypothesis is that the filtering detects short diachronic semantic change in the corpus, which is most likely linked with an event relayed by the press. This first possible path is referred to as the "topic-related connotational dift" in the diagram. It happens within very short time spans (years and sometimes months). In the second path, tagged "connotational drift" in Figure 12[115], only co-occurrence frequency variation is detected, leading to the hypothesis that the variation affecting the semantic networks is not related to the corpus and the events it describes but rather to changes that may have started to unfold before the time of corpus. The second hypothesis is thus that these changes belong to long diachrony processes. It seems to be related to larger sociolinguistic and linguistic phenomena.

[115] Presented on a poster at the International Conference on Computational Semantics (IWCS) 2011, in Oxford. Graphic Design by Anne-Lyse Renon.

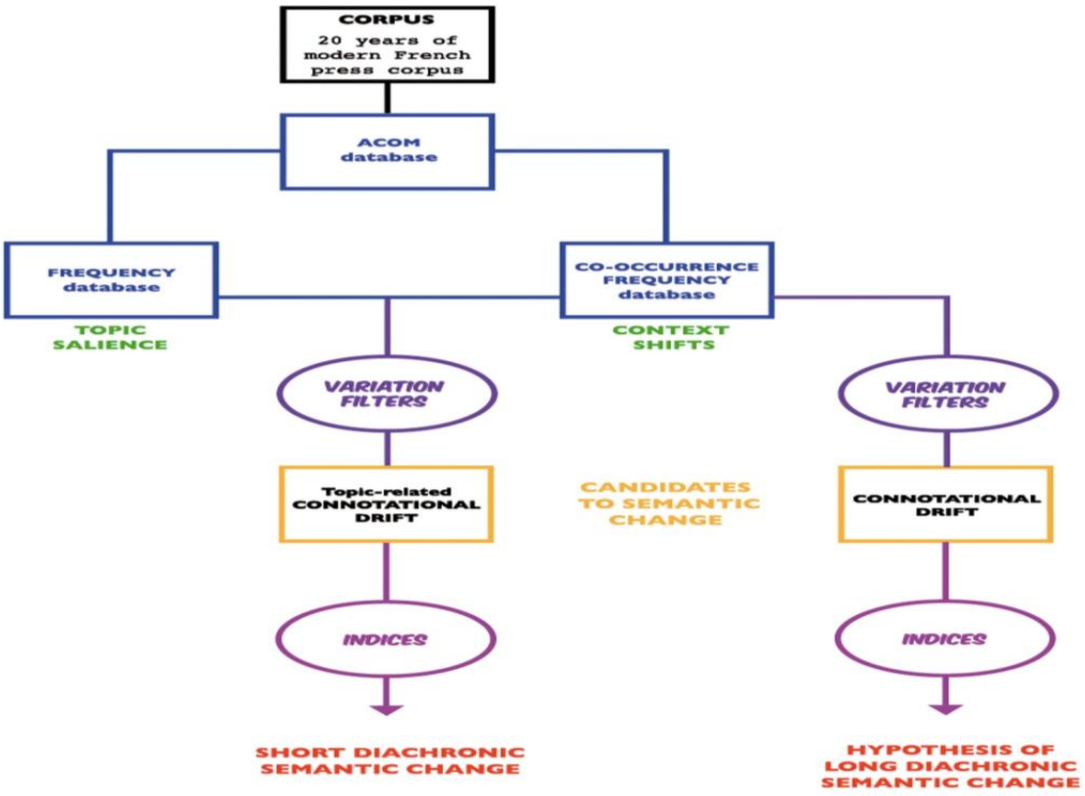


Figure 12 Schematic representation of the methodology

To further investigate the candidates that have been selected through this process, analysis tools allow for detailed exploration of words on a case by case level. These hypotheses will be tested on case studies in the third part of this work. The third Part of this work applies these tools to case studies.

PART III Corpus Trends and Case Studies

Two types of analyses were conducted: filtering on the whole corpora to delineate general tendencies and detailed analysis for case studies involving a single word or an element of composition. Theoretically, detailed studies shed light on mechanisms of change and, as such, are a starting point to the extraction, understanding and modeling of these mechanisms at large. Therefore, going into detailed analysis at word level is a necessary step of investigation in the perspective of mass extraction and detection. Case studies are not only about the entertaining story of a word. They are also and most importantly about the mechanisms at stake in that story. To model patterns of semantic change, the first step is to make these patterns come to light. The following studies are limited to this first step and are far from being exhaustive. They include some modeling attempts at the experimental level, but do not in any case offer a complete model.

Indeed, being exhaustive would mean that we know how many patterns, or "types", need to be dealt with, and this is not the case yet. Typologies define and classify them and there would be as many ways as there are existing typologies to extract mechanisms of semantic change. The extraction would depend on the number and nature of the patterns which are attested by them. However, as shown in Part I. there is no clear consensus in the scientific and linguistic communities about how many types there are and how they are organized. Systematic extraction by category is constrained by the choice of a set of categories and is therefore constrained by a theoretical view. The chosen perspective here is purposefully more naïve in that I try to observe, analyse, and ultimately model semantic change mechanisms. These mechanisms may pertain to the making of several "types" which are unanimously attested, even if sometimes viewed differently, but they may also characterize "types of profiles" which emerged via statistical filtering. The aim is rather to dissect their nature, structure and dynamics, without trying to evaluate a type's relevance or definition. In this sense, this work is not a typological one, but rather an experimental one.

Within that scope, the most accessible indicator of change is variation in use and context. The latter can only be measured in terms of a norm or baseline, which in this case is the corpus (taken as a semi-closed linguistic system for the purpose of simulation). Language is subjected to "natural" variation in synchrony, which is not semantic change but rather a state of flexibility, as well as variation which produces semantic change in diachrony. Separating

the two phenomena is not an easy task since both might be rooted in similar mechanics operating under and above undefined thresholds of variation, diffusion and time.

In the first chapter of this section, tools for filtering and indices are presented. Results are detailed in the second chapter.

Chapter III.1: Tools for filtering and indices

3.1.1 Corpora

The most explored corpus is an archive of the French newspapers *Le Monde* covering the years 1997 to 2007.[116] Work has been conducted on two formats: stemmed and unstemmed.

In the unstemmed version the corpus has 221 580 827 tokens and 795 628 types.

In the stemmed version the corpus has 219 378 433 tokens and 87 304 types. *Le Monde* was used as the major test corpus, since its access was easier[117] and many other corpus linguists work with it, which makes results comparable in the future. Ideally, indices tested in the corpus *Le Monde* could be tested in other corpora.

The English corpus is the American newspapers *The New York Times* (one of the closest equivalent to *Le Monde*) covering the years 1987-2007. In stemmed version it has 973 044 833 tokens and 223 048 types. In unstemmed version it has 1 015 620 131 tokens and 1 952 913 types.

Le Monde and *The New York Times* were tagged with Tree Tagger and exploited in ACOM format directly built from raw text[118].

---

[116] My early works were based on *Le Monde* (1997-2001) and *Le Monde* (1997-2003), as the corpus was acquired step by step.

[117] The L2C2 had access to the 1997-2001 corpus when I started this work. Access to "The New York Times" was acquired later.

[118] Both corpora came with pre-applied linguistic treatments. Since these treatments are different and bias corpus result comparisons, both were discarded, and all work was done from raw text and ACOM format.

Two other corpora were browsed uniquely for trends for comparison purposes, and their access was granted by two institutions. They were explored in untagged version. The first is a French general long diachrony corpus (with a wide time period covering 1180 to 2009), including not only press but also literary, philosophical and scientific sources, called *Frantext* [119]. The second is a Spanish and Catalan press corpus from which I used only a subpart (the most general one) with irregular coverage from 2000 to 2007 which has a total of 16 154 125 words.[120]

3.1.2 Indices

As summed up by Renouf (2007), there are three major elements to assess a word's life in lexical diachrony :

"With the diachronic perspective on lexis, the degrees of frequency, productivity and creativity indicate how active and important a word is in the language at a given point in time, and they provide the means whereby its "life-cycle may be charted across the years."

To filter out candidates to semantic change, I relied on a series of indices that use and assess frequency patterns, productivity and creativity. The indices are of three types: mathematical and computational indices, as well as linguistic and extra-linguistic ones.

3.1.2.1. Mathematical and computational indices:

FREQUENCY VARIATION

- The variation of raw and normalized frequencies for a head word. Normalized frequencies are obtained by dividing the number of frequencies by the total number of words in the time chunk and multiplying the result by the mean number of words per time chunk.

  Frequency patterns show the stability of use for a target word. When the variation of frequency is unusual, meaning may be affected through events that mobilize the word or through linguistic trends.

[119] http://www.Frantext.fr. Access was granted to the team by the ATILF.

[120] Access was granted by the IULA.

- The regression coefficients on frequencies (the slope of the regression line).
- The coefficient of variation of individual word frequencies. The coefficient of variation is the ratio of the standard deviation to the mean.

CO-OCCURRENCE NETWORK STRUCTURE

- Frequency patterns of the co-occurrent words for a given word, independently of that word.
- Frequency patterns of the co-occurrent words for a given word, relatively to that word.
- Normalized co-occurrence frequencies. Raw frequencies of co-occurrence divided by the frequency of the target word and multiplied by the mean frequency of the target word.
- Hierarchically sorted networks of co-occurrence for a target word. The mean for each co-occurrent word is calculated and co-occurrent words are classified hierarchically in a table, with a mean per month for each word.
- Ranks, showing the evolution of the hierarchical importance of co-occurrent words over several time periods. Ranks show the internal structure of the context network.
- The density and cohesion index, based on the variation of the ratio between the number of cliques and the number of associated words for a head word over time. This measure was created in terms of the SA paradigm. The ACOM model generates a series of associated terms for a target word and a series of cliques in which these terms are organized in sets. A high number of new cliques shows the necessity to build continuity with the existing set of cliques, either because the existing contexts are enriched or because new contexts appear and have to integrate the existing system in a coherent way.

- The time frames may be re-chunked according to frequency variations, and indices re-applied to the re-chunked corpus.

3.1.2.2. Linguistic indices

- Part of speech.
- Punctuation and spelling (a new or unusual word often appears between quotes, with or without a hyphen, or followed by an explanation in parenthesis)

- Morphological productivity. When the word is prefixed or suffixed, or giving birth to a compound. Morphological productivity is an index of the lexical "life" of the word: if a word is morphologically productive, it is in high use. The created neologisms may impact retroactively the source terms. The trend effects produced by high frequency use may impact it as well. The production of derivational and compositional neologisms is also related to real-world realities, by naming unnamed, but existing referents.
- Synonymic competition. When two or several words are close to pure synonymy, one of them is expected to "win the competition" while this process implies a re-distribution of semantic content across the different candidates.

3.1.2.3. Extra-linguistic indices

- The information about the topic, author, section and genre contained in the corpus, (a specialized section uses specific terminology, as some authors do…)
- Supplementary real-world information is gathered manually when needed, gathered over the Internet and in various resources (books, articles, etc.). Since the press corpus is tightly connected with the events it depicts, the analytical dimension of events is dealt with manually in post-treatment to preserve linguistic objectivity.

Keeping function words

While applying these filters, all function words have been kept. Numerous studies in NLP use stop-lists to remove them in order to optimize calculation. However, function words can show a lot about word use and meaning. I have chosen to keep all this information for the sake of precision and detailed analysis. In effect, one of the debates evidenced in NLP is the choice between optimal and efficient programs versus heavier calculations that preserve data. Optimization issues are beyond the scope of this work and at this stage, the preservation of details is favored. In the perspective of refining tools later, compromises can be made between the computationally efficient approach and the detailed linguistic one.

3.1.2.4. Similar indices in the literature

Other authors are working with similar indices combined in different ways. Dury and Drouin (2009) offer a strikingly similar list of indices to detect neology but also "necrology" or how words die out:

"- **Les marqueurs linguistiques**, c'est-à-dire l'utilisation d'expressions et de formulations caractéristiques par les auteurs pour parler d'un terme désuet ou disparu (par exemple : « *formely called* », « *previously known as* », etc.),

- **La ponctuation et la typographie**, c'est-à-dire l'utilisation des parenthèses, des guillemets et des italiques par les auteurs pour parler d'un terme désuet ou disparu,

- **La distribution**, c'est-à-dire les changements pouvant se produire dans les cooccurrents d'un terme au fi l du temps, indiquant alors peut-être un changement de sens de ce terme,

- **La variation synonymique**, c'est-à-dire le foisonnement de termes concurrents qui peut se produire lorsqu'un terme disparaît d'un lexique, tout comme il se produit lorsqu'un néologisme apparait dans un lexique,

- **La grammaire**, c'est-à-dire le changement de nombre ou de catégorie grammaticale d'un terme au fil du temps (par exemple un terme utilisé comme substantif dans la partie « ancienne » du corpus, ensuite utilisé comme adjectif, et dont la forme substantivée disparaît au fil du temps).

- **La morphologie**, c'est-à-dire la disparition ou la modification d'un ou de plusieurs affixes dans une forme lexicale,

- **La fréquence**, c'est-à-dire l'observation statistique des changements de fréquences d'apparition d'un terme dans l'ensemble des sous-corpus."[121]

This list confirms the validity of my assumptions, since I had no knowledge of Dury and Drouin's work when I complied lists of possible indices to detect semantic change.

121 The linguistic markers, that is to say the use of characteristic expressions and formulations by the authors to speak of an obsolete term or a term no longer in use (eg. "formerly called," "previously known as", etc..)
- Punctuation and typography, that is to say the use of parentheses, quotation marks and italics by the authors to speak of an obsolete term or a term no longer in use.
- The distribution, that is to say, the changes that can occur in co-occurrent words of a term, perhaps indicating a change in the meaning of this term. -Synonymic variation, that is to say the proliferation of competing terms that can occur when a term disappears from a lexicon, in the same way that it occurs when a neologism appears in a lexicon,
- Grammar, that is to say the change in number or grammatical category of a term in time (eg a term used as a noun in the "old" part of the corpus, used as an adjective afterwards, and whose substantive form disappears in time).
- Morphology, that is to say, the loss or modification of one or more affixes in a lexical form,
- Frequency, that is to say, the statistical observation of changes in frequency of occurrence of a term in all sub-corpora.

Chapter III.2 :Results

In the first part, overall corpora trends are outlined, along with examples. In the second part, three case studies are presented: the first study deals with formal neology in the case of the French word *malbouffe*, as well as "paternity" issues in the creation of a new meaning. It is then extended to morphological productivity issues in *mal-* . Drawing on this first approach, the second study focuses on morphological productivity, with the examples of *crypto-* and *cyber-* as well as a detailed study on *bio-* in the third section.The fourth study focuses on detailed analysis of connotational drift and involves a graphical dimension, anchored in the SA model and in collaboration with Anne-Lyse Renon (former trainee at the L2C2 and PhD student) specialized in graphic design. The evolution of the word *mondialisation* is studied and rendered graphically by the model and made more readable and intuitive by the design work. This collaboration also benefited from the programming skills of Charlotte Franco (who is also a former trainee at the L2C2). This gave birth to an attempt at dynamic visualization, with the question in mind of what would be a suitable visualization if it were to be applied to a whole model, and not just a candidate word. The visualization was shown at the colloquium "Néologie sémantique et analyse de corpus"[122] and the study has been published in French in (Boussidan et al. 2012).

The examples are grounded in sociolinguistic variation within the journalistic genre. There is a debate among linguists about the value of sociolinguistic change, based on the fact that what changes is the way the sign is received by a certain linguistic community, rather than a brutal and complete change (necessary to define neology for certain authors, for instance Sablayrolles, in the same volume). However, if this debate is justified, sociolinguistic change is clearly a solid index for the detection of semantic change. This debate is well summed up by Gérard and Kabatek (2012: 18)

"… la dimension du changement qui intéresse la variation sociolinguistique n'est qualifiable de « sémantique » qu'à la marge, car il concerne moins le contenu du signe, c'est-à-dire le signifié, que le *signe lui-même*, dans sa totalité, en tant qu'unité appariant les deux plans de l'expression et du contenu. […] Ce dernier changement n'intéresse pas la néologie sémantique, ni même sans doute les évaluations qui l'accompagnent. En revanche, les

[122] Semantic neology and corpus analysis

évaluations peuvent être conçues comme des indices du changement sémantique (Rastier 2000)."[123]

3.2.1. Corpus Trends

Frequency variation is an indicator of a word's use in corpus. It is not an indicator of semantic change *per se*; however, words that undergo semantic change often combine an unusual frequency profile with a variation in their network of co-occurrent words. Therefore, it is useful to use measures of frequency variations even if they do not suffice on their own. Apart from raw and normalized frequencies for a given word, frequency measures can be applied to all words to browse a corpus and extract major tendencies and variations. To do so, regression coefficients and coefficients of variation are calculated.

3.2.1.1. Regression coefficients: tendencies in the corpus

Generating the regression coefficients on normalized frequencies for each word, and classifying results in hierarchical order, gives us a picture of the major tendencies of vocabulary change in the corpus. These changes characterize the corpus as a global context of occurrence for the case studies that follow. The regression coefficient allows for an assessment of the general tendency of frequencies. If the coefficient is positive, it indicates an increase in use, if it is negative, a decrease. However, when words have very low frequencies, these values are too low for the regression coefficient to capture a tendency at all. They are therefore removed from a selection if they are equal or under a mean frequency of 1 per time chunk. The strongest absolute values for each corpus, in stemmed and unstemmed version, for this selection and for the entirety of the corpora, are the following:

-in *Le Monde*, in unstemmed version: 30.89 to 2.49e-07 in the selection

30.89 to 1,75e-09 for all words

-in *Le Monde*, in stemmed version: 39.60 to 7, 238 e-08 in the selection

[123]"The dimension of change which is interesting to sociolinguistic variation is only definable as "semantic" at the margin, it is less related to the content of the sign, that is to say the signified, than to *the sign itself*, in its wholeness, as a unit matching the two planes of expression and content. […] This latter change is not interesting to semantic neology, nor is the evaluation that goes with it. Nevertheless, the evaluations can be conceived as indices of semantic change"

39.60 to 1,128 e-08 for all words

-in the *NYT*, in unstemmed version: 109.14 to 1.057e-07 in the selection

109.14 to 5.595e-10 for all words

-in the *NYT*, in stemmed version: 93.57 to 2.048e-07 in the selection

93.57 to 1.493e-09 for all words

-in *Frantext*, in unstemmed version 199.37 to 6.766e-07 in the selection

199.37 to 8.586e-08 for all words

-in the Spanish corpus, in unstemmed version 9.076e+00 to 6.901e-07 in the selection

9.076e+00 to 2.842e-08 for all words

Words with extreme patterns are useful to characterize the ranges and profiles in which frequency change is meaningful. For instance, Figures 13 and 14 show normalized frequencies and subsequent linear regressions for the nouns *euro* and *franc* in *Le Monde.* The first shows drastic increase while the second shows drastic decrease on the whole. The word *euro* was highly discussed before being introduced, and therefore its frequencies are high before the currency change.

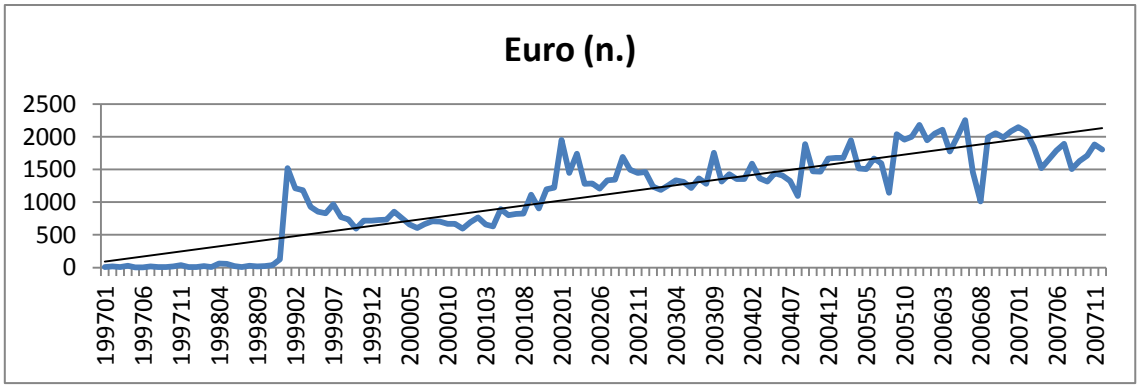


Figure 13 Normalized frequencies for the noun *euro*[124], with linear regression, in the corpus *Le Monde* (1997-2007) in stemmed version.

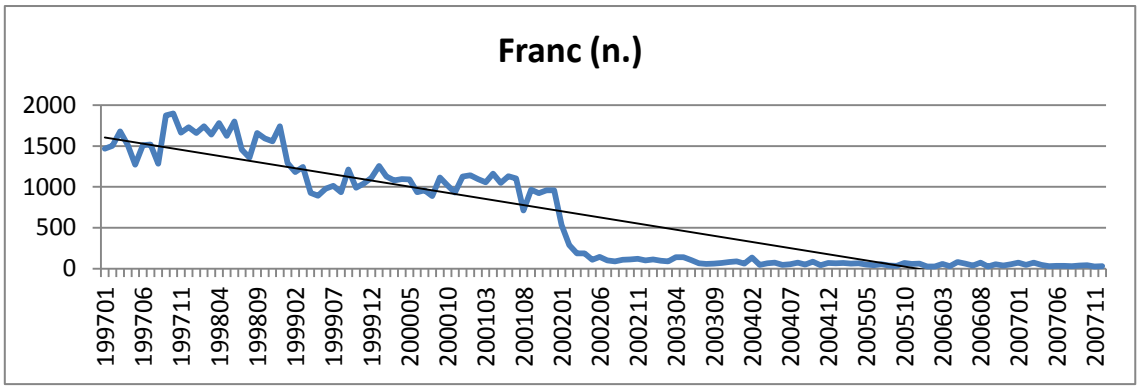


Figure 14 Normalized frequencies for the noun *franc,* with linear regression, in the corpus *Le Monde* (1997-2007) in stemmed version

These results are extreme (appearance/quasi-disappearance), and show the general tendencies that are to be looked for in other words with more subtle frequency behaviours to detect changes in use.

Table 12, an excerpt of the strongest coefficients in the corpus "*Le Monde*, 1997-2007" shows major social changes as well as linguistic ones. It is subdivided according to part-of speech sets: Table 12a contains nouns and adjectives, Table 12b contains prepositions, pronouns, adverbs and conjunctions, and Table 12c contains proper nouns.

[124] Raw frequencies are divided by the total number of stemmed words per month and then multiplied by the mean number of stemmed words.

Nouns and adjectives			
positive		negative	
NOM euro	15,624	NOM franc	-15,037
NOM monsieur	7,990	ADJ national	-3,052
NOM président	3,797	NOM heure	-2,542
NOM candidat	3,347	NOM temps	-2,219
NOM pays	3,168	NOM emploi	-2,122
NOM madame	2,977	NOM livre	-2,048
NOM ministre	2,916	NOM travail	-2,010
ADJ présidentiel	2,495	NOM banque	-1,997
NOM million	2,489	NOM monnaie	-1,824
NOM novembre	2,409	ADJ général	-1,742

Table 12a: Excerpt of the strongest regression coefficients in *Le Monde* (1997 – 2007) in stemmed version for nouns and adjectives. Parts of speech tags: ADJ= adjective, NOM = common noun.

Prepositions, pronouns, adverbs and conjunctions			
positive		negative	
PRP en	13,170	KON et	-8,648
PRO:PER il	9,515	PRP de	-7,080
DET:POS son	8,347	PRO:PER on	-5,051
PRP pour	6,896	PRP:det du	-2,982
PRO:PER je	6,756	PRO:IND tout	-2,288
ADV pas	6,273	PRO:PER se	-1,916
PRO:PER elle	4,418	PRP sans	-1,542
PRO:PER nous	4,274	PRO:IND quelque	-1,449
ADV ne	3,815	ADV encore	-1,328
PRP selon	3,781	ADV donc	-1,296
PRP contre	2,587		

Table 12b: Excerpt of the strongest regression coefficients in *Le Monde* (1997 – 2007) in stemmed version for prepositions, pronouns, adverbs and conjunctions. Part of speech tags:PRP= preposition, KON= conjunction, DET: POS= possessive noun, PRO:PER= personal pronoun, ADV= adverb, PRO:IND= indefinite pronoun.

positive		negative	
sarkozy	7,840	jospin	-4,030
ump	5,133	jean	-2,608
nicolas	4,496	lionel	-2,227
irak	3,889	kosovo	-1,793
bush	3,034	claude	-1,395
royal	2,830	otan	-1,194
villepin	2,182	serbes	-1,068
iran	1,754	philippe	-0,999
etats-unis	1,734	aubry	-0,995
chine	1,494	maurice	-0,952

Table 12c: Excerpt of the strongest regression coefficients in *Le Monde* (1997 – 2007) in unstemmed version for proper nouns. The unstemmed version is used to avoid TreeTagger's occasional tagging mistakes.

Table 12 Excerpt of the strongest regression coefficients in *Le Monde* (1997 – 2007) in hierarchical order according to part-of speech sets, in a), b) and c).

In Table 12a, containing nouns and adjectives, the French money the *Franc* is replaced in this period by the *euro*. At a more subtle level, the adjective *national* decreases in use while *pays* ("country") increases. While these two expressions refer to similar concepts, *national* is coloured by the noun derived from it *nationalisme* ("nationalism") while *pays* is more neutral. In positive values, the nouns and adjectives *candidat*, ("candidate"), *président*, ("president"), présidentiel, ("presidential") and *ministre (*"minsister, secretary of,") show the growing importance of domestic policy as well as national elections. In negative values, vocabulary related to work, money and time prevail: *emploi* ("job", "work") *travail* ("job", "work"), *banque* ("bank"), *monnaie* ("currency"), *temps* ("time") and *heure* ("hour"). The case of *heure* ("hour") seems more mysterious at first glance, however normalized frequencies show three peaks of use in the corpus in February 1999 and January and March 2001, corresponding to vigorous debates about the 35 hour-work week in France. The debate then wears off.

Table 12b mostly shows stylistic trends, with a growing use of negation for instance (*ne…pas*) while the increase in use of the prepositions *pour* ("for") and *contre* ("against") may be interpreted as a growing tendency to be "for or against" policies. The preposition *selon* ("according to") may be interpreted as an increase in use of reported speech and referencing.

In Table 12c, within proper nouns, the major front-scene media topics related to wars, political parties and politicians emerge. Actors of domestic policy appear in positive values (*Nicolas Sarkozy*, *UMP*[125], *Royal*, *Villepin*) along with international wars and politics (*Irak*, *Bush*, *Iran*, *Etats-Unis*, *Chine*[126]). The same is true for negative values, in which we find names of politicians who left the media front scene, such as the prominent left-wing politician Lionel *Jospin*[127], along with references to ending international conflicts (such as *Kosovo*), and the notable presence of *Otan* (NATO).

Such a simple extraction of frequency trends shows how complex sorting frequency results can be. Indeed, for a fully automatic treatment to be efficient, the ideal programme must possess encyclopedic knowledge about the money in Europe, real world events, political personalities and parties, and general knowledge about the relationship between them. This analysis also applies to the subsequent tables and justifies a semi-automatic approach.

With a much longer time span, we can observe more profound changes. In Table 13 the same method is used on a corpus including French literary text from the XIXth century to today[128].

[125] UMP : « Union pour un Mouvement populaire », right-wing French party to which former president Nicolas Sarkozy belonged.

[126] Irak, Bush, Iran, United States, China

[127] This politician left politics for about ten years and consequently the media as well, to the point that recent newspapers commenting on his renewed political responsibilities in 2012 talk about a *come-back*, see « Le *come-back* de Lionel Jospin n'est pas une surprise »(“Lionel Jospin’s come-back is not a surprise”),http://www.lexpress.fr/actualite/politique/cinq-choses-a-savoir-sur-la-commission-jospin_1138769.html, accessed on 25/07/2012.

[128] Thanks to the ATILF for granting access to this corpus.

pas	106,500	vous	-139,295
tu	32,540	point	-37,941
temps	15,190	homme	-30,238
visage	11,160	cœur	-24,946
maman	11,037	roi	-18,863
guerre	10,013	dieu	-16,449
petit	9,715	nature	-16,437
train	8,986	esprit	-16,279
années	8,582	âme	-15,000
fin	8,016	madame	-14,589
début	7,697	duc	-13,506
nuit	7,697	hommes	-13,266
sens	7,125	idées	-13,087
rue	7,112	grand	-12,174
travail	6,726	monsieur	-12,072
problème	6,684	empereur	-11,711
type	6,644	lettre	-11,691
jeu	6,511	peuple	-10,864
place	6,278	paris	-10,700
question	6,111	comte	-10,424

Table 13 Excerpt of the strongest regression coefficients in the corpus *Frantext*, a compilation of French literary texts from the IXth Century to today.

Vous shows a clear decrease; the third person pronoun in the plural used for politeness in French becomes more and more obsolete, as the second person singular *tu* takes over. The negation mark *point* follows the same fate, as *pas*, its less formal equivalent, replaces it over time. Nonetheless, both *vous* and *point* survive in spoken French, being restricted to more literary and formal contexts. These words show a general shift in register, towards a more familiar written style and a change in status of words, being less frequently used and in more specific contexts. They have been subject to social and stylistic changes, and testify of a more general register shift from formal to informal in French over the past centuries. This phenomenon is reinforced by the negative values for *monsieur* (“mr, Sir”) and *madame* (“Ms, Madam”).

The strong positive value for *travail* (“work” also used in the sense “job” and “piece of work”) is paralleled by the strong negative value for *ouvrage* (“work” still used in the sense “piece of work”) of -6.392 (in the 0.0379 % highest values). This parallel shows how a word

"wins" the synonymic competition over time, adjusting to the social changes related to it. In spoken and written French, *Ouvrage* gains a connotation of artistic and intellectual work, as its register becomes more literate, while *travail* takes on a more general meaning. The value of *train* reflects the introduction of a -once- new referent in everyday life. As for the negative values of *cœur* ("heart ") , *roi* ("king"), *dieu* ("god"), *nature* (idem.) *esprit* ("spirit" or "mind"), *âme* ("soul"), *duc* ("duke") *empereur* ("emperor") and *comte* ("earl"), they clearly show that these subjects and referents have fallen in disuse in more recent texts, while they were central in older ones. On the contrary, the positive coefficients for *rue* ("street") and *guerre* ("war") show the emerging importance of these concerns

Applying the same method to the New York Times corpus (1987-2007) provides the following output:

Nouns and adjectives			
positive		negative	
NN street	6,095	NN share	-21,210
NN pm	5,714	JJ net	-21,170
NN notice	5,479	NN company	-18,653
NNS death	5,258	NNS earning	-8,276
NP iraq	4,955	NN today	-9,968
NN family	4,623	NN percent	-9,793
NN web	4,473	NNS sale	-9,607
NN internet	4,195	JJ soviet	-9,418
NN site	4,013	CD million	-8,537
NN center	3,494	NN loss	-6,673
NN world	3,481	NNS share	-6,348
JJ online	3,203	NN president	-5,359
NN show	3,042	NN year	-5,329
NN art	2,975	NN ms	5,435
JJ beloved	2,940	NN corporation	-4,660
NN security	2,754	NN government	-4,447
NN wife	2,633	NN revenue	-5,146
NN way	2,530	JJ old	-4,014
NN sunday	2,452	NN quarter	-3,922
NNS attack	2,415	NN stock	-3,788

Table 14a Excerpt of the strongest regression coefficients calculated in the corpus *The New York Times* (1987-2007) for nouns and adjectives (and one number dealt with as a noun). Parts of speech : CD= cardinal number, NN = singular noun, NNS = plural noun, NP= proper noun, JJ= adjective

Verbs			
positive		negative	
VBZ be	12,257	VBZ report	-15,942
VBD say	7,829	VBZ earn	-15,388
VBN pay	5,206	VB be	-6,921
VBD do	3,845	VBN be	-5,156
VBP do	3,111	VBN unite	-3,892

Table 14b Excerpt of the strongest regression coefficients calculated in the corpus *The New York Times* (1987-2007) for for verbs. VBZ= verb, present tense, third person singular , VBD = verb, past tense, VBN=verb, past participle, VBP= verb present tense other than third person singular, VB= verb, base form/

Table 14 Excerpt of the strongest regression coefficients calculated in the corpus *The New York Times* (1987-2007) in stemmed version, in hierarchical order according to part-of speech sets, in a) and b).

When looking at the New York Times in a twenty-year time span (the double of *Le Monde*), the majority of decreasing coefficients are related to the semantic fields of corporate life and money-making (*share, net, company, report, earn, percent, sale, million, earning, loss, revenue, corporation, stock*), while increasing coefficients are related to the semantic fields of the internet (*web, Internet, site, online*) family and private life (*family, beloved, wife, Sunday)* as well as war and security issues (*Iraq, security, attack*), and the notable presence of *death*. As *earn* and *earning* show a decrease in use, *pay* shows an increase, showing the transition from the centrality of the notion of earning money to the centrality of paying. This type of transition may reflect the effects of the economic crisis in the choice of vocabulary. The adjective *Soviet* clearly goes out of use, reflecting a historical transition, as the Cold War's effects progressively fade away from American political life and as *Iraq* comes to the front scene, as in the French corpus. *Security* becomes a major concern (this is also evidenced in the French corpus as *sécurité* has a strong positive coefficient of 1,858 (comprised in the 0,085 % highest values). The presence of *art* and *show* in the positive values shows that there is a stronger focus on art events, and is more characteristic of the corpus than of linguistic trends.

While conducting a similar extraction on a shorter and smaller Spanish and Catalan corpus (2000-2007), comparable phenomena emerge:

positive		negative	
euros	1,997	pesetas	-2,220
internet	0,795	estadística	-1,518
zapatero	0,676	imprimir	-1,393
personas	0,647	interesa	-1,375
catalunya	0,534	enviar	-1,367
web	0,522	impresa	-1,172
barcelona	0,514	noticia	-1,111
google	0,473	edicíon	-1,057
compañia	0,466	utilidades	-1,028
red	0,453	paises	-0,700
.com	0,428	contra	-0,694
usuarios	0,413	internacional	-0,678
explica	0,398	estados	-0,636
catalan	0,353	aznar	-0,581
proyecto	0,326	país	-0,573

Table 15 Excerpt of the strongest regression coefficients calculated in the unstemmed Spanish corpus, including Catalan and Spanish text and covering 2000 to 2007[129], in hierarchical order.

Pesetas and *euros* follow the same trend as *Franc* and *euro* in the French corpus. Positive values include Internet related vocabulary (*internet, web, Google, .com, red,* "network", *usuarios,*"users") as in the American corpus. Vocabulary related to print edition processes decreases (*impresa*, "printed", *imprimir*, "print", *edición*, "edition"). As in both American and French press, front line politicians change (*Zapatero, Aznar*). Vocabulary connected to the perception of the country with negative values (*país, países, internacional, estados:* "country, countries, international, states") echoes the French results in which the word *pays* ("country") has a positive value but *national* (idem) has a negative one. As for *compañia* ("company"), it has a strong positive value while *company* had a negative one in the *NYT*. The references to Catalonia in the positive values may be due to the fact that the corpus is partly of Catalan origin (*Lavanguardia*).

[129] This corpus is taken from the IULA corpus. I would like to thank the IULA team in Barcelona for granting access to it.

Regression coefficients are therefore useful to extract the roughest frequency tendencies in a corpus. However, they do not show variation within these tendencies. To capture variation, we calculate the coefficients of variation for all words in each corpus.

3.2.1.2. Coefficients of variation and their distribution

The coefficient of variation is the ratio of the standard deviation to the mean for a given word frequency. It assesses variation in use and has no semantic value intrinsically. However, it filters out words that have an unusual frequency pattern. As stated earlier, unusual frequency variations are to be taken as a starting point for semantic analysis, and not as a result in itself.

Similarly to the selection operated for regression coefficients, very low frequency words are removed from the selection. Words with this profile have too few occurrences for the coefficient of variation to be significant. Coefficients of variation, like regression coefficients, are useful to show coarse grained tendencies in substantial amounts of data. However, the majority of types in the corpus have frequencies which are too low to be captured by this measure. Therefore, to make the data readable in terms of distribution and avoid working with a majority of noisy data, all words which have a frequency equal and inferior to a mean of 1 per time chunk are removed. Among the removed words, unusual spellings, idiosyncratic inventions and compositions are to be found among typographic errors. In rare cases, these items may be the first of their kind, either as formal neologies or as indicative of the unfolding of a semantic process. Therefore, they cannot be called "errors" and are rather "unusual uses" or idiosyncratic ones. Indeed, among them, many words are simply used once in a specialized article or a few times by the same author who likes an expression. It is therefore interesting to generate the figures for all words in the corpus to assess creativity, but the result includes too much noise to provide a readable overall distribution. These low-frequency words are studied separately subsequently, under the perspective of morphological productivity and creativity. As regards the coefficients of variation, the maximum values are the same before and after removing these words. However, the minimum values drop to zero when all words are considered. The ranges of coefficients of variation are the following:

- in *Le Monde*, in unstemmed version: 6,603 to 0,01524

- in *Le Monde*, in stemmed version: 5,651 to 0,01944

- in the *NYT*, in unstemmed version: 5,718 to 0,01307

- in the *NYT*, in stemmed version: 5,716 to 0,01093

- in the Spanish corpus, unstemmed version 6.121 to 0,03222

In the version of *Frantext* used for regression coefficients, the coefficients of variation cannot be relied on since the corpus is heterogeneous, and the coefficients capture differences in use related to the different styles and registers of excerpts that compose the corpus. Therefore there are no figures for *Frantext*.

Figure 15 shows the distribution of the coefficients of variation per head word for the corpus *Le Monde* in unstemmed format, after clearing out numbers and errors generated by special characters, and very low frequency words.

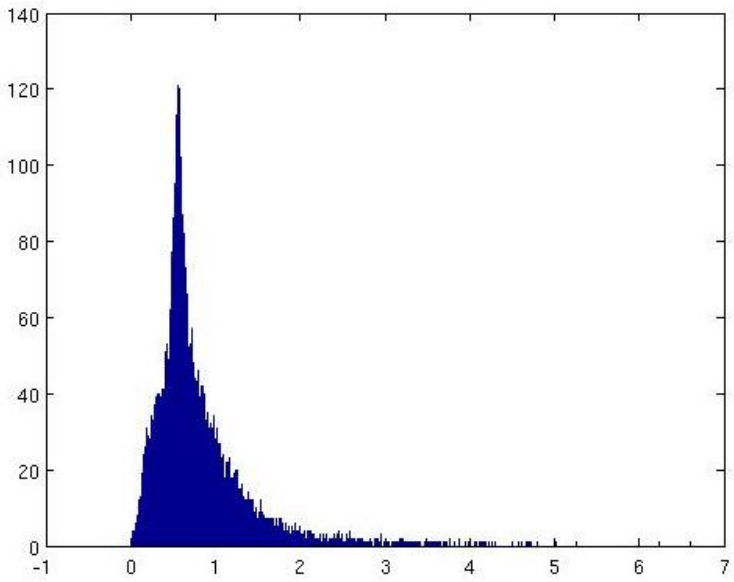


Figure 15 Histogram of selected coefficients of variation (very low frequency words have been removed) per head word in the corpus *Le Monde* (1997-2007) in unstemmed version, x shows the coefficient of variation values, y the number of words which possess this value.

The distribution of coefficients of variation shows that a majority of words are stable. Most words between 0 and 1 are stable function words (determiners, conjunctions, pronouns, adverbs etc.) as well as nouns, adjectives and verbs with stable use. The peak is at 0.582.

The highest coefficients of variation are fewer and more dispersed. There are only 1066 values over 2 in unstemmed version and just 664 when low frequency words are removed. They correspond to words whose frequency behavior varied substantially. While the regression coefficient extracted the major themes and trends in the corpus, the coefficient of variation extracts items with unstable behaviour: themes, ideas, objects, people, places and things that may have had a sudden importance and/or suddenly lost all importance thereafter in the newspaper. Therefore, measuring the coefficients of variation in the corpus is useful to extract strong unstable themes. The word *éclipse* ("eclipse", 4.29) typically has such a profile. In *Le Monde* the highest coefficients of variation apply mostly to proper nouns, generally politicians (e.g. *Lepage* : 4.61, *Boutin:* 4.58, *Taubira:* 3.952), political parties *(*e.g *UDF-D:* 4.71, *LO-LCR:* 4.12), individuals involved in scandals and affairs (*Lewinsky:* 3.58) and names of places (*Larzac:* 3.29). They also include words that were central to media and political scandals, like *caricatures* (idem) with a coefficient of 4.07, referring to the scandals of Mahomet's caricatures or the word *anthrax* (idem), with a coefficient of 3.54, referring to the scandal of anthrax letters after the 11th September 2001. Words that correspond to sudden media coverage like *canicule* ("heatwave", 3.25), referring to a specific summer and the fears it induced the following summers, also have high coefficients of variation. Words like *dioxine* ("dioxin", 2.63) and *tsunami* (idem,1.82) can be said to have similar profiles: they are extremely frequent for delimited periods of time.

Results in stemmed and unstemmed versions concur, however some words are not captured in unstemmed version and show stronger coefficients in stemmed version like *intermittent* ("casual" in "casual worker", 2.66) taken as an adjective, mostly used in the expression *intermittent du spectacle* ("casual worker in the sector of arts and entertainment"*). Indeed casual workers in the sector of arts and entertainment have been demonstrating vigorously in French streets with subsequent media coverage. The unstemmed version separates plurals and singulars and counts adjectives and nouns altogether, which bias the calculation for words with this type of profile.

In the *NYT*, the results are very similar in terms of content, even though the *NYT* covers a time period which is the double of *Le Monde*. In the highest coefficients of variation, words such as *anthrax* (5.12), *tsunami* (4.40*)* and *Lewinsky* (4.03) echo the French newspaper. The highest coefficient, of 5.71, corresponds to the word *canvassing*, in reference to U.S. Election

campaigns. *Impeachment* (4.55) and the verb *impeach* (3.96 in stemmed version) may echo an unstable treatment of legal affairs in the corpus. Once more, numerous names of politicians emerge (e.g *Kerry*: 3.53 st.[130]), names of political parties and movements (eg. *Hezbollah* 3.97) and vocabulary related to politics (*electors* 5.36 st.).

In the Spanish corpus, words with the highest coefficients of variation are heavily political: (*dictadura, reemplazar, resisten, democraticas, democraticos, prohibido, superioridad, capitalismo* …[131]).

In French and in English, while selecting words with the highest coefficients of variation that also have the highest mean frequency, two major semantic fields appear: politics, war terrorism and violence..

In Le Monde (stemmed version), the highest figures are for politics and politicians as well as names of places (cities and countries, generally related to political decisions and war), for instance:

- FN[132] 2.18 ; circonscription (n.) ("electoral district") : 1.78 ; OTAN ("NATO") : 1.72 ; RPR[133]: 1.48 ; PS[134]: 1.04

- Royal[135] : 2.10 ; Milosevic : 1.96; Juppé[136] : 1.07

- Kosovo : 2.08; Liban ("Lebanon") : 1.97; Belgrade : 1.80; serbe (adj., "Serbian"): 1.77

---

[130] "st." stand for stemmed hereinafter.

[131] Dictatorship, replace, resist (third person plural), democratic (masculine and feminine plurals), forbidden, superiority, capitalism

[132] "Front National" ("National Front") French extreme right party

[133] " Rassemblement pour la République" (« Gathering for the Republic") former French right wing party.

[134] "Parti Socialiste" (« Socialist Party ») French Socialist Party

[135] French politician

[136] French politician

The semantic field of war and terror is almost as strong, with, for instance:

- bombardement (n.,“bombing”) 1.47; otage (n.,”hostage”): 1.00; terrorisme (n. “terrorism”) : 0.95; terroriste (adj., “terrorist”) : 0.94 ; attentat (n., “ attack”) : 0.85, missile (n.,idem) : 0.79

The word *guerre* itself (n.,“war”) is stable (0.52).

The notion of *virus* (n., idem) also has a strong coefficient of variation of 0.88, and associated fears about health and safety also show significant variations, as is analyzed in the following case studies.

Other strong values point to major issues and scandals which touched society in this period. For instance *voile* (n., “headscarf”, “Hijab” 0.77) refers to a major social debate after the prohibition for Muslim women to wear a headscarf in public places. *Immigration* (n., idem, 0.65) and *pollution* (n., idem, 0.64) are also highly debated topics.

The last type of strong values are for cyclical events in sports for instance (*olympique*, adj., “Olympic” 0.89), names of months and the like. The variation in use for these items is predictable, as is detailed in the next section.

The same patterns emerge in the *NYT*, even if the time period the newspapers cover is the double, as show the words *terrorist* (adj.,1.42), *war* (n., 0.49) *hostage* (n.,1.40) *bombing* (n., 1,08) *virus* (0.53). In the same way that home issues come out with high values in French, American home issues follow the same pattern, with, for instance *hurricane* (n., 2.48) and *earthquake* (n., 1.56).

3.2.2. Predictable variations in use: Sample word profiles

It is not because a word has an unusual frequency variation and/or a high coefficient of variation, that it is a candidate for semantic change. Some patterns, however, can be characterized to sort “false” candidates. For instance, some words have a regular seasonal profile (as expected, the press mentions the summer in the summer, and the winter in the winter), others have profiles related to events in the calendar that come with some regularity, such as Olympic Games and political elections.

3.2.2.1. Seasonal and event based variations

Words with single and multiple frequency peaks come out with high regression coefficients and high coefficients of variation. However, this is due to the gap between periods with almost no occurrences and sudden periods with very high frequencies. For instance, in *Le Monde*'s stemmed version, the word *avril* has a high regression coefficient of 0.50 among the 0,50% highest values, as do all months names. The coefficient of variation of 1.45 is comprised in the 0,85 % highest figures. This is due to the sudden frequency ups and downs every ear. In effect, the word *avril* regularly oscillates between a minimum raw frequency of 98 and a maximum of 4454 in the short time span of a month, which makes it a perfect candidate to a high coefficient of variation. All candidate words following this seasonal profile are therefore discarded, among which all names of months.

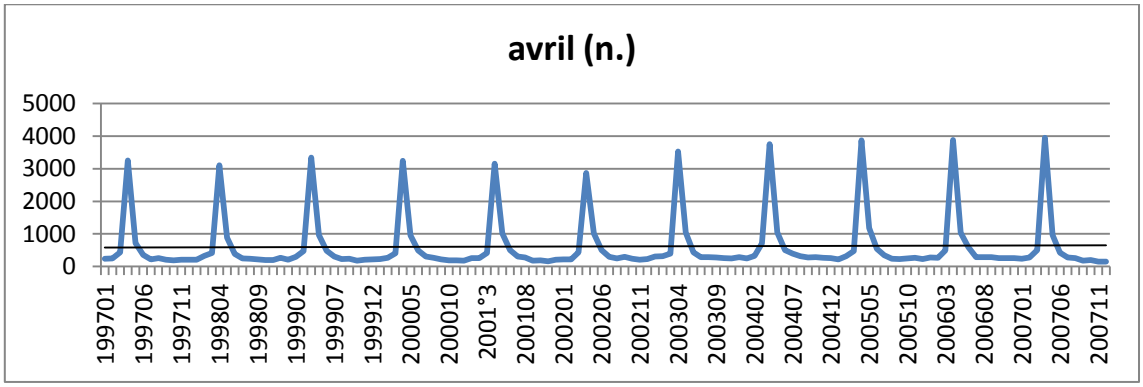


Figure 16 Normalized frequencies of the word *avril* and linear regression in the corpus *Le Monde* (1997-2007) in stemmed version.

In the same way that some words have a very regular seasonal profile, other words related to real-world event with some regularity show high coefficients. All events that appear cyclically in the newspapers provide their share of "false" candidates. Most of them are related to politics and sports, since these domains are highly reported on in the press.

Figure 17 shows the normalized frequencies for the word *circonscription* ("electoral district") in stemmed version. *Circonscription* is characterized by a high coefficient of variation (1.783, among the 0,36 % highest values), but this figure only reflects the "natural" use of this word. The regression coefficient for this word is also strong, with a value of -0.06 (among the highest 4%). Three frequency peaks appear in 1997 for the legislative elections, in 2002 for both the legislative elections in June and the presidential elections in April and in 2007 again for the legislative elections. The word *circonscription* is very rarely used outside of the

context of a French election, and therefore the high coefficient of variation gives us information about word use. Similar patterns of use are found for words which follow cyclical patterns linked to regular events, such as *olympique* ("Olympic", with a coefficient of variation of 0.89, among the 5,59% highest values).

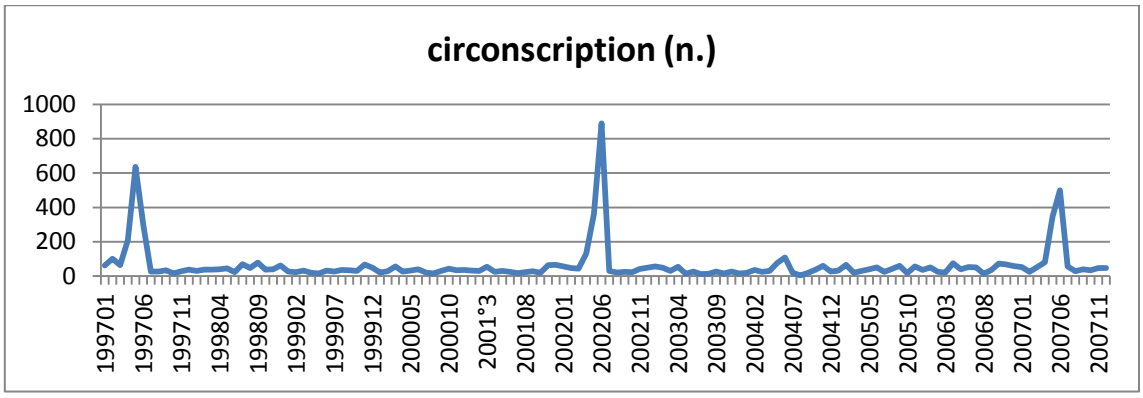


Figure 17 Normalized frequencies of the word *circonscription* in the corpus *Le Monde* (1997-2007) in stemmed version.

Other less evident words appear in peak patterns: for instance the word *terroriste* ("terrorist") rocketed in the press, on and after the 11th of September 2001, and adjectives describing the population involved in conflicts and wars also rocket abruptly every time there is an attack.

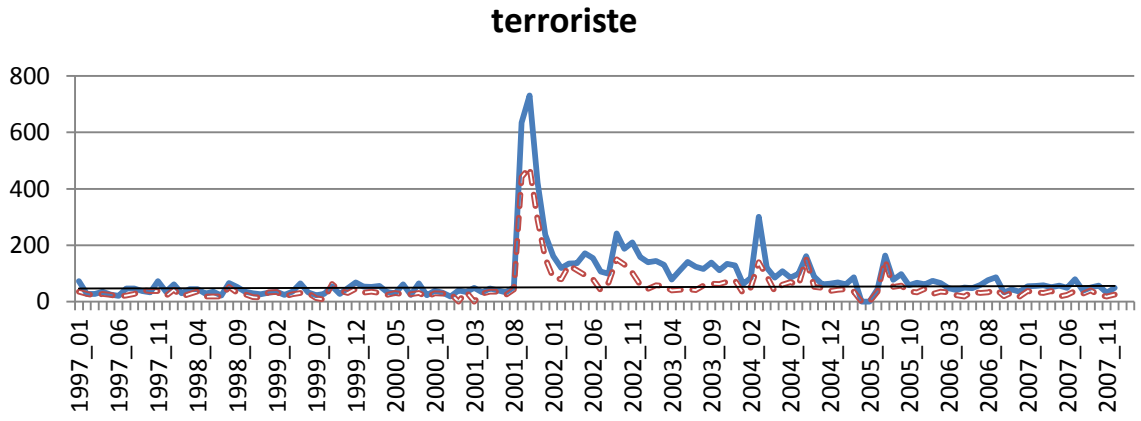


Figure 18 Raw frequencies of the adjective (in red dots) and the noun *terroriste* (blue line) in the corpus *Le Monde*, (1997-2007) in stemmed version.

However, with regards to the word *terroriste*, the peak of the 11th of September triggered a stable increase in use, reflected by the positive regression coefficient of 0.72 for the adjective and of 0.31 (highest 0.77%)for the noun, and debate about the definition of this word arose in

reaction to media over-use. The noun has a coefficient of variation of 0.94 (highest 4.76%) and the adjective of 0.88.

Before the 11th of September the mean of the added frequencies of the adjective and the noun is 68,5. During September and October 2001 it is 1141, and after that 158,51. The shock of the event and the overwhelming media coverage triggered this increase in use. Every time an expression comes to the front scene in the media, it is played with, questioned, and, as such, evolves in the collective perception of its meaning. One of the linguistic clues which show this process is the morphological productivity of the word. *Terroriste* has 51 compound forms in the corpus. Most of them emerge directly during and after the 11th of September (40 out of 51), as shown in Figure 19.

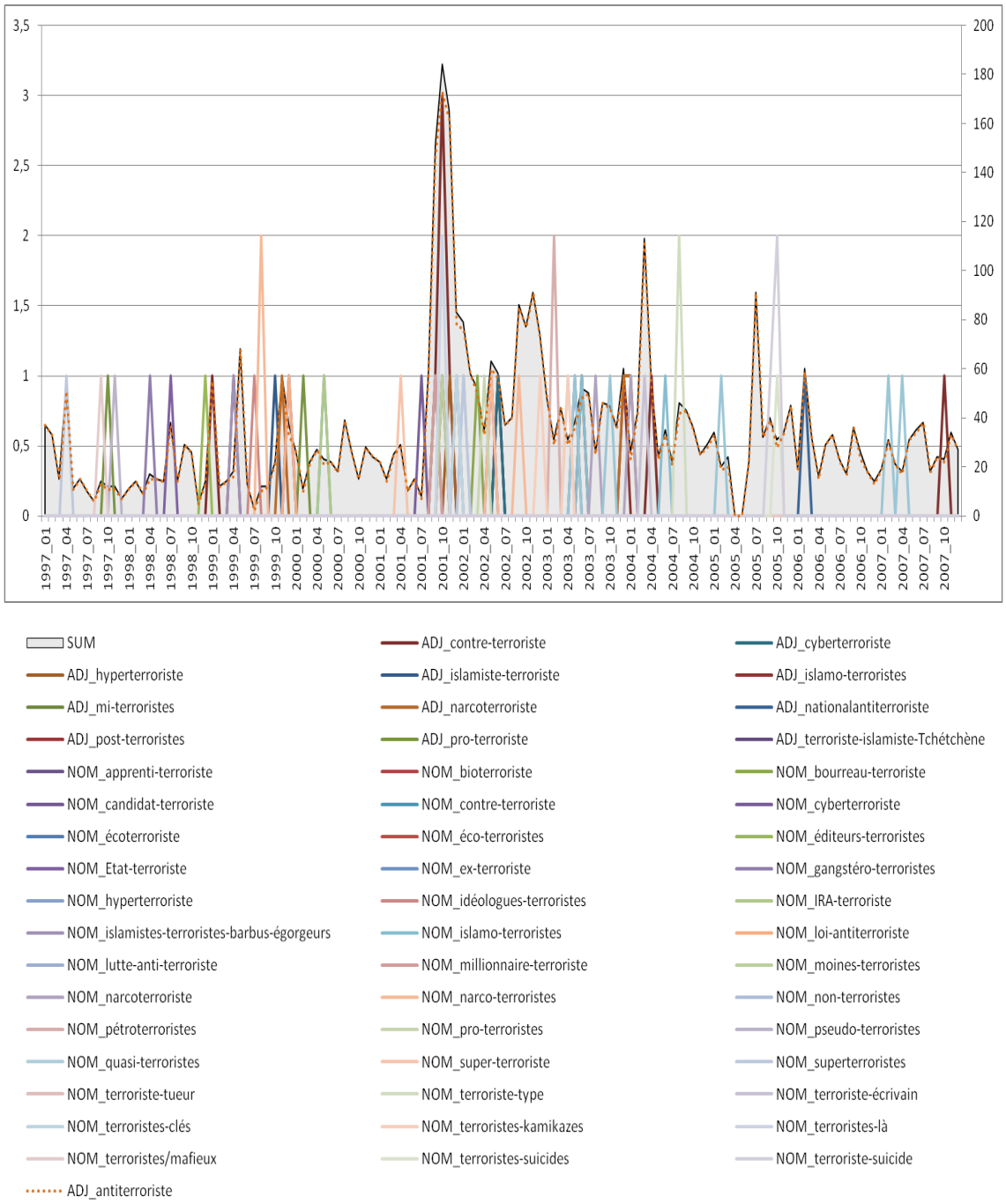


Figure 19 Raw frequencies of compound words based on *terroriste* in *Le Monde* (1997-2007) in stemmed version.

The dotted line for the adjective *antiterroriste* as well as the sum (filled in gray), have graduations of the z axis. (Adjectives and nouns are mentioned in the legend in the singular but counts include both singular and plural occurrences.)

Besides the frequency pattern of the adjective *antiterroriste* (4443 occ. in total, with a strong regression coefficient of 0.318), there are numerous creations linked with the event such as *islamo-terroriste*, and productive prefixation and composition in *bio-*, *hyper-*, *super-*, *post-*, *pro-*, *non*, *ex-*, *quasi-*, *pseudo-*, *cyber-*, *éco-*, *narco-*, *etc.* *Bioterroriste* (with and without a hyphen) shows the greatest frequencies with 123 occ. Neologisms show frequencies under 70

occurrences per month whereas *antiterroriste* reaches 173 occurrences at its highest. Moreover, creative compound words with very few occurrences appear, such as *idéologue-terroriste* ("ideologist-terrorist"). Therefore, all words that show abrupt peaks of use should not be discarded, since this peak is sometimes the reflection of a truly marking event that can induce debate on word definition, as well as creativity.

3.2.2.2. Idiomatic and polysemy variations. Network co-occurrence and ranks

High variations can also be indicative of word use at the levels of polysemy and idiomatic range. The detected variation is then a rather natural one, since highly idiomatic and polysemous words are at the highest threshold of natural variation. However, within this natural variation, specific lexical units can take on more importance and salience than others, and the target word can undergo increase or decrease in use whilst this shift takes place.. Measures like the coefficient of variation or the regression coefficient are not informative enough, since they are only based on word frequency. These measures cannot show the difference between natural variation and variation indicating semantic change because they take every variation as a unit, regardless of their semantic and linguistic status. In this case, it is the variation of the network of co-occurrence that can provide information.

For instance, the words *barre* (Figure 20) and *bouquet* (Figure 21) have strong idiomatic and polysemous uses. They have a volatile but stable frequency pattern. Measuring the coefficient of variation does not grasp their variability between several meanings and uses.

Barre is used to refer to buildings *(barre d'immeubles, HLM*), in *barre de fer* as a metal bar, in the expressions *un coup de barre* ("to get really tired suddenly") and *atteindre/passer la barre des* …("to reach/go over the threshold of…"). *Barre* has both a high positive regression coefficient (0.05 a value comprised in the highest 4.95 %) and a low coefficient of variation of 0.21 (a value comprised in the highest 76.91%).

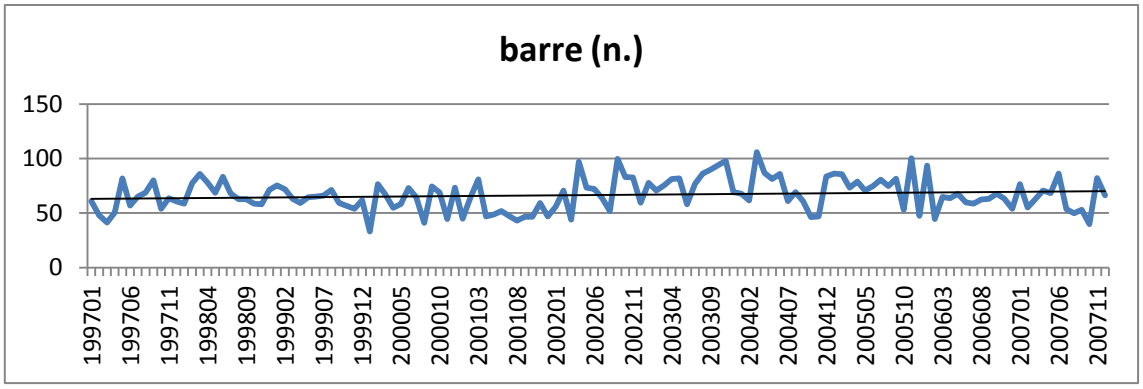


Figure 20 Normalized frequencies of the word *barre* (n.) in the corpus *Le Monde* (1997-2007) in stemmed version, with linear regression

Bouquet can be used to refer to a bouquet of flowers, and by extension to a television channels package; it can also refer to the taste of wine and the smell of perfume and it appears in idiomatic expressions such as *c'est le bouquet!* (approximately translatable as "that's all it needed!"). Its use clearly decreases in the corpus while its frequency pattern remains volatile. *Bouquet* has a very high negative regression coefficient of -0.27 (a value comprised in the strongest 1.12 %), and a stable coefficient of variation of 0.47 (a value comprised in the 41.94% highest).

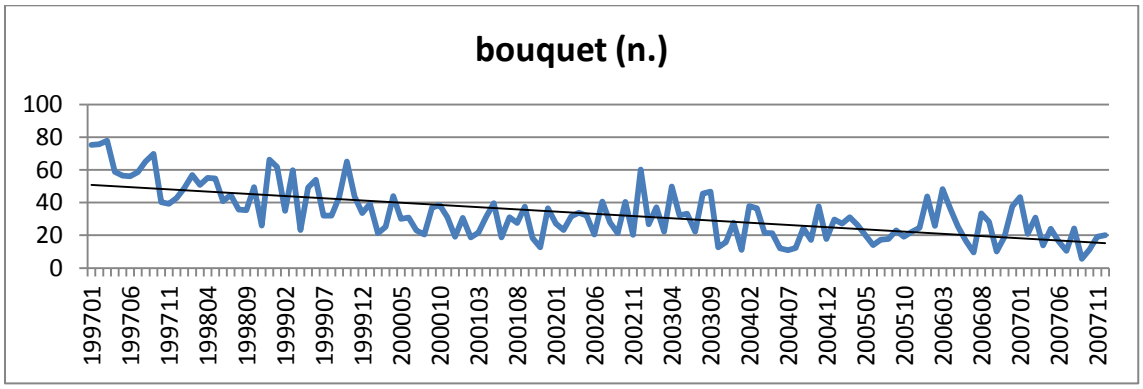


Figure 21 Normalized frequencies of the word *bouquet* (n.) in the corpus *Le Monde* (1997-2007) in stemmed version, with linear regression.

In the case of *barre*, the meaning "to reach a threshold" shows enhanced salience in a limited time span, which may explain the volatility of its frequency behavior , while in the case of *bouquet* it is the meaning "television channels package" that is progressively becoming more salient as the word decreases drastically in frequency.

Both *barre* and *bouquet* have strong regression coefficients showing a shift in use, however their coefficients of variation are low. Both words have a rich idiomatic profile. *Barre* can be used by extension and analogy of its core meaning in domains as diverse as ships, law, games, mechanics, music, anatomy, physics, building, etc. *Bouquet* has two major meanings (it has four in total but only two are common use), one related to a set of plants, and the second to the nature of a smell or taste.

To see whether these meanings undergo a shift and whether a new meaning, analogy or extension appears, the network of co-occurrent words is observed (as described in Part II, chapter 2). All co-occurrent words are sorted hierarchically, on the basis of the co-occurrence mean frequency per month, in a table.

The term *barre* is increasingly used in contexts in which it means "threshold", and the threshold is generally a numerical one, related to economic figures and percentages. While looking at the strongest co-occurrent words for *barre*, the term *déficit* ("deficit") shows higher frequencies from the end of 2001 to the end of 2005, as shown in Figure 22, probably due to the discussions raised by the economic crisis.

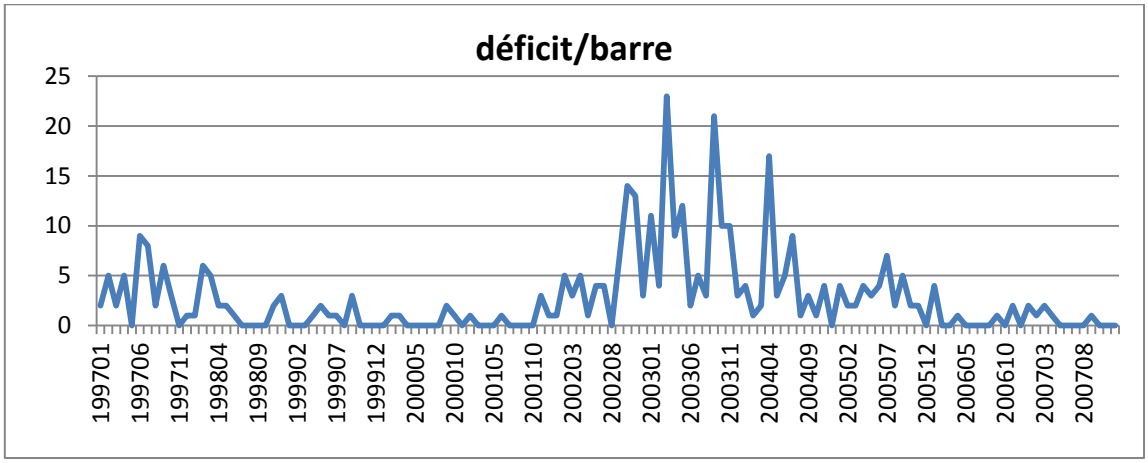


Figure 22 Raw co-occurrence of the nouns *barre* and *déficit* in the corpus *Le Monde* (1997-2007), in stemmed version.

To investigate the co-occurrence of the noun *déficit* with *barre*, ranks provide further information about the hierarchical order of co-occurrent words.

The ranks classify co-occurrent words by order of importance: ranks closest to 1 are more important. Therefore, Figure 23 should be read as an increase in importance of the co-occurrent word *déficit* compared to other co-occurrent words for *barre*. The graph shows a

change in status of *déficit*, since the ranks are high from 1997 to the end of 2001 (while co-occurrence is sparser) and low from 2002 to 2005 (with denser occurrences).

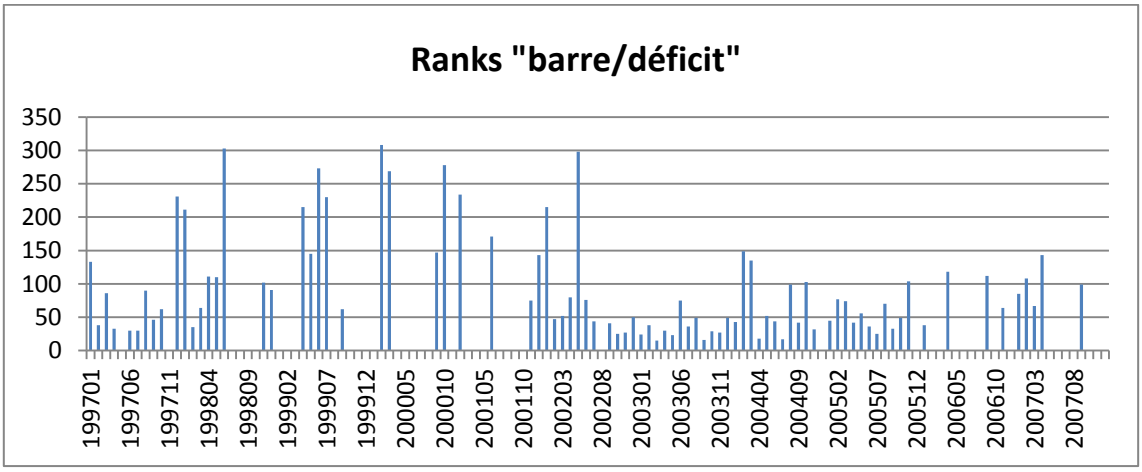


Figure 23 Ranks for the co-occurrence of *déficit* (n.) with *barre* (n.) in the corpus *Le Monde* (1997-2007) in stemmed version. Ranks closer to zero show a higher order of importance.

However, other co-occurrent words of *barre* show a stable use, such as *coup*, used in the expression *coup de barre*, as shown in Figure 24. The slight increase in the order of co-occurrent terms cannot be analyzed as a change in status, since the ranks count takes into account all co-occurrent words, including function words, and therefore a slight variation of this type is read as a natural variation of use.

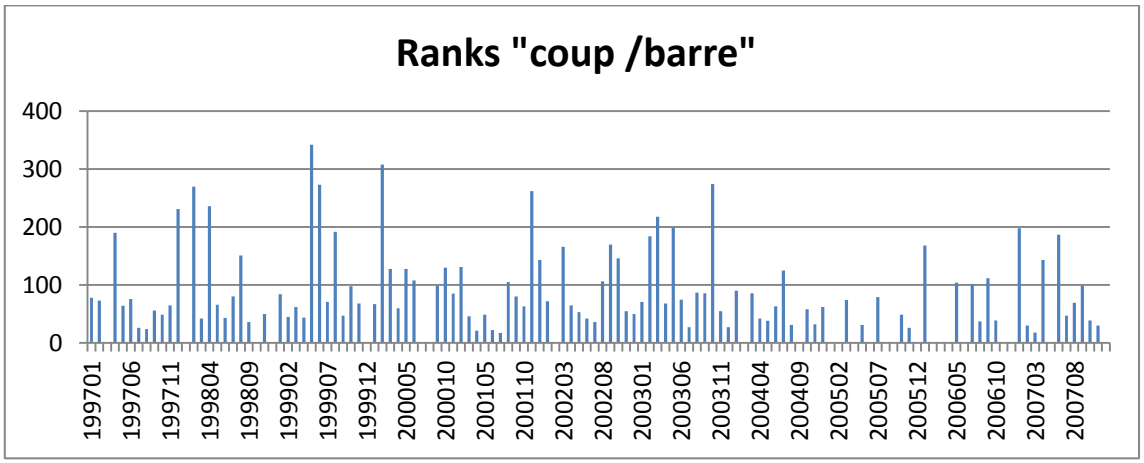
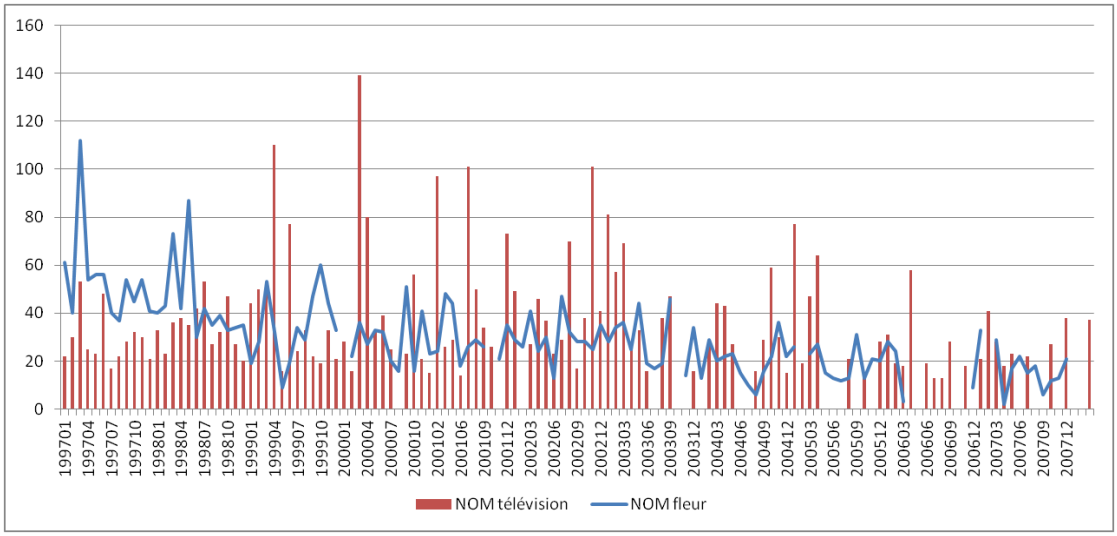


Figure 24 Ranks for the co-occurrence of *coup* (n.) with *barre* in the corpus *Le Monde* (1997-2007) in stemmed version.

The noun *bouquet* presents a similar idiomatic profile. As seen earlier, it is less and less used in the corpus (as shown in Figure 20). It is increasingly co-occurring with *télévision*



("television") and the ranks of *télévision* are stabilizing at a low threshold, showing that the idea of television channels packages is normalizing, while the association of *bouquet* with *fleur* ("flower") is reinforced, as shown by the ranks graph in Figure 25. The meaning of

Figure 25 Ranks for the co-occurrent words *télévision* and *fleur* with *bouquet* in the corpus *Le Monde* (1997-2007).

bouquet as a television package is not listed in the TLFI or in the Dictionnaire de l'Académie Française. In France, the first television packages appeared in 1992, and therefore this meaning is relatively recent from the terminological point of view. However, the GRLF[137] lists it as follows:

"5 Télév. Bouquet de programmes, bouquet numérique : ensemble de programmes télévisés payants, diffusés par le câble ou le satellite et proposés aux téléspectateurs par l'intermédiaire d'un opérateur. Bouquet de programmes thématiques regroupant plusieurs chaînes sportives, culturelles, musicales.[138]"

[137]Le GRLF. The GRLF was issued in 2001 and updated in 2011.Dictionaries published by Le Robert and Le Larousse benefit from yearly updates, and are more prone to include new meanings.

[138] 5. Television. Program package, digital package: a set of TV programs, available through cable or satellite and available to viewers via an operator; package of thematic programs containing several sports, cultural, and musical channels.

3.2.2.3. Fluctuation

As seen through these examples, it is difficult to distinguish natural variation (fluctuation) from semantic variation. Indeed, *circonscription* shows frequency peaks but does not change, in use or in meaning, while *terroriste* shows a major frequency peak and sees its use rocket and its meaning questioned. It is noteworthy that some acts that may have been called *mass murders*, *attacks*, *killings* or *crimes* before the 11th of September come to be referred to as *terrorist acts* after it. *Barre* and *bouquet* both show a rich idiomatic and polysemous use, but *barre* remains quite stable while *bouquet* acquires an additional meaning. The first "fluctuates" while the second integrates an additional meaning by extension of the concept "bunch", as in "a bunch of flowers", being applied to television channels. A bunch may group things that grow together, like flowers, or items that are alike, like channels. I take the expression "fluctuation" from Stern (1931) who defines it as follows:

"Every utterance is the expression of the momentary state of mind and purposes of the speaker. Every man has –within reasonable limits- thoughts and feelings to express that are peculiarly his own, and also his own way of expressing them (…) strictly speaking the psychic processes always vary from instance to instance, and from individual to individual. The variations are largely variations of context –of "setting"- due to the constantly shifting circumstances in which the word is employed, and to the similarly shifting apprehension and purpose of the speaker." (Stern 1931: 162, 163)

Fluctuation starts with the personal relationship to language and encompasses context variations. Moreover, it is my contention that the nature of words, in terms of their degree of polysemy and idiomatic range, add a linguistic layer to this fluctuation process. When a certain type of fluctuation becomes widespread enough, it can modify a word's meaning. Both fluctuation and semantic change rely on the same process: the variability of connotation and the multiplicity of use of several connotational values for a word, depending on context, taken at the linguistic and sociological levels. Fluctuation is very volatile and cannot be captured fully with measures. However, measures assessing the variation of the network of co-occurrences of a word provide a profile that can be further investigated with analysis involving real-world knowledge. Some research in NLP directly integrates tools such as Wikipedia in their programs to provide for this knowledge. This field of study could be complementary to the type of indices presented in this work.

Assessing fluctuation also means that semantic change processes can be observed one step before they are identified while capturing epiphenomena that may impact meaning later or not. The use of the word *terroriste* has changed in the French and American press, but whether its meaning will continue to be subjected to strong fluctuations is not yet predictable with our current means. This phenomenon is not limited to the press, but it is yet another object of study to assess its diffusion in society and the nature of that diffusion.

The changes in connotation are a type of fluctuation. At the basic level, every word is subjected to change in connotation, from use to use and individual to individual. At another level, when a certain connotation gains enough power to modify meaning (and that threshold cannot be grasped), fluctuation becomes the source of meaning change. The next step is the recognition of that process, via lexicalization. Between idiosyncratic variation and lexicalization, all we observe is the dynamics of fluctuation. From then on, linguists have to set a threshold as to when meaning is said to have been modified, and that threshold is notorious for the disagreement it generates among experts.

3.2.3. Case studies

3.2.3.1. *Malbouffe* and *mal-*

One of the ways to test indices for detection is to work on attested cases. This is the method encountered in Cook and Stevenson (2010) and Sagi, Kaufmann and Clark (2009). In French, an Appendix to the doctoral thesis of Martinez (2009) provides a list of all new words which entered (and exited) the dictionaries Le Larousse and Le Petit Robert, from 1997 to 2008[139]. Both dictionaries are known for being prompt to integrate changes. Martinez's (2009) doctoral thesis is primarily concerned with spelling variations. The latter is one of the indices of semantic change and neologisms, since, as new meanings and forms emerge, there is hesitation on the part of writers as to their spelling, before consensus takes place. From Martinez's work, the word *malbouffe* (*la malbouffe*: "eating junk food") was selected since it is a neologism undergoing semantic change and a word undergoing spelling variation. This item was selected because it enters the corpus as well as the dictionary suddenly. Its meaning shift was integrated in its definition and use at defying speed. The fact that *malbouffe* contains

[139] At the time of access

the element *mal-* was at the origin of a study on the morphological productivity of this blurred element of composition in the corpus[140].

Neologisms generally take a long time to enter dictionary updates (overall this process was measured in decades until recently, but this time span may come to be reduced in accordance with current creativity and borrowings trends.) New terms are generally submitted for consideration by terminologists a year before the publication of a dictionary. In the quickest cases, it takes one to three years for the word to be attested in use, discussed by terminologists and included in the dictionary. In the case of *malbouffe*, the time span for this process is extremely short in terms of lexicographic rules: two years between its first occurrences in our corpus and its publication in dictionaries. In effect, *malbouffe* appears in the corpus at the end of 1999, and is included in *Le Petit Larousse* and *Le Petit Robert* in the 2001 editions. It is my contention that this sharp acceleration process may be observed for other words.

The PR defines *malbouffe* as follows:

> "Malbouffe : 1999 ; « mauvaise alimentation » 1979
>
> de 1.*mal* et 2.*bouffe.* FAM. Aliments dont les conditions de production et de distribution nuisent à la qualité et à la sécurité de l'alimentation (pollution, épizooties, hormones, OGM...).[141]"

The *Larousse* gives a shorter definition and notes "on écrit aussi mal-bouffe" (a spelling variation is *mal-bouffe*), but this comment is removed in the 2002 edition. The fact this comment is removed shows that the word's spelling has reached consensus. The reference to 1979 corresponds to the publication of a book titled "La Mal bouffe, comment se nourrir pour mieux vivre" ("Junk food, how to eat to live better") by Stella and Joël de Rosnay (1979). The book wass re-edited in 1981 with the spelling *malbouffe*. The authors drew the concept of malbouffe from the existing word malnutrition (idem) and the idea of "grande bouffe" ("big

---

[140] The original study on *malbouffe* was published in French in (Boussidan, Lupone, and Ploux 2009)

[141] Malbouffe : 1999, " bad quality diet" 1979

fom 1.*mal*("bad") et 2.*bouffe* ("food"/ "junk food"). Familiar. Foods whose conditions of production and distribution affect the quality and security of one's' diet (pollution, epidemics, hormones, GMOs ...).

meal") taken from the movie "La grande Bouffe[142]" ("The Blowout") by Marco Ferreri, depicting a group of four men who decide to commit suicide by eating until death. When the De Rosnays use the term *malbouffe*, they use it in a dietetic meaning, to underline the effects of bad food on people. What they call bad food at the time, is mostly food with high sugar and oil contents. This meaning is also released in the media by Jean Pierre Coffe,[143] a famous French cook. The meaning related to the conditions of production of foods appears later, as discussed subsequently. In a sense, the De Rosnays can be considered the first "parents" of this word.

3.2.3.1.1. ***Malbouffe***

3.2.3.1.1.1. Frequencies

The initial study on *malbouffe* was conducted on the corpus *Le Monde* 1997-2001[144] and published in French in Boussidan, Lupone, and Ploux (2009). Here, the corpus covers 1997 to 2007. The period from 1999 to 2001 alone contains all the elements of understanding of the phenomena. Raw frequencies of *malbouffe* and *mal-bouffe*, shown in Figure 26, show that from the beginning of 1997 to August 1999, the word does not appear. The first occurrence at this date is under the spelling *mal-bouffe*. The next month, there are 11 occurrences, followed by 7 occurrences in October and 19 in November. These are the highest frequencies in the corpus. After that, the term keeps on appearing and undergoes frequency variations comprised between 1 and 12 per month. Even though a new word is created and enters the dictionary, its frequency drops:

"Typically, in the life-cycle of a word or phrase in text, a rise in lexical creativity is followed by a fall." (Renouf 2007 : 72)

According to Renouf, this fall is observable after the birth of a word and its peak. After this frequency fall, the word stabilizes.

---

[142] Original Italian title : « *La grande abbuffata* ».

[143] Creator of the SCMB ("Société contre la malbouffe")

[144] At the time I conducted this study in 2009 we had not acquired the corpus Le Monde from 2001 to 2007 yet.

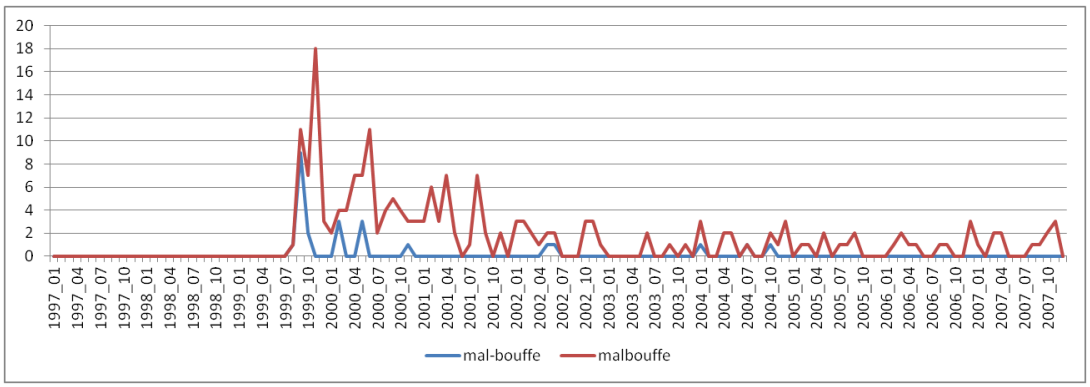


Figure 26 Raw frequencies of *malbouffe* and *mal-bouffe* in the corpus *Le Monde* (1997-2007) in unstemmed version,.

On the whole, the frequency figures for *malbouffe* are very low, with a total of 197 occurrences over ten years (173 for *malbouffe* and 24 for *mal-bouffe*). These low numbers could make such a word go undetected by programs which filter out extremely low frequencies to clean the data from typographic errors and epiphenomena.

The low frequencies also make the low figures for the regression coefficient (+0,00307 unst[145].) and the coefficient of variation (0,819 unst.) meaningless.

The term first appears with a hyphen but this spelling virtually disappears afterwards, with a total of 5 occurrences from 2001 to 2007, compared to a total of 52 occurrences for the spelling *malbouffe* in this period. The period between the end of 1999 to the end of 2001 sees the highest frequencies. After that time, the frequencies become very low (1 to 4, cumulating both spellings) but they are regular.

The spelling variation is one of our linguistic indices. Moreover, the hesitation on the part of writers is shown by the use of inverted commas. Punctuation marks are also a linguistic index. Both the hyphen and the inverted commas almost disappear over time, as the word enters common usage.

If figures say something about words, the source text primarily explains the sudden apparition of the term: as in many cases, it is related to an event, covered by the media. The first time the word *malbouffe* is employed, it is in reported speech, by François Dufour, spokesman of the

[145] Unstemmed. Figures are given in unstemmed version since Treetagger misses out on neologisms.

farmers' confederation, after the "dismantling"[146]of a McDonald's in Millau on the 12th of August 1999. This time, it appears without inverted commas. The event was led by José Bové, and the press kept following its consequences since Bové ended up in jail, and a part of the public opinion rallied in support to him and his actions. The occurrences of September and October 1999 correspond to these events. This moment is marked with additional media coverage about food quality: the media talks about a "food crisis" related to mad cow disease, dioxin chicken, GMOs and hormones contained in meat. These elements participate in creating a generalized anxiety towards food quality. However, the term *malbouffe* is strongly connected to José Bové.

3.2.3.1.1.2. Co-occurrence network

Indeed the highest co-occurrent word of *malbouffe* within the period of higher frequencies (end of 1999 to end of 2001) is *Bové* (39 co-occ.[147]) and the second *José* (35 co-occ.). Figure 27 shows how strongly interrelated the two terms are, by extracting the raw frequencies for *Bové* and superimposing them with those of *malbouffe*.

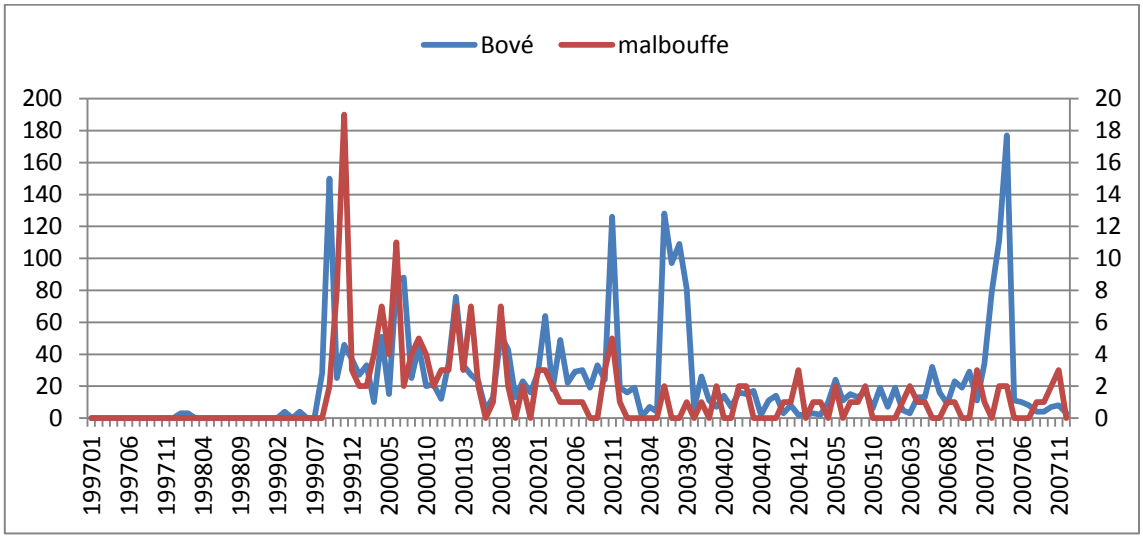


Figure 27 Raw frequencies in *Le Monde* (1997-2007) in unstemmed version, for *Bové* , measured on the y axis and *malbouffe*, on the z axis.

[146] "Taking down" is the term used by the actors of this event ("démontage" in French).

[147] co-occ. is used as an abbreviation of "co-occurrence" hereinafter.

Two months after *Bové* abruptly enters the corpus with its highest frequency, *malbouffe* does too, after a first frequency peak where both words appear together. This shows that the entrance of *malbouffe* in the corpus is tightly linked with Bové's actions and his trial. The word patterns are clearly related from the beginning of the corpus to 2002, date after which the word *Bové* has frequency peaks unrelated to malbouffe. Therefore, the semantic networks that define *Bové* may influence those of *malbouffe* in this first time period. Moreover, José Bové actively participated in defining and spreading the word *malbouffe*, which was (and maybe still is) a central concept in his discourse. He employed it to a great extent, and gave it connotations that it did not previously have. The fact that Bové was (and maybe still is) considered a leader by some, helped spread his idiosyncratic associations to a larger number of people. The co-occurrence networks show a tight link: among the co-occurrent words for *Bové* are all the major co-occurrent words for *malbouffe*, including *José, confédération* ("confederation"), *paysanne* (as a fem. adj."farming"), *monde* ("world") and *Mcdonald*, which are part of the ten strongest co-occurrent terms in both cases:

> *Bové:* José (1480), confédération (295), paysanne (285), contre (183), Millau (174) monde (96), prison (96), procès (94) McDonald (85), mois (73) porte-parole (70), agriculture (69), France (69) mondialisation (68), mouvement (67) Seattle (67), paysan (66) été (66) août (66) militants (63)… combat (48), McDo (45), OGM (42)… pays(39)… lutte (34)[148]
>
> *malbouffe* : contre (89) lutte (26), mondialisation (23), paysanne (17), confédération (16), combat (15), monde (15) pays (14) mcdo (14)

This shows that these two networks are intertwined. Figure 27 shows raw frequencies of major co-occurrent words for both *Bové* and *malbouffe*, independently of their frequencies (*contre, lutte, mondialisation, paysanne, confédération combat, monde, pays, McDo*). They follow a coordinated pattern from the end of 1999 to the end of 2001, when *malbouffe* enters the corpus. The coordination of the whole semantic network is visible in Figure 28. When the frequencies for *malbouffe* are high, the whole network is active. This is particularly visible in September 1999,

[148] Confédération : confederation ; paysanne : farming ; contre : against ; prison : jail ; procès : trial ; mois : month ; porte-parole : spokesperson ; mondialisation : globalisation ; paysan : farmer ; été : summer ; août : August ; militants : activists ; combat : fight ; OGM : GMO ; pays : country ; lutte : struggle.

May 2000 and March, September and November 2001. *Malbouffe*'s co-occurrence network is clearly interwoven with Bové's.

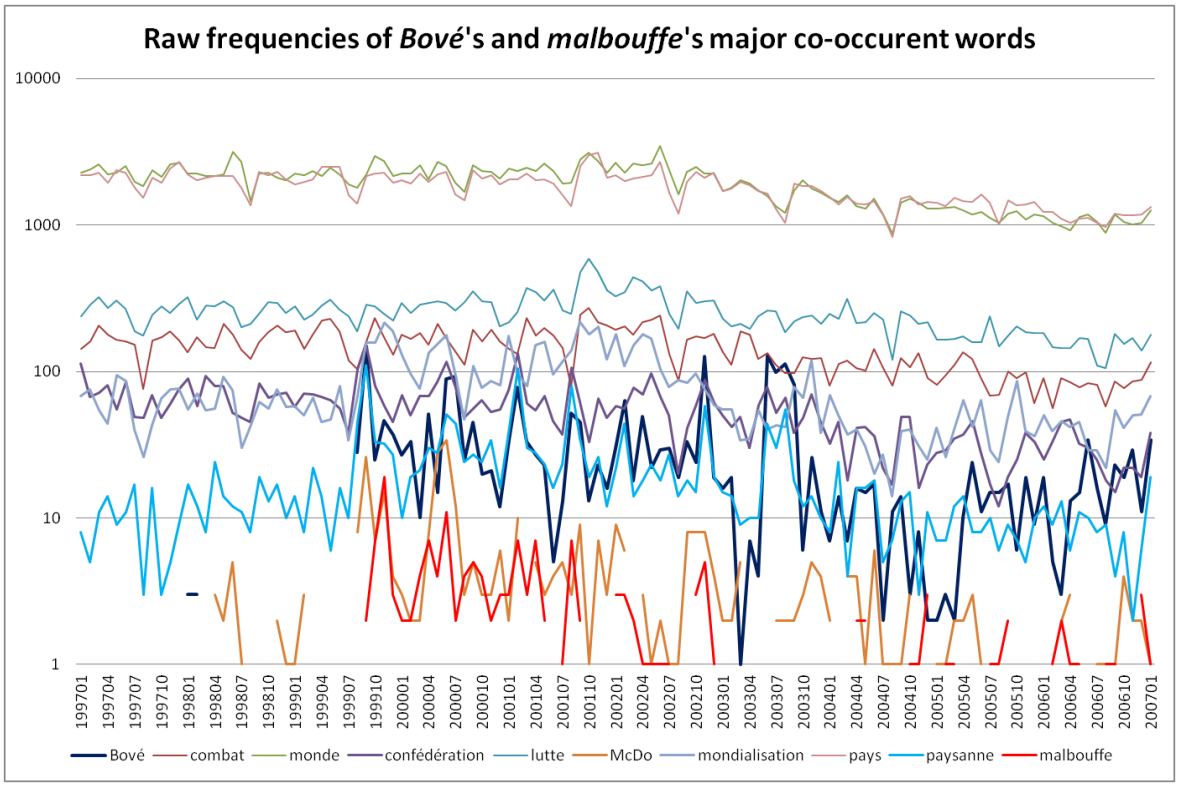


Figure 28 Raw frequencies for the words *malbouffe* and *mal-bouffe* (added up) and their strong associated co-occurrent words in the corpus *Le Monde* (1997-2007), in unstemmed version, taken separately: *Bové, mondialisation, lutte, paysan, confédération, combat , monde, pays* and McDo.

3.2.3.1.1.3.Semantic shift and terminological stabilization

The word *malbouffe* is subjected to a superimposition of new meanings onto preserved older meanings. This superimposition takes place as new criteria emerge in the public opinion, due to a societal crisis regarding food, sustainability, and relationship to nature and its exploitation. This crisis is different from José Bové's discourse which includes ideas about what could be summed up as an activist's way of eating, inscribed in the framework of a conscious citizen act showing disapproval of the current economic system (and the establishment, at a larger scale). As for the food crisis, it creates a need to redefine the criterion for food quality, and generates discussion. The word *malbouffe* and the concept of "bad diet" that is associated to it change hand in hand. The original meaning, "bad diet", is

enriched, evolving from referring to the nutritional value of foods to additionally including a relationship to the conditions of production of these foods in the context of the industrial society. The dictionary keeps a trace, albeit withered, of the political implications which surround the term in Bové's discourse. To eat well is additionally defined as eating GMO-free foods. The 2001 one editions of the pre-cited dictionaries list GMOs among conditions of production; the acronym GMO entered the dictionaries in 1992. The original stable definition, anchored in dietetics, is enriched with a new definition which adds notions related to the conditions of production and distribution as well as security, and are reflected in concrete examples such as pollution, epidemics, hormones and GMOs[149]. The cited examples are all part of Bové's discourse. It seems that the time period of terminological stabilization of meaning takes place after a double "crisis": Bové's conviction and the food crisis.

There are therefore two lexical units, each with its associated semantic network: The first is restricted to dietetics and refers to "eating badly", too fatty and sugary foods, it is symbolized by the co-occurrence of *mieux* ("better"). The second is impacted by both crises and is symbolized par the co-occurrence of *contre* ("against"). It is connoted by its co-occurrent words: globalization, the United States, the struggle against consumerism, the industrial society, McDonald's, etc. This second lexical unit rehabilitates notions which are contained in the first lexical unit, by giving them a new role in the semantic network. Dietetics and health are joined by tradition, authenticity, and attachment to the French land and soil (fr. "terroir"). The semantic network of the first lexical unit is weak. Indeed, the first plain words[150] to co-occur with *malbouffe,* belonging to the semantic network of dietetics are *qualité* ("quality", 6 co-occ.), *produits* ("products", 6 co-occ), *santé* ("health", 5 co-occ.), *obésité* ("obesity", 5 co-occ.) and *alimentaire* ("dietary", 5 co-occ.)

The word's meaning is quickly set and integrated into language. The contexts of use can help explain this speed of terminological stabilization. *Malbouffe* is employed repetitively, taking on the status of a motto. The motto effect is both in Bové's discourse and in journalistic writing which repeats Bové's habit of use (either to support or criticize it) and cements the relationship between the *malbouffe* and Bové. Additionally, the press depicts Bové with

[149] Genetically Modified Organisms

[150] After removing function words and common verbs.

nicknames such as “héro mondial de la lutte contre la 'malbouffe '” (“world hero of the struggle against junk food”) “porte-parole du combat contre la 'malbouffe'” (“spokesman of the struggle against junk food”) and “symbole de la lutte contre la 'malbouffe' ”(“symbol of the struggle against junk food”), among other examples. In 2000, Pierre Georges writes ironically about the fictitious situation of helping an American journalist recognizing Bové among the French. He writes that if the numerous characteristics that define him are not sufficient (strange, wears a moustache, wears a beret), they would have to go looking for the “guy who invented the concept of ‘malbouffe’”[151]. This ironical statement bears its share of truth, since it underlines that a process comparable to “paternity” ties Bové with the word *malbouffe*.

3.2.3.1.2. ***Mal- : morphological productivity***

The quick integration of the term is linked with events and the media, but it is also facilitated by other factors of cognitive nature. First, it is a compound word. It is psychologically easier to welcome and memorize a new word made from two known words than a completely new form. Second, its morphological structure in *mal-* evokes other known terms that are similarly built. The fact that *malbouffe* is built on *mal-* raises morphological questions. Indeed, morphological productivity is one of the key loci to assess words in diachrony.

“Productivity is the term used to refer to the word formation processes wrought upon a lexeme. If a word is “productive”, it means that associated grammatical and derivational variants are being produced.”(Renouf 2007:63)

It is questionable whether the *mal* in *malbouffe* comes from the adjective *mal*, from Lat.*malus* as in the compound *mal(-)connaissance* (“bad, incomplete knowledge”), or from the adverb *mal* form Lat. *male* in adjectival and nominal position as in *mal(-)aimé* (“badly loved, unloved”) or from the noun *mal* from Lat. *malum*.

In De Rosnay’s original title the word *malbouffe* was used as a two word expression (“mal bouffe”), in which “mal” could be interpreted as an adjective or a noun. However, the construction mimics other words built on the element *mal-* associated with a noun. Therefore,

[151] In French in the article: “le type qui a inventé le concept de la ‘malbouffe’”.

several explanations are considered, and several of them could have contributed to the formation of the word.

Most compound words in *mal-* are old and are confixations (also called "neoclassical combinations", Fr; "composition savantes"). Confixations are compounds based on elements of composition of Greek and Latin origin, in specialized domains (such as science, medicine, technology, and the like). An example of confixation in *mal* is *malhabile* ("clumsy"). *Malhabile* is the association of an adverb and an adjective, which is one of the frequent compound patterns of *mal-*, along with associations of adverbs and verbs, in the past participle and present, as in *mal-voyant* ("visually impaired" a politically correct version of "blind"), and *mal-logé* (lit. "badly housed").

Barraud (2008) notes that *malbouffe* is the first construction associating *mal-* with a familiar word. In her opinion, the creation of *malbouffe* triggers a "new wave of compound words"[152] in *mal-*. The author mentions: *malboulot*,malconsommation, malrépartition, malurbanisation, malinformation, malpolitique, malgestion, malutilisation, malmenage*, malintérprétation, malconnaissance, malconception, malfréquentation, maldéveloppement, maléconomie, malagriculture.*[153]

This idea can be investigated by extracting all compound words in *mal-* and *mal* in the corpus. The extraction brings up a series of innovative words, such as:

- Malboulot, mal-logés, mal-logement, mal-vie, mal-vivre, malbonheur, mal-développement, mal-comprenant, malformatives, mal-fondé, mal-croyants, mal-protégés, mal-informés, mal-mariés, malfonctionnement, mal-administration/

[152] Original expression in French is "une nouvelle vague de mots composés".

[153] Items with a * were also found in the corpus. *Mal* and *mal-* are translated as "bad", "badly" and "mis-" in the following made-up or existing expressions and simply try to grasp meaning and respect the original form. malboulot* :"badwork" ; malconsommation : "badcomsuption" ; malrépartition : "baddistribution" ; malurbanisation : "badurbanisation" ; malinformation : "badinformation" ; malpolitique : " badpolitics" ; malgestion : " badmanagement" ; malutilisation : " misuse", malmenage* : "mistreatment " ; malintérprétation : " misinterpretation"; malconnaissance : "misknowledge" ; malconception : "misconception"; malfréquentation: "badcompany"; maldéveloppement: "baddevelopment"; maléconomie: "badeconomy"; malagriculture: "badagriculture".

maladministration, mal-emploi, mal-eBay, mal-penser, mal-pensant, mal-pensance/malpensance, mal-Bové…[154]

Most of these innovative words appear with and/or without a hyphen. If these words follow the same process as *malbouffe*, these spelling variations may reflect a process of lexicalization. Some are new morphological forms based on pre-existing terms, such as *malformative*, based on *malformation* (idem, or "deformity"). The latter, a semantically stable word lexicalized in 1867, becomes productive. Other new terms appear in various morpho-syntactic forms that do not follow the rules of word formation inherent to their category, such as *mal-vie* and *mal-vivre*. *Mal-vivre* has the form of a verb but it is only employed in the infinitive form and *mal-vie* seems to mimic the structure of *mal-être*, in a nominalized form (*le mal-être* and not *mal-être* in the infinitive verbal form). *Mal-vivre* may be in a process of diffusion and come to be conjugated later, in the same way that it could simply disappear from use, if the trend comes to an end. All these words can be said to be productive in terms of use. However, they have not (yet) entered the dictionary, with the exception of *mal-logement* ("badhousing") which entered the dictionary in 2006. This fact corroborates the idea that compounding in *mal-* is productive and leads to lexicalizations.

Figure 29 is an analysis of the quantitative behaviour of these terms. It shows an increase in use of all innovative compound words in *mal-* and *mal*. The graph also shows that in September 1999, not only the word *malbouffe* appears, but the frequencies of other words in *mal-* and *mal* rocket. It could be that *malbouffe* triggered this phenomenon, as Barraud (2008) contends, or that the fate of *malbouffe* is part of a larger movement of productivity in *mal/mal-*. The two analyses are not mutually exclusive. The sinusoidal form of the polynomial curve suggests that the phenomena operate in cycles. The pattern reminds us of the "word-waves" observed by Clarke and Nerlich (1991).

---

154 *Same as in previous note for translation issues.* Malboulot : "badwork" ; mal-logés : "badly-housed" ; mal-logement : "bad-housing"; mal-vie : "bad-life"; mal-vivre : "bad-living"; malbonheur : "bad-happiness"; mal-développement : "bad-development"; mal-comprenant : "bad-understanding/er"; malformatives : "badformative"; mal-fondé : "badly-founded"; mal-croyants : "bad-believers"; mal-protégés : "badly-protected"; mal-informés : "badly-informed"; mal-mariés: "badly-married"; malfonctionnement: "badfunctionning"; mal-administration/ maladministration: "bad-administration/badadministration"; mal-emploi: "bad-employment/job"; mal-eBay: "bad-Ebay"; mal-penser: "to misthink"; mal-pensant: "misthinking"; mal-pensance/ malpensance: "mis-thinkingness/ misthinkingness"; mal-Bové: "bad-Bové".

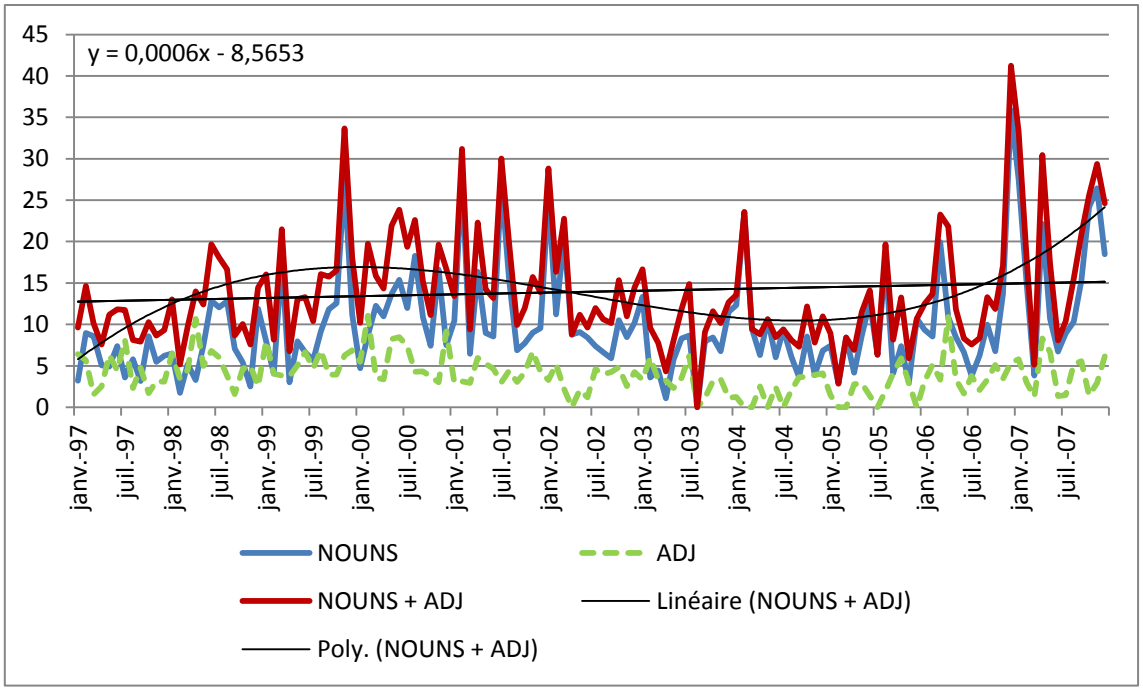


Figure 29 Normalized frequencies for all innovative nouns and adjectives in *mal-* in the corpus *Le Monde* (1997-2007), with linear regression and 3rd degree polynomial on the sum nouns+adj.

The full list of 172 words can be found in Appendix 1. The following selection, however, gives a sufficient overview of the phenomenon:

- mal-pensant
- mal-logés
- mal-logement(s) 2006
- mal-vie, mal-vivre
- malbonheur
- mal-développement
- mal-assis
- mal-comprenant
- malformative
- malpropreté
- mal-fondé
- malvisée
- mal-croyants
- mal-protégés
- mal-informés
- mal-mariés
- mal-lotis
- mal-perçu
- mal-dire
- malfonctionnement
- mal-classés
- mal-administration maladministration,
- malgouvernement
- mal-aimants
- mal-emploi
- mal-eBay
- malnutries
- mal-loti
- mal-pense, mal-penser
- mal-pensance, malpensance
- mal-entendant
- mal-mort
- mal-perceuses
- mal-boire
- mal-partis
- maltraitement
- mal-gouvernance
- mal-comptants
- mal-modernes
- mal-mutants
- mal-Bové

At the sociological and philosophical levels, one may wonder what the meaning of the increased productivity in *mal/mal-* is. One possible explanation is that this emerging vocabulary reflects a societal crisis, in which the traditional values (food, work, government, marriage…) are all of a sudden "badly" done, as if things were not done "properly". An interesting item is the word *malbonheur* ("badhappiness"). In French, the opposite of *bonheur* ("happiness") is *malheur* ("unhappiness"), thus the word *malbonheur* stands for something different. *Malbonheur* is not the opposite of *bonheur*, but rather an expression depicting a state in which *happiness* is pursued in a wrong or bad way. This seems to be a judgment of the way happiness (b*onheur*) is being searched for and represented. Therefore, a word like *malbonheur* shows that the conceptual definition of happiness is undergoing discussion.

A wider phenomenon of morphological productivity in *mal-* and *mal* also emerges. It raises several questions: what are the other morphologically productive patterns? How can they be detected without starting with a neologism in the first place? The fact the neologism is built on a familiar word also raises the question of whether this is an isolated phenomenon or part of a wider one, in which familiar words come to take on more importance in language. This would coincide with what regression coefficients showed: that language registers are becoming more familiar in the press. Moreover, the fact that *malbouffe* does not follow classical compounding rules raises other hypotheses. This could be explained by the influence of the English language in French, since English rules of word formation and creativity are more lax than French ones, and the French as well as the French media are increasingly in touch with the English language.

This case study shows the emergence of a neologism which is subjected to semantic change, since the original meaning is enriched via the context of use. In the current context of an "information society" in which communication media are getting quicker and denser (as described in Part I, Chapter 2 of this work), the hypothesis that this phenomenon is spreading and losing its status of exception can be put forth.

3.2.3.2. Morphological productivity, neology and semantic change: *crypto-* and *cyber-*

A press corpus such as *Le Monde*, in French, is undoubtedly a mine of innovations, neologisms and free-forms. However, a great number of so-called "neologisms" are only spontaneous innovations that may never spread, or stay confined to an extremely limited set

of users. Some of them are one-shot creations. But even if they do not spread as such, they have a semantic function when they are part of a more global phenomenon. As seen above, a word like *mal-mutant* may be treated as the result of personal fantasy on its own, but when looking at it as an instance of composition in *mal-*, within a set of 172 words, it acquires a different status. It shows the creativity associated with the prefix or compound, and how productive the creativity is.

The linguistic creativity observed in the press is developed enough to extract phenomena and yet it is constrained by clear stylistic boundaries. The press is therefore a very adequate place to evaluate these mechanisms. In an article from January 1999 in the corpus (excerpt in Appendix 2), *Le Monde* reports on a lexicographic study about its own tendency to have generated neologisms in 1998. Feminization, creation of adverbs, thematic creations (like football-related words) and prefix productivity are dealt with, mentioning the productivity of auto-, eco-, euro-,bio-, dé and cyber (sic.):

" Eco - comme d'ailleurs euro -, se met à toutes les sauces : écotourisme, écoguerrier, écotaxe, écoconseiller... L'évolution des techniques fait écrire biofibre, bionique, biométrique, mais aussi biojeu, bioterrorisme, biovigilance, bioprospection. Très prisé également dans *Le Monde*, le préfixe dé, qui semble illustrer un délitement général (déliaison, déprotection, décivilisation, déspectacularisation, désintermédiation). Quant à cyber, il n'a sans doute pas dit son dernier mot, après cybercitoyen, cybercriminalité et cybernétisation"[155]

[155] "New words are composed with fashionable prefixes, such as auto (*autodénigrement, autogénocide, autoputsch, autocongratulation, autofiction*, or even –in a more obscure way- *autopathographie*). Eco –in the same way as euro- is adapted for any purpose: *écotourisme, écoguerrier, écotaxe, écoconseiller...* The evolution of techniques makes people write *biofibre, bionique, biométrique*, as well as *biojeu, bioterrorisme, biovigilance, bioprospection.* The prefix dé is also very popular in Le Monde, and seems to illustrate a general splintering *(déliaison, déprotection, décivilisation, déspectacularisation, désintermédiation).* As for cyber, it has not yet said its last word, after *cybercitoyen, cybercriminalitéet cybernétisation."*

The phenomenon is therefore acknowledged by the newspaper itself. One of the processes the article lists is morphological productivity. Morphological productivity is interesting because it is an element of natural change in language evolution. Prefixes, suffixes and elements of compositions are expected to be productive, but at what rate? An element of composition like *non-* for instance is stable in the corpus (the regression coefficient is 0,08 for all compositions in *non-*). As elements of composition tend to condensate whole complex concepts, on the model of an element like *philo-* for instance, the rate of productivity they display is an indication of the fact that the concept is spreading or remaining stable. When it spreads, it is via a growing a number of combinations and the changes in the nature of these combinations. Moreover, if a known element which has had for some time a limited set of possible combinations, comes to acquire more creative combinations, or if one or several new meanings make it more polyvalent, the concept is also evolving. The degree of linguistic constraint put on word composition and use partly defines their semantic profile. As discussed subsequently, words built in *bio-* can mean several things. Wondering what "biogays" and "biopunks" are, one could come up with several plausible hypotheses. However, some elements seem to have no active plasticity. They have several meanings but the way these meanings are related to each other and organized has not changed an iota for decades or centuries, like *philo-*.

To observe the tendency for neologisms and creations based on the morphological productivity of prefixes and compounds, it is useful to take into account words with an extremely low frequency profile. The total number of forms gives a scale of the word's general productivity. These words benefit from being analyzed in sets, since their values (frequency, etc.), taken separately, do not constitute a basis for interpretation (statistically as well as semantically). Words that appear once, twice or thrice have something to say when they are part of a substantial set. This set may be viewed as a "cloud" of creativity which accompanies semantic change. In most NLP approaches, these items are discarded to "clean" the data, because this frequency range also contains the bulk of typographic errors and idiosyncrasies. Indeed, classifying manually such items is a long task, and no semi-automatic program is able to do it with our current means. To explore what this frequency range has to say, I manually sorted these elements in three case studies: *cyber-*, *crypto-* and *bio-*. What is at stake is the observation of the mechanisms involved in productivity. At this stage, the computational approach is important to extract large amounts of data (hundreds of candidates

for a single element of composition). Since these items are not recognized by the TreeTagger program, the figures are taken from the unstemmed version of the corpus *Le Monde* for *cyber-* and *crypto-*.

Words which are not attested in the dictionary are manually selected, and their use is manually checked over the Internet to evaluate their online use to complete the data. Words that have no online use are generally classified as idiosyncrasies. Since the frequencies are very low, the percentage of elements that will remain active is not possible to determine with our current means. However, it would be interesting to test these elements in the future to assess whether they survived or whether they were part of an ephemeral trend. As seen with *malbouffe*, the creation, entry in the dictionary and quick integration in journalistic style of *mal-logement* sets a path for similar forms to emerge.

A series of neologisms appear along with free-forms. Some of them can be said to be idiosyncratic innovations, since they seem to be related to a unique person. However, some of them may encounter success, diffusion and become lexicalized. The fact that some do not follow the expected rules in their formation makes them interesting candidates.

The formal neologies being morphological productions, two questions of semantic nature are raised: whether the forms undergo semantic shift and whether the elements of composition also do. As stated in Part I, specialized terminology is a constant source of neologisms, depending on the mainstream visibility of the domain. For instance, in medical terminology, when an illness is highly reported on in the media, words to describe it become common use. The most productive domain, at the time of writing, is technology. This commonly accepted statement is confirmed by statistical data as well as trends. For instance, in 2012 the OAD[156] consecrates as "word of the year" the term *gif* (and abbreviation of "graphic interchange format") an image and video format widely used on the Internet.[157]

[156] Oxford American Dictionary

[157] See the articles on this topic, in French: http://bigbrowser.blog.lemonde.fr/2012/11/14/lol-gif-elu-mot-de-lannee-par-le-dictionnaire-oxford/; in English http://www.huffingtonpost.com/2012/11/12/word-of-the-year-gif-oxford_n_2119349.html. Both last accessed on 14/11/2012.

Several words written with a hyphen or in two words follow the same fate as *malbouffe*: instead of two separate words they become one, or the two spellings coexist. Words that are highly used online by users and the independent press are neologisms in the process of diffusion. It does not mean they will eventually enter the dictionary but they have a chance to do so. If their process of formation includes another index, as the shift from a two word expression to a unique hyphened one, then they are even more likely to be on the path of integrating mainstream language. Once they appear in the press corpus, they are generally already used online. Most of the neologisms and free-forms are found online, some with an extremely widespread use and definitions offered by independent sites or users, and some with very few hits. The number of hits in Google is sometimes given as an indication of use, along with the number of occurrences in the corpus and the part-of-speech for each selected word. Google automatically offers "corrections" for the words which exist in several forms: with or without a hyphen, or as a collocation (here as strings of two words). For words that have too few hits, Google does not offer this distinction.

3.2.3.2.1. ***Crypto-: domain shift***

C*rypto*- comes from Gr. *kruptos* ("hidden") and allows compounding in French, but it might also be the case that French compounds are borrowed from English, which has a richer attested productivity in *crypto-*. *Crypto*- (the element of composition) gave way to *crypto* (the noun) in the 1950s allowing to build two word expressions, which sometimes become one word expressions, and can include a hyphen on the model of *crypto-*.

There are 90 units in *crypto*- in the corpus[158], after clearing errors and merging plurals and singulars on a total of 112 units. All words are presented in Table 16, classified by frequency range and domain. 19 words are attested among which two are used with an incorrect spelling: *crypto-communisme* and *crypto-communiste* (attested without a hyphen) and 70 are unattested[159] (above 77 %). These attested words seem to provide a model for the new creations. Most of these creations do not belong to the domains of biology and cryptology which are the most attested. They are rather used in political, social and religious compounds

[158] This study was conducted on the unstemmed version of the corpus to make sure all low frequency words were extracted properly.

[159] The difference of one unit in the table is due to the plural use of *crypto-communistes* in the low range.

in the sense "hidden" on the model of *crypto-communisme*. If only two words are attested in French on this model, there are more of them in English (10 attested). Composition in *crypto-* has been gaining momentum in political critique, on the model of *crypto-communist* (1924) and *crypto-fascist* (1927) in English[160] and *cryptocapitalisme* and *cryptocommunisme* in French, which find their way into dictionaries much later (1960)[161]. The bulk of creations with extremely low frequencies (under and equal 3) seem to rely on this model: *crypto-fasciste, crypto-socialiste, crypto-marxiste, crypto-vert* ("crypto-green" referring to the ecological party*),* etc. A total of 46 items follow this model, among which only 2 have frequencies higher than 3 over the whole corpus. All of them are unattested at the exception of *cryptocommuniste* and *crytocommunisme*.

For instance, the term *cryptojuif* exists online and is defined by Wikipedia as the secret adherence to and practice of Judaism while overtly practicing another religion. Under this spelling it has close to no hits in Google (1510), but under the spelling *crypto juif*, we find 86 200 hits. In the corpus, the term is used in the same article in 2007. The term can be created in French, but an equivalent word is attested in English by the OED under the spelling *crypto-Jew*, and its structure is echoed in similar terms based on religions such as *crypto-Christian*, and *crypto-Catholic*. The French equivalents *crypto-catholique* and *crypto-chrétiens* also appear in the selection along with creations such as *crypto-taoïsme.*

However, *crypto-* has long been around in natural science, in words like Fr. *cryptobiose* ("cryptobiosis") referring to living organisms that do not show outward signs of life. The meaning "hidden" is transposed from the predominant fields of natural sciences and didactics to politics, religion and social classes in general (see *crypto-gay, crypto-gothique*). In natural sciences, the compositions can be said to be confixations, defined earlier as compounds based on elements of composition from Greek and Latin origin in specialized domains. However, as we move to politics, religion and social classes, one may wonder whether the compositions are confixations or if they emulate the confixation mechanism used in the natural sciences and didactics. In English, most natural science compounds in *crypto-* are rare and obsolete (e.g. *cryptocarp, cryptocephalous*).The shift from the domain of the natural sciences to politics is

[160] Source: OED

[161] Source: GRLF

happening cross-linguistically. Either the French are borrowing the expressions directly, or their mode of composition is replicated in French independently. When an element of composition adds a new domain to its meaning as an effect of the nature of the previous compositions it has generated, it can be said to undergo widening, a type of semantic change. *Crypto-* can combine with biology, religion and politics, and the last two are highly predominant in the selection. Moreover, a few free creations appear, such as *cryptoromantique* ("cryptoromantic"), in which *crypto-* is freely used as "hidden". Therefore, in low frequency ranges, we find more words and more unattested words than in the high frequencies, confirming that low frequency ranges are a good locus for creativity and productivity.

	Total frequency range				total
	< and =3		>3 and < 50	> 50	
biology	cryptomère	cryptoclidus	crypto		19
	cryptoméria	cryptobranchus	cryptogamique		
	cryptobiose	cryptogame	cryptogamie		
	cryptologique	cryptozoologiste	cryptographe		
	cryptonyme	cryptosporidium	cryptozoologie		
	cryptobranchidé	cryptozoologue	cryptochromes		
	cryptochrome				
cryptography	cryptonumérique	cryptographe	cryptographique	cryptologie	12
	cryptoprocesseur	cryptologues	cryptogramme	cryptographie	
	cryptosystème	cryptography	cryptologue		
			cryptome		
politics, religion, nationalities and social classes	cryptocagoulard	cryptocommuniste			46
	crypto-maréchaliste	cryptocommunisme	crypto-communiste		
	crypto-polynésien	crypto-communisme	crypto-marxiste		
	crypto-judaïque	crypto-communistes			
	archéo-crypto-socialiste	cryptojuif			
	crypto-biblique	crypto-catholique			
	crypto-républicain	crypto-chrétien			
	crypto-vert	cryptofasciste			
	crypto-socialiste	crypto-fasciste			
	crypto-thatchérien	crypto-fascisme			
	crypto-nazi	cryptofascisme			
	crypto-boyard	crypto-raciste			
	crypto-puccinienne	crypto-lepéniste			
	crypto-judaïsant	crypto-rocardien			
	crypto-taoïsme	crypto-gay			
	crypto-révisionnisme	crypto-candidat			
	crypto-penseur	crypto-gothique			
	crypto-soutien	cryptocollectiviste			
	cryptomarxiste	cryptoriches			
	crypto-marxiste	crypto-expressionniste			
	crypto-totalitaire	crypto-irakien			
		crypto-militantisme			
other	crypto-autobiographique	cryptoportique	cryptonline		14
	crypto-armée	cryptophonos			
	cryptoromantique	cryptonomicon			
	cryptogénique	cryptopuzzle			
	cryptomnésique	cryptobio			
	crypto-volcanisme	crypto-ésotérique			
	cryptobiographique				
total	75		13	2	90

Table 16 All words in *crypto-* classified according to frequency range and domain in the corpus *Le Monde* (1997-2007), in unstemmed version. Words in red are attested, words in green are attested under a different spelling (without a hyphen).

3.2.3.2.2. ***Cyber-***

Cyber- enters the French dictionary in 1945 and is taken from *cybernétique*, itself calqued on English *cybernetics*. *Cybernetics* is of Greek origin, from *kubernêtiké*, meaning "science of government". *Cyber-* denotes the regulation of human and machine activities. In the corpus, the highest figures are for attested words, such as *cyberspace* (365 occ.) *cybercafé(s)* (287), *cybercriminalité* (121) *cyberculture* (84) or *cybermonde* (59)[162].

Figure 30 shows added normalized frequencies for all (unsorted) words in *cyber-*. The concentration of creativity and use in the period from 1997 to the end of 2002, shows that words in *cyber-* follow a trend.

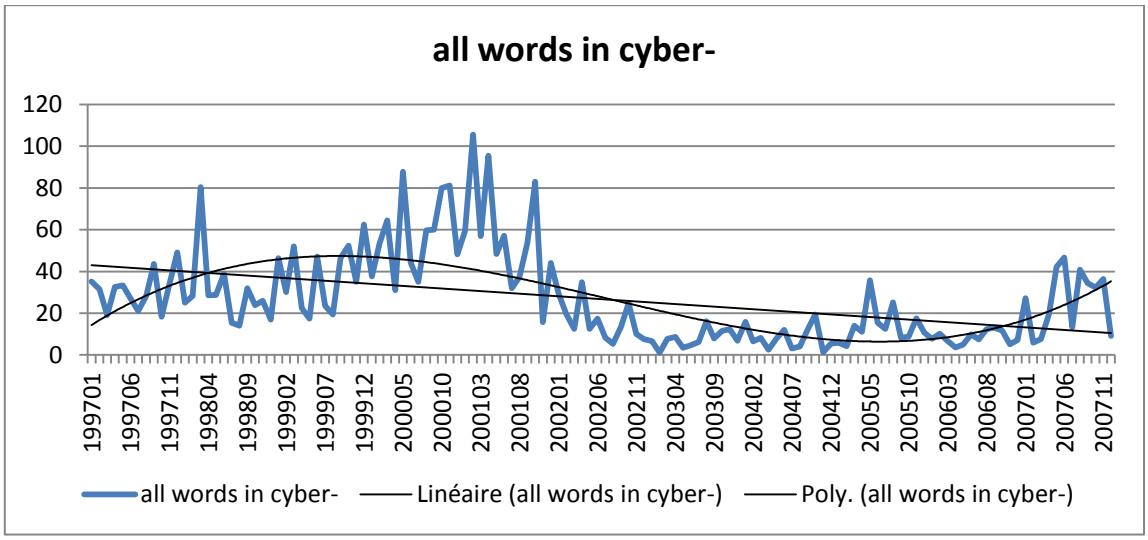


Figure 30 Added normalized frequencies for all 913 (unsorted) words in *cyber-* in *Le Monde* (1997-2007), in unstemmed version, with linear regression and 3rd degree polynomial .

All attested words are recent words having entered the dictionary in the 1990s, except *cybernétique* (1945), including for instance, *cybernaute* (1995), *cyberculture* (1995) and *cyberguerre* (1993; "cyberwar").

As newly attested forms in *cyber-* become part of everyday life, free-form creations are stimulated. In fact, it is the whole set of words in *cyber-* that should be considered and not just a set of attested words versus a set of neologisms and free-forms. The dynamics are within

[162] Cyberspace, cybercafé, cybercriminality, cyberculture and cyberworld*

that whole set between words that are used more and/or lexicalized and creations appearing in that context, the first process facilitating the second.

In the high frequency range 7 words out of 9 are attested such as *cybercafé* (1994) and *cyberspace* (1995). *Cyberpunk* and *cybermarchand* ("cybertrader") are not attested even though they are widely used, the first as the name for a whole movement which includes postmodern science-fiction.

In the middle frequency range (>3 and < 50) there are 8 attested words (7 after sorting out plurals) and 85 unattested words that come down to 81 after sorting plurals and singulars. Words like *cyber-sexe* and *cyberterrorisme* sound acceptable to most people however they are not attested.

The full list of words in these two ranges can be found in Appendix 3. In the very low frequency range (<=3), there is a plethora of free-formations (768 items).

To check whether free-formation is still productive in the low phase observed for all words, the frequencies of words in the lowest range are added in Figure 31.

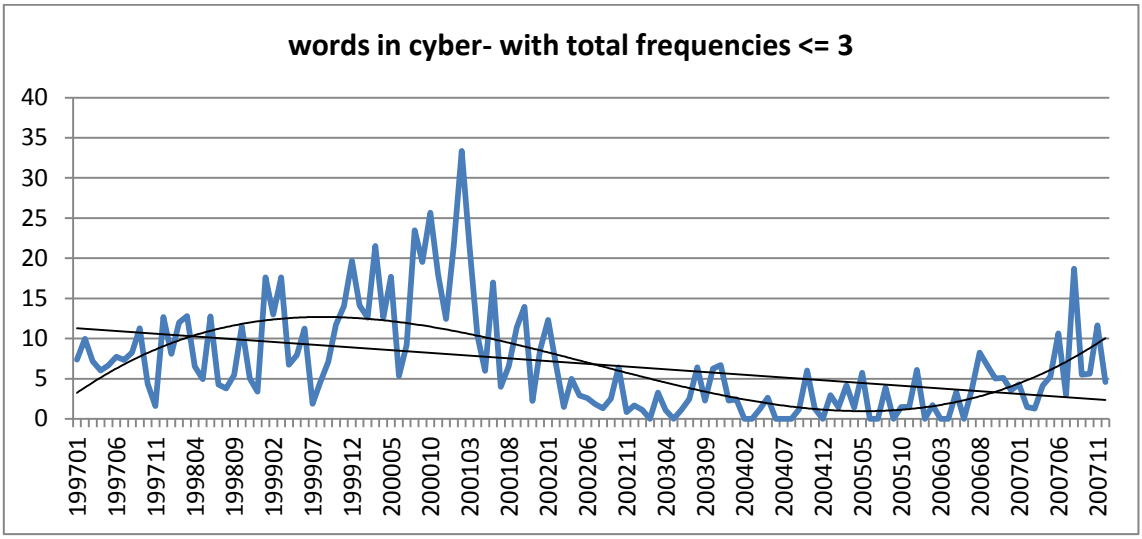


Figure 31 Added normalized frequencies of words in *cyber-* with a total raw frequency under or equal to 3 in the corpus *Le Monde* (1997-2007), in unstemmed version, with linear regression and 3rd degree polynomial.

The graph shows that even in very low frequencies where creativity unfolds, the use of *cyber-* follows a trend. This is replicated in the middle frequency range and the graph shows the same pattern, therefore confirming that the trend applies to all frequency ranges.

Non-attested words show relatively high frequencies within the middle range, for instance *cybermarchand* (44), *cyberconsommateur* (38), *cybersurveillance* (34) *cybercrime* (32) *cybermarché* (23) *cybercommerce* (21), *cyberdémocratie* (21) *cyberterrorisme* (21) and *cybertattaques* (19)[163].

Creation in *cyber-* is still productive and numerous compounds seem to be here to stay. For instance, the word *cyberattaque* ("cyberattack"), despite having only 26 occurrences from 1997 to 2007 (plural and hyphened forms included) appears in 181 articles in *Le Monde* from 2008 to today[164] (131 in the hyphened form, 50 in the non-hyphened one). Among innovative forms, some words point to new social referents, such as the composition *cyberpapy* ("cybergrandpa", 9 occ. in the singular, 6 in the plural), and the emerging *cybercitoyen* ("cybercitizen") 4 occ. in the singular, 8 in the plural). We may be facing the first generation of cybergrandpas and cybercitizens. However, it is questionable whether we will still need to name them this way now and in the future, when most grandpas will be cybergrandpas and most citizens, cybercitizens.

Renouf (2007) finds comparable results in English newspapers while observing a more complete set of elements of composition and prefixes:

	89	**90**	**91**	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**00**	**01**	**02**	**03**	**04**	**05**
Cyber	8	9	8	9	8	5	3	2	2	2	3	2	3	4	4	5	5
Euro	1	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	4
Mid	2	3	3	3	4	4	3	3	3	4	4	4	5	5	5	5	5
Techno	6	7	6	6	5	5	3	3	4	4	5	4	5	5	7	6	6
Under	2	2	3	3	3	3	3	3	3	3	4	4	4	4	5	4	5
Dis	3	4	5	4	5	5	5	5	6	6	6	5	6	6	6	7	6
Inter	2	3	3	4	4	4	4	4	4	5	5	5	5	5	5	5	6
Mis	2	3	4	4	5	5	5	5	5	5	6	5	5	6	5	5	6
Mock	4	4	5	5	5	5	5	4	5	5	5	5	5	6	6	6	7
Faux	9	9	8	9	7	8	6	5	6	6	5	6	7	6	5	5	5
Trans	4	6	5	6	5	5	6	5	5	7	7	6	6	7	6	6	8
Poly	4	5	5	5	6	7	6	6	6	6	7	7	9	7	7	7	7
Uber	10	10	10	9	9	9	7	7	6	6	5	7	7	7	6	6	5
Fore	6	7	7	7	8	8	9	7	8	7	8	8	8	8	9	7	9
Vice	6	7	8	8	9	7	9	9	9	8	9	9	9	10	8	9	9

Table 17 Prefix productivity in UK broadsheet newspapers from 1989 until the end of 2005, taken from Renouf (2007: 3)

[163] Cybertrader, cyberconsummer, cyberwatch, cybercrime, cybermarket, cybertrade, cyberdemocracy, cyberterrorism and cyberattaks

[164] Number extracted on the 17th August 2012.

She notes "a noticeable rise for the vogue items *cyber(-), faux(-)* and *uber(-).*" (Renouf 2007:65).

3.2.3.3. Semantic change, polysemy and ambiguity of *bio-* : natural vs. artificial life

"BIO- : Élément tiré du grec *bios* « vie », qui entre dans la composition de nombreux mots savants, dès le XVIII[e] s. (➞ aussi -bie). Les composés récents (noms et adjectifs) sont didactiques et servent généralement à désigner le rapport entre une science, une technique et la biologie*. — Ex. : bioastronautique [bjoastʀonotik] n. f. (1966) ; biocybernétique [bjosibɛʀnetik] n. f. (1964) ; biopolitique [bjopɔlitik] n. f. (1969) ; biospéléologie [bjospeleɔlɔʒi] n. f. (1964) ; biosocial [bjosɔsjal] adj. (1965) ; biosystématique [bjosistematik] n. f. (1964) ; biothéologie [bjoteɔlɔʒi] n. f. (1943). D'autres composés désignent des phénomènes et des objets propres à la biologie : *bioconversion* [bjokɔvɛʀsjɔ] n. f. (→ Biocatalyse, cit.) ; *biosmose* [bjosmoz] n. f. (1958) ; *biotactisme* [bjotaktism] n. m. (1970) ; *bioplaste* [bjoplast] n. m. (1965)."[165]

Since *bio-* comes from Gr. *bios* ("life"), it is first used in older compounds. In more recent compounds, words like *biocybernétique* point to the relationship between biology and a science, and words like *biosmose* refer to concepts specific to biology. *Bio-* compounds related to the relationship of biology, sciences and techniques are recent according to the GRLF.

The full list of words starting in *bio-* was extracted (785 units in stemmed version),

---

[165] Source : GRLF. The equivalent entry in the OED is more up to date: < post-classical Latin bio- < ancient Greek βίο- , combining form (in e.g. βιόδωρος life-giving) of βίος life, course or way of living (as distinct fromζωή 'animal life, organic life'), probably ultimately < the same Indo-European base as quick adj. Found in borrowings and adaptations of Latin words from the 17th cent. onwards (earliest in biotic adj., and in biographer n., biographist n., biographical adj., biography n., and related words). Formations within English are found from the early 19th cent. (compare biometer n., bioscope n., etc.). Compare French bio- , German bio- (formations in both of which are found from the early 19th cent. or earlier).The ancient Greek suffix -βίος 'having a specified manner of life' is also represented in the ultimate etymology of a number of scientific words which entered English chiefly via scientific Latin forms in -bius , -bia , or -bium (or via corresponding forms in French ending in -bie or -be or in German ending in -bie or -be); English words so derived hence show a variety of different endings, none of which has given rise to a productive suffix in English, although some analogous formations are apparently found. **1.** With the sense 'biographical', 'comprising elements of biography'. **2.** a. Forming temporary and nonce words relating to life and living organisms (real and fictional), and (in later use) to biotechnology or environmental sustainability.b. In more established (chiefly scientific) words with the sense 'biological', 'concerning organic substances, life, or life processes', 'concerning biotechnology or sustainability of the environment', as bioclimatic, biomanufacturing, bionanotechnology, bioresearch, biorobotics, etc.

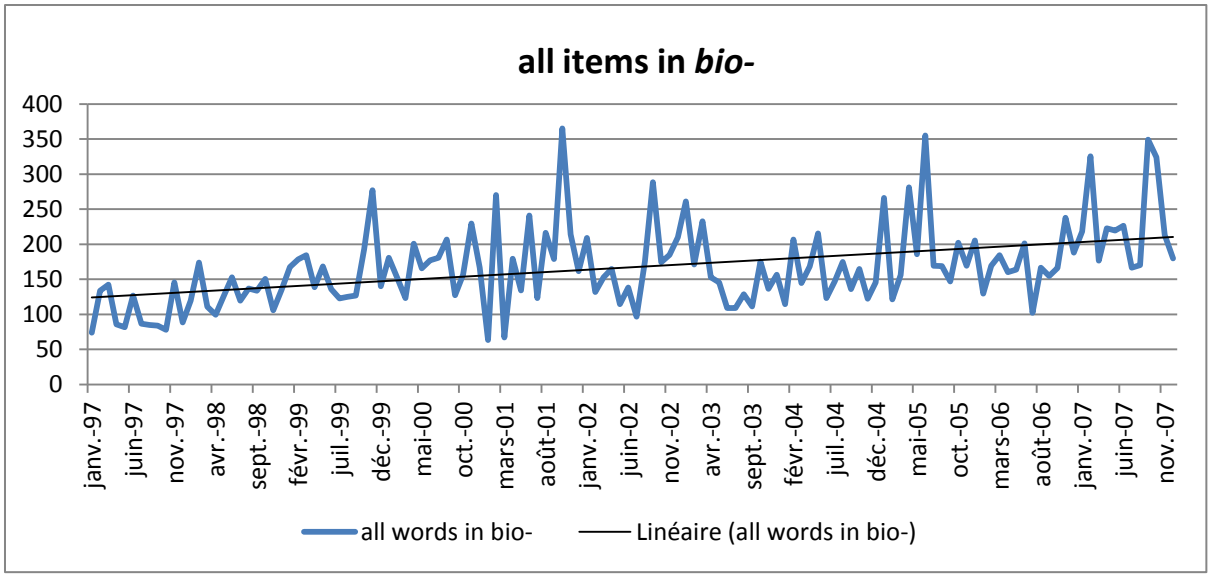


Figure 32 Added normalized frequencies with linear regression for all 785 units in *bio-* in *Le Monde* (1997-2007), in stemmed version.

Words were manually sorted. Out of the 785 units, 304 are unattested or recently attested words (375 units in total due to the Tagger's mistakes while encountering new words, and 304 after corrections). This count includes unattested words as well as 16 recently attested words (since 1975) which participate in the dynamics of new creations. These figures attest that *bio-* is productive.

attested word	Date of entry according to the GRLF
NOM_bio-diesel	1992
NOM_biopuce	1997
NOM_bioscience	1982
NOM_biocatalyse	1979 >biocatalyseur
NOM_biocompatibilité	1980
NOM_biodesign	1987
NOM_bioénergie	1975 > Eng. "bioenergy"
NOM_biomatériau	1982
NOM_bioréacteur	1982
NOM_biorythme	1972 > Eng. "biorhythm"
NOM_bioterroriste	1998
ADJ_biopesticide	1988
NOM_biodynamie	1976
NOM_bioéthanol	1987
NOM_biosécurité	1990
NOM_biotechnologie	1980
NOM_bioindustrie	before 1979 built on > biotechnologie

Table 18 Recently attested words in *bio-*, since 1975, kept in the selection.

Words were then manually classified by semantic field. The full list of words by category can be found in Appendix 4. Items were manually classified in the following categories:

- Life sciences /biology/from *bios*, life 220 words
- Ecological (made from organic substance/biodegradable) 33 words
- Organic 24 words
- >Biotechnology 14 words
- Ambiguous 6 words
- Other (idiosyncrasies and meanings related to biography and biosphere) 7 words

The attested definitions of *bio*- 1.) from *bios*, life, 2.) the relationship of biology with sciences and techniques and 3.) related to biology, were gathered under the heading "life sciences/biology/from *bios*". The other categories were created in terms of the data. The heading "ecological" groups words in which *bio*- means "made from organic substance" and/or "biodegradable". Items connected to "biography" and "biosphere" were discarded.

The hypothesis is that the high production of neologisms in *bio* - (1) , (2) (3) in the fields of biology, medicine and life sciences in general may have an impact on the production of words in *bio*- with other unattested meanings.

The categories "life/science", and "biotechnology" mostly generates authorized morphological production. However, the categories "ecological" and "organic" generate unauthorized and hybrid compositions.

3.2.3.3.1. ***bio- as organic: creative compounding***

The meaning "organic" is taken from the element *bio*- as an (unattested) element of composition taken from the adjective *biologique(4). Biologique* has five meanings:

> "1. Related to biology (science).
> 2 Pertaining to life and living organisms
> 3 Characterized by life. Biological beings
> 4 (Non-scientific use). Of spontaneous and natural life. ➞ Ecological . Organic Farming. - Organic and natural products, which are manufactured or grown without artificial and chemical substances.

Abbr. fam. : Bio [bjo]. Organic farming. Organic products. - Adv. Eat organic.

5 N. m. The biological: all phenomena specific to life"* [166]

The category "organic" was created according to the data. In high frequencies, words with this meaning are the attested *biodynamie* referring to the biodynamic school of thought, *bioproduit* ("bioproduct") and *biodéchet* ("biowaste"). In very low frequencies (under 3), 20 items correspond to this meaning, for instance *bioscosmétique* ("biocosmetics"), *bio-équitable* ("bio-fair"), *bio-attitude* (idem), and *biorelaxant* ("biorelaxing").

In French, there seems to be four definitions of *bio-*, three are attested: from *bios*, life, and, by extension related to biology and sciences, or restricted to biology, and the fourth, the abbreviation of the adjective *biologique* meaning natural and organic, is not attested as an element of composition. To simplify, I summarize it as:

Bio-

(1) from *bios*, life
(2)) the relationship of biology with sciences and techniques.
(3) related to biology
(4) From the adj. *biologique*: organic and natural products

In other items, the idea of a natural product is found. However, it is not in the sense that they are organic products, obtained via organic agriculture, but in the sense that the products are made from (undefined) natural and organic source instead of being made from synthetic or chemical source. This is the case for the idea of bio-petrol for instance.

---

[166] Source: GRLF. Original definition in French :
« 1 Relatif à la biologie (science).
2 Qui a rapport à la vie, aux organismes
3 Qui est caractérisé par la vie. Les êtres biologiques
4 (Emploi non scientifique). De la vie spontanée, naturelle. → Écologique. Agriculture biologique. — Produits biologiques, naturels, fabriqués ou cultivés sans substances chimiques artificielles.
Abrév. fam. : bio [bjo]. L'agriculture bio. Produits bio (ou *bios*). — Adv. Manger bio.
5 N. m. Le biologique : l'ensemble des phénomènes spécifiques de la vie. »

3.2.3.3.2. ***Ecological petrol : synonymic competition***

In the range of highest frequencies (over 50 in total), there are two words, the recently attested word *biodiesel* (1992) along with the unattested *biogaz.* However, the attested spelling of *biodiesel* is *bio-diesel*. In middle frequencies (3-50 in total) the attested *bio-carburant,* (1977) as well as *bioéthanol* attested earlier than *biodiesel*, are also a reference to bio-petrol. In the range of extremely low frequencies (3 and under 3 in total) as series of unattested forms also denote biopetrol: *bioessence, biofioul, biofuel, biométhane, bio-combustible*. There seems to be a process of synonymic competition for this notion even though attested forms have already entered the lexicon.

The meaning of *bio-* here can be summarized as "made from organic substance and/or biodegradable to some extent". At least, *bio-* products are regarded as more biodegradable than non-*bio-* products. This meaning also includes the notion of respect of the environment, since the motivation for using *bio-* products is ecological. Therefore, in this context *bio-* means "made from a natural substance" and "respectful of nature"

The noun *bioplastique* (with a high frequency) relies on this meaning as well as *biopolymère* (high frequency). *Bioplastique* stands for ecological plastic in the corpus, and a search online shows the expression is widely used as such.

3.2.3.3.3. ***bio-plastique: a case of polygenesis?***

Bio-plastique raises a few hypotheses. The equivalent English term "bioplastic" is defined as follows by Wikipedia:

> "Bioplastics are a form of plastics derived from renewable biomass sources, such as vegetable fats and oils, corn starch, pea starch, or microbiota,"

And as follows by the OED:

> "Biosplastic, n.
>
> **Etymology:** < bio- comb. form + plastic n.

1. A type of transparent plastic in which biological and palaeontological specimens can be embedded for the purposes of preservation, display, or manipulation. Freq. *attrib*[167].

2. Any of various biodegradable plastics derived from biological substances (as opposed to petroleum). Freq. *attrib.*"

However, English possesses a rare homonym adjective:

"Bioplastic, adj.

Etymology: < bioplast n. + -ic suffix.

Biol. Now *rare*.

Of, relating to, or concerned with bioplasts."

"Bioplast,n.

Biol. Now *hist.*

In the terminology of L. S. Beale: a unit of bioplasm as an independently existing entity capable of growth and reproduction; the living part of a cell. Cf. bioblast n."

In French *bio-plastique* is found online without a hyphen and in two words, with the same definition. The meaning "taken from biodegradable substance" is not contained in the definition of *bio-* in the GRLF.

However, similarly as in English, there is an attested adjective *bioplastique* which was formed on the basis of the original meanings of *bio-* (as in "related to life" and "related to biology") and has long been attested in biology as an adjective referring to the property of living cells to regenerate (1896):

« bioplastique [bjoplastik] adj.

ÉTYM. 1896, cit.; de *bio-*, et *plastique*.

Biol. Se dit de la propriété qu'ont les cellules vivantes de se régénérer. | « une accélération des processus bio-plastiques morphologiques » (*Année sc. et industr.*, 1897, p. 246; 1896). »[168]

[167] Frequently attribute.

However, the English term is built as "bioplast+ic" whereas the Fr. term is built as "bio+plastique". Meanings correspond to each other. Therefore, both English and French have the same homonyms: *bioplastique*, noun and adjective and *bioplastic* noun and adjective. However, in English there is a supplementary meaning in *bioplastic*, n. 1, as a type of plastic used in biological preservation, which is not echoed in French.

Could the Fr. adjective *bioplastique,* based on *bio-* (as in "related to life" and "related to biology"), influence the creation of the noun *bio-plastique*, based on the meaning "made from organic substance/ecological"? The central concept in common is the one of living organism, itself contained in *bio-* as *bios ("life")*. The second is regeneration. *Bioplastics* are –supposedly - made out of natural organisms[169] and are easier to recycle. In this sense, the idea of regeneration is kept. But the feature of "life" is partially lost since corn reduced to plastic form is not exactly alive, and it certainly cannot regenerate itself, except through transposition, where regeneration is possible to some extent via recycling. However far-fetched the transposition of meanings in common, these are contained in the etymology of the elements composing these words. It is difficult to say whether these two words may have been in contact at all in terms of time scale and domains. The adjective dates back to 1896 whereas the noun only emerges a century later, in the late nineties (first occurrence in the corpus: 1998). The 1896 meaning belongs to the domain of biology, the 1998 one to the domain of ecology. It is therefore likely that the new term has been formed independently, in complete blindness as to the pre-existence of the adjective. This process corresponds to what Geeraerts (1997) calls polygenesis, defined as the multiple and disconnected creations of meanings for homonymous words. This raises the question of the polysemy and possible ambiguity of *bio-*. Indeed, these two homonyms coexist, and while the attested adjective is almost in disuse, the unattested noun is in use. The meaning of *bio-* in both is clearly different: the adjective relies on the notion of "life" while the noun relies on the notions of "natural and ecological". This meaning is echoed in other compositions, such as *biopesticide* (attested in 1988, pesticides based on natural products respectful of the environment), *bioprocédé* ("bioprocess") and

---

[169] …or they are claimed to be so and perceived as such. Currently most of them are hybrid mixes of bioplastics and petroleum derived common plastics.

bioconstruction (“biobuilding”). There is therefore a fifth meaning to add to the definition of *bio-* encountered in the corpus:

(1) from *bios*, life
(2) the relationship of biology with sciences and techniques.
(3) related to biology
(4) From the adj. *biologique*: organic and natural products
(5) Made from organic and/or natural substance, biodegradable to some extent. By ext: ecological

Meanings (4) and (5), “organic” and “ecological”, are unattested for word-formation in *bio-* and share a common feature: nature and the natural. To observe the frequency tendency of words with these meanings (more often comprised in very low frequency ranges), their frequencies are added in Figure 33, showing an increase over time. The production of new meanings of *bio-* , for the notion of “natural and ecological” therefore seems active.

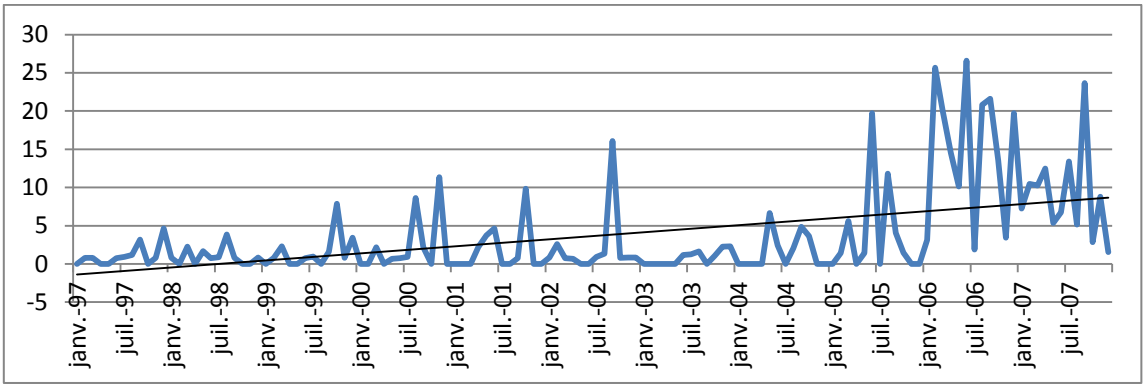


Figure 33 Added normalized frequencies of new words in *bio-* meaning "ecological" and "organic" in the corpus *Le Monde* (1997-2007) in stemmed version, with linear regression.

3.2.3.3.4. ***Biotechnology, a meaning extension?***

14 words are built on *bio-* taken from *biotechnologie* (high frequency), itself taken from *bio-* in biology and *bios-*, “life” and from the relationship between biology and technology (meanings (1), (2)and (3)). I hypothesize that it is an extension built from meaning (2) : “the relationship of biology with sciences and techniques”. This feature characterizes items *bio-* with low frequencies (apart from the abbreviations *biotech*) among which some are directly built on a prefixed form of *biotechnologie*. Examples include *bioindustrie* (attested with a

hyphen as a hybrid word in 1979 to denote the industry based on biotechnologies), *bio-techno, biopunk* and *bionumérique*.

The word *biopunk* is the same in En. and Fr. In English, it seems to be a portmanteau word based on *biotechnology* and *punk*. Biopunks are defined online and define themselves "as biohackers experimenting with DNA and other aspects of genetics."[170] The French version of Wikipedia describes it as a neologism base on *cyberpunk* and *biotechnologie*. *Techno-* disappears from the compounding process, but its semantics are retained in *bio-*. In fact, *biopunks* are semantically *biotechnopunks*. Therefore, in both hypotheses, either *cyber-* and/or *techno-* disappear from the word; however their meanings have been passed on. The meaning "technology" contained in *biotechnologie* seems to have integrated and contaminated the element *bio* -to some extent.

Bioindustry uses *bio-* resources, and transforms them via biotechnology. One might ask what exactly a *bio-* resource is, in this context. In the same way that words in *bio-* (2) refer to the relationship biology has with other fields in sciences, *bio-* (6) seems to do the same for the relationship biotechnology has with other sciences. This is exemplified in the very low frequency term *bionumérique* ("biodigital").

There is therefore a sixth definition in addition to the five meanings listed so far, leading to the following proposition of definition for *bio-:*

[170] http://en.wikipedia.org/wiki/Biopunk last accessed 15/11/2012. Reliability is unsure.
http://www.biopunk.org/what-is-biopunk-t37.html last accessed 15/11/2012

Bio-
(1) from *bios,* life
(2) relationship of biology with sciences and techniques
(3) related to biology
(4) From the adj. *biologique*: organic and natural products. By ext: ecological
(5) Made from organic and/or natural substance, biodegradable to some extent. By ext: ecological
(6) Built on the word *biotechnologie*

Table 19 Hypothesis of new meanings for *bio-* . 1, 2 and 3 are attested while 4,5 and 6 are hypotheses.

This production is related to the introduction of the noun *biotechnologie* in the dictionary (1980):

> "ÉTYM. 1980; de *bio-*, et *technologie*, probablt d'après l'angl. *biotechnology*.
>
> Didact. Technique qui met en œuvre les propriétés biochimiques d'êtres vivants pour améliorer la production agricole ou certaines fabrications industrielles. La biotechnologie utilise des micro-organismes (enzymes) pour réaliser des transformations ou des synthèses (en chimie, en pharmacologie…)." [171]

Since the GRLF states it is probably a borrowing from English, the En. definition of *biotechnology* follows:

> "biotechnology, n. Etymology: < bio- comb. form + technology n., after German Biotechnologie (K. Ereky, 1919 or earlier). Compare earlier biotechnics n. and also biotechnologist n.

[171] "Etym. 1980 ; from bio- and technologie, probably after En. Biotechnology. Didact. Technique that implements the biochemical properties of living beings to improve agricultural production and some industrial production. Biotechnology uses microorganisms (enzymes) to perform transformations or synthesis (chemistry, pharmacology ...)."

> 1. The application of science and technology to the utilization and improvement of living organisms for industrial and agricultural production and (in later use) other biomedical applications; a technique or technology used in this way. In later use: spec. = genetic engineering n. Cf. earlier biotechnics n. 1.
>
> 2. The application of science and technology to practical problems of living; the study of the interaction of human beings and technology. Cf. bioengineering n., and earlier biotechnics n. 2."[172]

The second En. meaning is not listed in the Fr. definition. *Biotechnologie* is a relatively young word (1980) spawning derivations (*biotechnologiste*) and abbreviations (*biotech, biotechno*), as well as a variability of spellings in the corpus. However, there are a few differences between the En. and the Fr. definitions. It seems that the idea of *the interaction of human beings and technology*, especially in words such as *biopunk*, is imported from En.

The new disciplines and realities in the field of biotechnologies are morphologically productive in *bio-*. The neologisms retain the original meaning "related to life", however, this meaning is extended to artificial life.

3.2.3.3.5. ***Ambiguities in bio-***

Some words were difficult to classify since they showed some degree of ambiguity, and therefore a category was created for them.

Recently attested words include *biosécurité* (1990) which does not only involves GMO control, but also the control of artificial lives and genetics as the *bio-* in *biosécurité* carries the meanings contained in the word *biotechnologie* ("biotechnology") in addition to the meaning "related to life forms"(1). Therefore, in *bio-* (6) the idea of artificial life is added to the idea of "life" found until then in (1), (2), (3) and (4) by extension.

In the same vein, *biosurveillance* ("biowatch") refers to the sanitary control of living organisms and foods (based on the meaning derived from the Greek root), and includes GMO watch. However, GMO presence also defines *bio-* (4) but at the opposite: "GMO free". The same pattern applies to *biovigilance* ("biowatch") attested since 1989. In the corpus, it is used in the sense of controlling GMO presence for health and safety.

---

[172] OED

3.2.3.3.6. ***Contradictions crystallizing emerging meanings***

It is therefore jarring to determine without a context whether a *bioentreprise* refers to a company involved in activities related to biology and biotechnology or to a company involved in activities related to organic products. In the corpus the first meaning prevails, while online both coexist. It is difficult to say whether these words undergo polygenesis or if a linguistic community is borrowing these words from another linguistic community (thereby attributing a different meaning to them on some occasions).

This state of affairs creates confusion, and journalists are sometimes mistaken. For instance the very low frequency item *bioraffinerie* ("biorefinery") in which *bio-* means "made from organic substance" is described by the author of a corpus article (09/2002) as an organic product, as opposed to a "traditional" (mass consumption) product, in the same way that one would distinguish an organic carrot from one produced using pesticides.

The meaning categories delineated above tend to become blurry, and generate ambiguities. Either similar words are employed in two different meanings, on the model of *bioplastique* (adj) and *bio-plastique* (n.), or their meaning is ambiguous. There are four hypotheses of how these ambiguities come to being:

(1)Neologisms of this kind can either be the result of polygenesis, or

(2) they may be re-employed under a different meaning by different linguistic communities, which give them different meanings.

(3) Several meanings of *bio-* can blend or merge, as a result of (1) and/or (2) and/or socio-linguistic ambiguity.

(4) The new meaning categories allow for ambiguity as they emerge, (the level of ambiguity of new meanings may also influence the attested meanings and meanings associated to them).

The ambiguities rely on combinations of attested meanings (1,2 and 3 especially via 2) with unattested ones.

Although these meanings are dealt with in separate categories, there are conceptual links between them: Scientific ecology relies on the advances of sciences and technologies in the

domain of ecology. As such, words in *bio-* pertaining to the domain of scientific ecology belong to the definition of *bio-(*2) in which biology is in contact with other sciences: here both technology and ecology, resulting in a paradigm merging biology (bio-), ecology (eco-) and technology (techno-). The postulated *bio-(*6) is closely related to meanings in *bio-(*1), (2) and (3) since technologies are in touch with the domains of biology and sciences in general and are themselves part of sciences at large, therefore somehow included in meaning (2), as long as the meanings "*the interaction of human beings and technology"*(from En.) does not emerge. The very low frequency word *biosenseur* ("biosensor") is a good example of bridging meaning between *bio-* (1), (2), (3) and (6). A biosensor is a neurochip that detects the effects of gaz on living organisms. The term *biopuce* ("biochip") is attested since 1997 and is a calque of the En. word (1981). As biology and technology meet, *biosenseur* fits perfectly the definition of *bio-(*2), however biology and technology have given birth to biotechnology in the meantime, and this specific sub-category of (2) has been enriched with supplementary meanings in (6). The term is used in the domain of biotechnology and shows how *bio-* (2) extends to (6).

The hypotheses of the emergence of the meanings bio-(4), (5) and (6) by extension and of ambiguity of the element *bio-* are based on several observations. Summarized below, they are:

- *bio-* (2) generates the word *biotechnologie,* and the later generates *bio-(*6)
- the abbreviation of *biologique (adj.* fam) generates *bio-(*4)
- *bio-*(4) encompasses the notions "biodegradable" and "ecologically produced" to some extent and generates *bio-(*5) in conjunction with the original meaning of *bios* "life" with the idea of protecting natural life via ecology.

This results in the presence of words in *bio-,* in which *bio-* may mean one thing and its opposite. In *bio-jeu* ("biogame") the concept is the digital simulation a life. In *biopesticide,* the concept is developing natural and ecological products. How do we get from one meaning to the other? While *bio-* (4) relies on the concept of the "natural", *bio-* (3) relies on the study of "nature". Even though the two meanings seem to be separated by domains: one is scientific, the other familiar; they share common associated concepts. Moreover the familiar and unattested *bio-(*4) makes a connection with ecology via agriculture. Biology is interested in GMOs, ecologists are too, except the first use *bio-* to refer to any type of "life" including artificial life, whereas the second refer to "natural life". Moreover, with *bio-* (6) the idea of

"life" is extended to "artificial life". Therefore, ambiguous words could mean one thing and its opposite: artificial life or natural life, GMO or non-GMO, organic and ethical or not (see *biosurveillance*). Since people who choose to buy organic products buy non-GMO products, they are often anti-GMOs, and defenders of "natural" products. The non-GMO feature of that sense embeds it in the discourse of ecologists. *Bio-* (4) is not limited to products; it is also related to a life style in touch with the preoccupations of popular ecology (5). Ecologists are already in touch with the vocabulary of ecological sciences, in which they are confronted with many words in *bio-* (1), (2), (3) and (6) as they themselves create words in *bio-* (4) and (5). It is difficult to tell how much is borrowed and how much is the result of polygenesis. Words that are based on the meaning "ecological" are extensions from this that do not respect the original meaning of *bio-*. This complex network of ambiguity and contact between specialized communities results in a kind of homonymous competition for some words (see *bioentreprise, biovigilance…*)

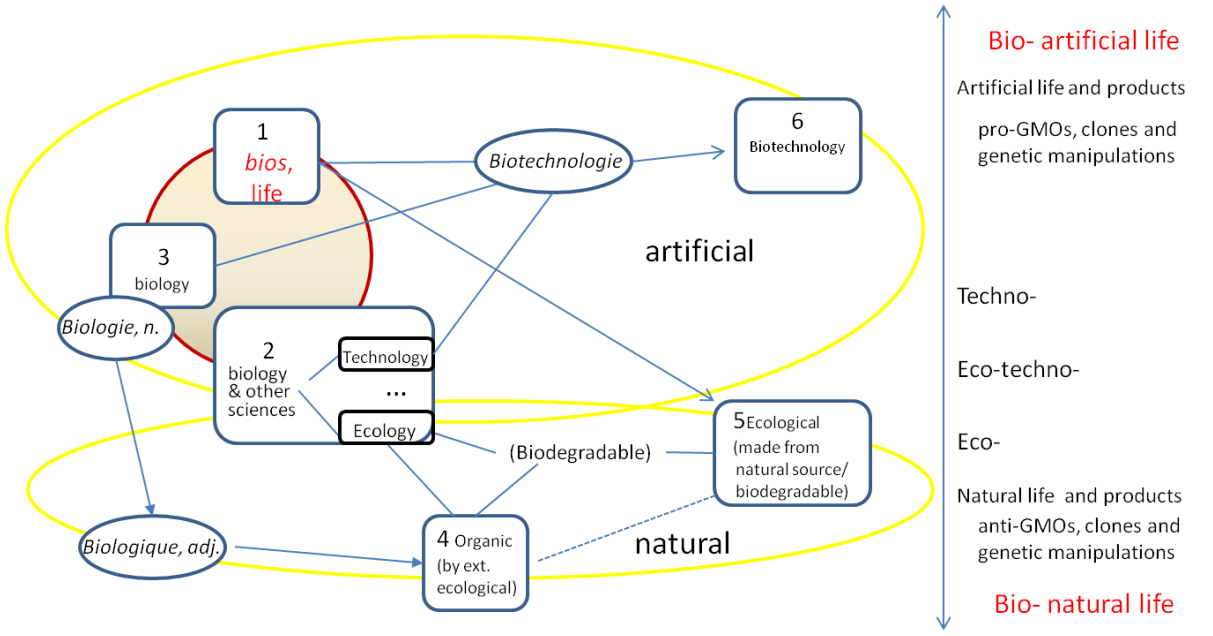


Figure 34 Schematic representation of the hypothetical meanings of *bio-* in the extracted neologisms. The area circled in red corresponds to the attested core meaning and the two areas circled in yellow to the new hypothetical meanings emerging. Concepts have square labels while words have circular labels.

Overall, and as illustrated in Figure 34, it seems that the original Gr. root *bios* gives birth to two sub-meaning sets in the process of structuring:

-**artificial life**, genetic manipulation of living organisms, cloning, human amelioration via machines, the possible fusion of humans and machines, TECHNOLOGY [BIO-technology]

(bridging :ECO-TECHNOLOGY)

-natural life, eating organic and natural products, making products ecologically from organic and biodegradable substances (foods, textiles, plastics, fuels,…) ECOLOGY [BIO-Ecology]

Artificial and natural life are in opposition, however, the definition of what is considered artificial or natural is unclear. This may be because the two meanings are a reflection of an emerging major social debate about the status of life. *Bio-* (2) seems to generate two semantic extensions, one via technology (*bio-6*) and one via ecology (*bio-*5). This process is influenced by the emergence of the abbreviation of the adjective *biologique*, *bio-(*4), which both has strong associated meanings with ecology and some associated meanings with technology via eco-technology. The interaction of these words creates ambiguities as they sometimes unfold as polygenesis and result in homonymous words with opposite competing meanings.

Cyber- and *bio-* seem to crystallize an emerging debate: the relationship of humans to technologies and nature. However, this process is entangled into other processes due to ambiguity, as can be said of *eco-*, which can refer to ecology and economy, or both.

After having looked at formal neology and morphological productivity, the next study zooms into more detail to look at more subtle traces of change, through the analysis of semantic shift (also called semantic drift) for a single word shaken by events.

3.2.3.4. *Mondialisation vs. globalisation* : synonymic competition and connotational drift

3.2.3.4.1. ***Analysis***

In the state of the art (Part I), I have put forth the hypothesis that if two words are quasi-perfect synonyms, one of them or both is/are likely to undergo a type of change. This hypothesis is built on the Saussurean assumption that within language taken as a system, elements owe their existence to their relationship of opposition. The Fr. words *mondialisation* and the Anglicism *globalisation* are in a relationship of synonymy. This process goes hand in hand with the creation of derivations built on *mondialisation*, and one for *globalisation*. Two linguistic indices enter into play: synonymic competition and associated morphological production. Out of the two, only *mondialisation* undergoes a subtle and progressive semantic shift.

3.2.3.4.1.1. Definitions

First, the definitions for the two words are compared:

> **"Mondialisation**
>
> n.f. – 1953 de mondial. Le fait de devenir mondial, de se répandre dans le monde entier. *La mondialisation d'un conflit.* SPECIALT Phénomène d'ouverture des économies nationales sur un marché mondial libéral, lié aux progrès des communications et des transports, à la libéralisation des échanges, entraînant une interdépendance croissante des pays. globalisation (ANGLC.) altermondialisme, antimondialisation."

> **" Globalisation**
>
> n.f.-1968 de globaliser. 1. Action de globaliser, son résultat. 2. (de l'anglais *globalization*)
> ANGLC. Mondialisation. *La globalisation des marchés.* "[173]

Both words are grounded in the same notion. However, *mondialisation* is more precisely defined than *globalisation*, to which it is redirected as a synonym. *Globalisation* is a calque from En. g*lobalization* and entered the French press via finance and translated news dispatches. It is noteworthy that many writers and speakers find it necessary to give precisions when they use *mondialisation* as in, for instance, *mondialisation économique* ("economic globalization") *mondialisation financière* ("financial globalization") and *mondialisation politique* ("political globalization"). Opinions diverge as to the exact coverage of *globalisation*: for some it goes further than *mondialisation*, as for others, *globalisation* is restricted to the financial domain. This uncertainty in usage is also a sign of semantic instability, as it shows that there is no established consensus about these definitions.

---

[173] Source : Le Petit Robert 2010. "Mondialisation: From worldwide. Becoming worldwide, spreading to the world. *Globalization of a conflict.* Specialt phenomenon of national economics opening to a global liberal world market, related to the progress of communication and transports and trade liberalization, resulting in a growing interdependence of countries. globalisation (ANGLC.) altermondialisme, antimondialisation"

"Globalisation: from to globalize. 1. Action of globalizing, its result (from English globalization) ANGLC. Mondialisation. *Market globalization*."

3.2.3.4.1.2. Frequencies

Frequency of use is stable for *globalisation* which is used 5.5 times less than its competitor *mondialisation.* In total, *globalisation* appears 1879 times, whereas *mondialisation* appears 9859 times.

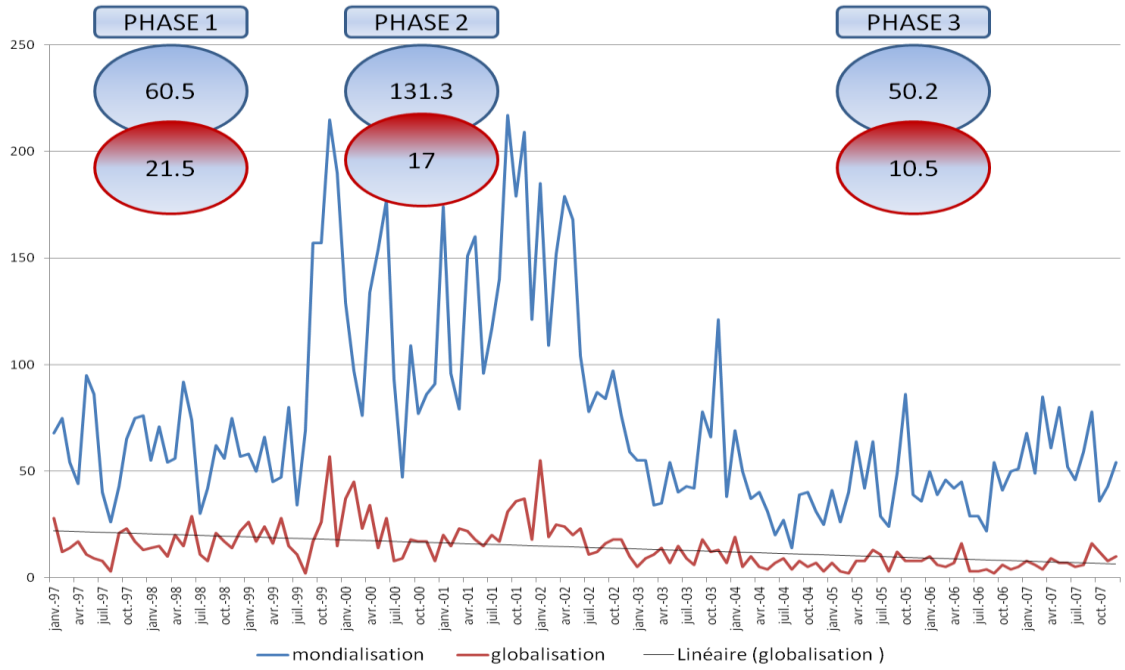


Figure 35 Raw frequencies for the words *mondialisation* and *globalisation*, and mean by phase, after considering three frequency phases, in the corpus *Le Monde* (1997-2007) in stemmed version.

In Figure 35, showing raw frequencies for both words, three phases appear in the frequency behavior of *mondialisation:* The first from January 1997 to July 1999, with a mean frequency of 60.5, the second from August 1999 to June 2002 with a mean frequency of 131.3 and the third, from July 2002 to December 2004 with a mean frequency of 50.2.

The use of *globalisation* seems to decrease, and to check whether the decrease is not biased by the fact that the monthly corpus size by is reducing over time, the frequencies are normalized (using the total number of words per month), and even though there is a slight decrease (regression coefficient of -0,0013), *globalisation* can be said to have a stable use overall, with frequency peaks during the second phase, as shows Figure 36:

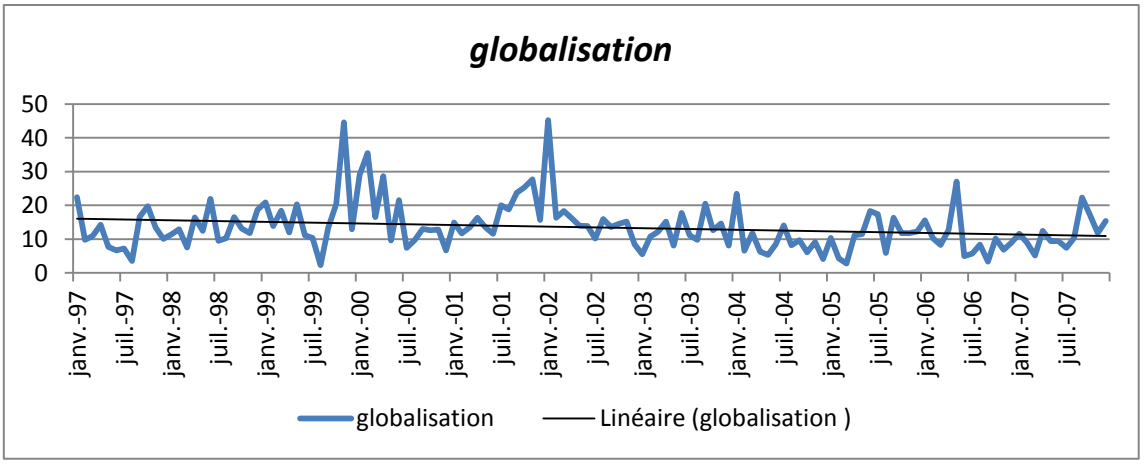


Figure 36 Normalized frequencies for the word *globalisation*, in the corpus *Le Monde* (1997-2007) in stemmed version, with linear regression.

The three observed phases are used in subsequent analyses. For *mondialisation*, the first and last phases are on the same scale while the second shows a frequency peak. I therefore try to determine what happens in the second phase and whether this frequency peak impacts semantics.

3.2.3.4.1.3. Sections

Both *globalisation* and *mondialisation* enter two major sections of the newspaper in March 2000: the sections "Editorial and analyses" and "Debates and deciphering" [174]. Before March 2000 there were no occurrences of these terms in these sections. This shows that the words and the topics attached to them moved to the forefront in the media at this time. This sudden entry in the sections happens six months after the second frequency phase has started.

3.2.3.4.1.4. Co-occurrence networks

To investigate the semantic networks of these two words in the corpus, all co-occurrent words are extracted for each candidate in each sentence in chunked units of the corpus in the ACOM format of a month. Co-occurrent words are then classified hierarchically according to their co-occurrence mean value per month as shown in table 20, an excerpt of the co-occurrence table for *mondialisation*:

[174] « Éditorial et analyses » and « Débats et décryptage ».

date	NOM mondialisation	DET:ART le	PRP de	PRP:det du	(…)	ADJ nouveau	NOM pays	PRO:IND tout	PRP contre	…
199701	68	273	154	68	(…)	9	7	4	0	…
199702	74	287	181	65	(…)	6	5	10	3	…
199703	54	239	127	58	(…)	5	5	7	4	…
199704	44	219	133	72	(…)	2	7	3	4	…
199705	94	404	230	90	(…)	3	15	13	8	…
…	…	…	…	…	…	…	…	…	…	…

Table 20 Excerpt of the co-occurrence table for the noun *mondialisation* in the corpus *Le Monde* (1997-2007) in stemmed version,

Overall and in hierarchical order, the twenty strongest co-occurrent plain words for the two candidates are the following[175]:

GLOBALISATION	MONDIALISATION
NOM économie	ADJ nouveau
ADJ économique	NOM pays
NOM marché	PRP contre
ADJ nouveau	NAM Europe
NOM pays	ADJ social
NOM monde	ADJ économique
ADJ mondial	NOM économie
ADJ financier	ADJ européen
ADJ social	NOM monde
PRP contre	ADJ grand
ADJ politique	NAM France
ADJ international	NOM marché
NOM mondialisation	ADJ international
ADJ européen	ADJ libéral
NAM Europe	ADJ politique
NOM effet	NOM effet
NOM entreprise	ADJ mondial
NOM politique	NOM monsieur
NOM société	NOM entreprise
NOM échange	ADJ français

Table 21 Comparison of the twenty strongest co-occurrent words for *globalisation* and *mondialisation* in the corpus *Le Monde* (1997-2007) in stemmed version, [176]

[175] After removing conjunctions, determiners, and most commonly used prepositions, pronouns, adverbs and adjectives

The two networks are almost identical, confirming the synonymic use of the two words. A few co-occurrent words differ: *financier* ("financial"), *société* ("society") and *échange* ("trade") for *globalisation*; and *grand* ("great, big"), *France*, *libéral*, *monsieur* ("mister") and *français* ("French") for *mondialisation*.

However, these items appear in both tables, in lower hierarchical positions, as in *mondialisation*'s co-occurrent list, the words *financier* (80th position) *société* (76th position) *échange* (104th position) appear. In the same way, in *globalisation*, the words *grand* (53rd position) *France* (92nd), *libéral* (154th) *monsieur* (110th) and *français* (118th) appear.

3.2.3.4.1.5. Co-occurrence networks after re-chunking the corpus according to frequency phases

Since *globalisation* is stable in terms of frequency, I focus on *mondialisation*. To determine

Janv. 97 – juil. 1999	Août 99 – juin 2002	Juil. 02 – déc. 07
économie	contre	Europe
pays	pays	pays
Europe	monde	nouveau
France	Europe	social
économique	économie	européen
face	économique	économique
monde	politique	monde
marchés	libérale	France
contre	autre	autre
politique	France	contre

Table 22 The 10 highest frequencies co-occurrent words of *mondialisation* within the three phases observed earlier, in the corpus *Le Monde* (1997-2007) in stemmed version,

---

[176] **Globalisation** : Economy, Economic ,Market, New, Country, World, World (adj. masc), Financial , Social, Against, Political, International, *mondialisation*, European, Europe, Effect, Company, Politics, Society, Exchange.
Mondialisation: New, Country, Against, Europe, Social, Economic, Economics, European, World, Big, France, Market, International, Liberal, Political, Effect, World (adj.) Mister, Company, French.

what happens in the three frequency tables observed beforehand, the corpus is re- chunked in three time sections. Table 22 shows that the network is stable overall.

There is no major semantic change, however there is reorganization within the hierarchy of co-occurrent words, showing instability in structure.

The word *contre* ("against"), in 9th position in the first phase, reaches the first position in the second phase and is back to the tenth position in the third phase. *Contre* is a pivot word and concept around which connotational drift unfolds. The word *économie* is the strongest co-occurrent word in phase one, the 5th in phase two, and gets to the 55th position in the third phase. However the adjective *économique* stays stable in the three phases. *Politique*, in tenth position in phase one and seventh in phase two, is not in the ten most frequent co-occurrent words in phase 3, reaching the 56th position. However, the interpretation of the change in status of *politique* may mean two different things: either *mondialisation* is less and less associated with politics or it is more and more so. The concept of politics is strongly embedded in *mondialisation*, as strongly as economics are, and it may not be necessary, after phase 2 to mention it in discourse since it becomes implicitly contained in *mondialisation*. To investigate *politique* more closely, ranks are calculated.

3.2.3.4.1.6. Ranks

To show the evolution of a single co-occurrent word's importance relative to the other co-occurrent words for the target word, ranks are calculated. The rank is the hierarchical position of a co-occurrent word in the table per month. It shows the position of a co-occurrent word in the network. If the co-occurrent word does not appear in a month chunk, it is noted zero. The word ranked 1 is the most common co-occurrent. The counts include all words, and therefore the thirty first positions are filled with determiners, pronouns, adverbs and conjunctions as well as most commonly used verbs. For a noun, a rank of 30 or 40 is a high position. All words are kept in the count to allow for the search of pivot function terms. The semi-automatic method makes the detailed search for patterns of word pairs easier when co-occurrent words are far in the hierarchy (in ranks of 200 for instance) at one point and close at another point in time. The graph should be read "upside-down", since words closer to 1 have a more important role in the contextual environment of the target word and therefore play a role at the level of connotation.

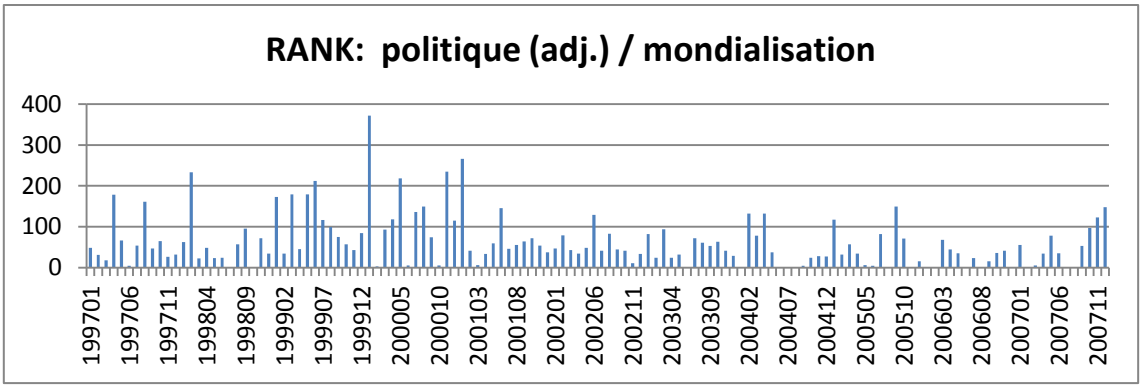


Figure 37 Ranks for the adjective *politique* co-occurring with *mondialisation* in the corpus *Le Monde* (1997-2007) in stemmed version.

Figure 37 shows that the word *politique* is more and more important in the direct environment of *mondialisation*, although it occurs less. This measure is distinct from frequency, since a co-occurrent word may not appear frequently, but become more important in rank at the same time when it is present. *Politique* does not appear in the third phase in Table 23 since it is too infrequent to be amongst major co-occurrent words, as it is in 56th position in terms of frequency. The fact that it becomes less frequent does not necessarily mean that it is becoming less important semantically. This is especially the case if it becomes integrated to its semantic field: it is less and less necessary to mention politics when talking about *mondialisation*, but when it is mentioned, it is done in an increasingly important way as shown by the diminishing values of ranks. At some point, when a co-occurrent word has become extremely significant, it is semantically "absorbed" by the target word, of which it becomes an implicit feature. Therefore there is no need to talk about "mondialisation politique" (as opposed to "mondialisation économique/financière" initially) since "mondialisation" now contains the idea of "politique".

3.2.3.4.1.7. Pivots

Co-occurrence networks and ranks showed that a few co-occurrent words operate as pivots, around which meaning shift is articulated. *Contre* becomes the strongest co-occurrent word in phase 2 and its rank is extremely low from the beginning of 1999 to the end of 2003, as shown in Figure 38. Ranks are high before and after that period, which indicates that the term is losing importance.

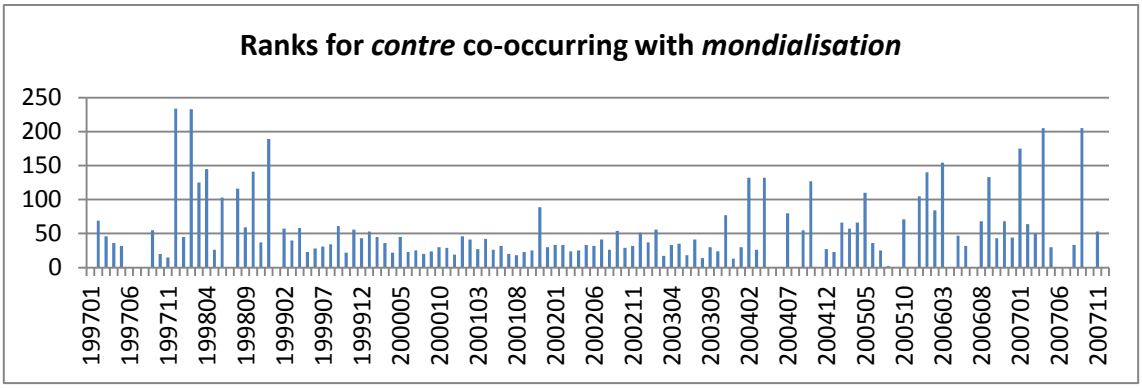


Figure 38 Ranks for the preposition *contre* co-occurring with *mondialisation*, in the corpus *Le Monde* (1997-2007) in stemmed version.

Ranks increase, while the number of occurrences drastically decreases in the third phase. However, if frequencies seem extremely low in the third phase while looking at raw co-occurrence frequencies, we see that they are still significant relative to the number of occurrences of the target word *mondialisation*, when generating normalized co-occurrence frequencies (see Figure 39).

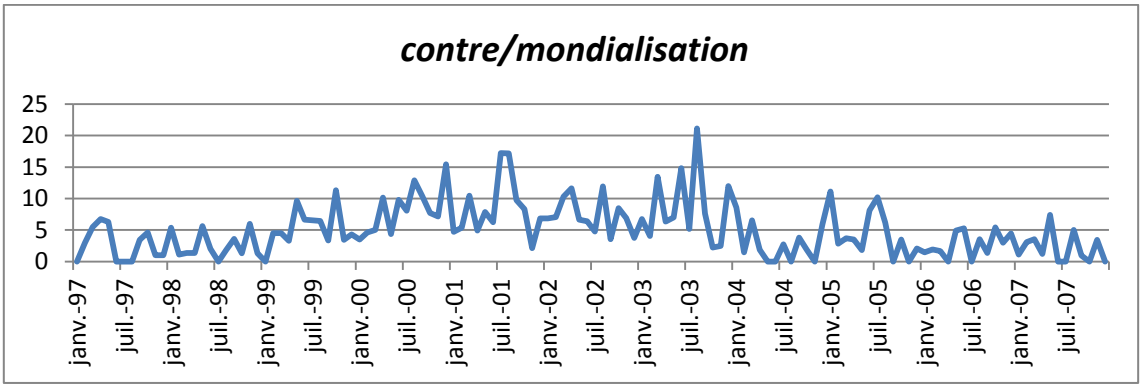


Figure 39 Normalized frequency of the co-occurrence of *contre* with *mondialisation* in the corpus *Le Monde* (1997-2007) in stemmed version.

The co-occurrence of *politique* increases in the second phase (see Figure 40), but as normalized co-occurrence frequencies show, this is due to the increasing frequency of the target word. However, *politique* becomes more important relatively to the frequency of the target word after this phase (see Figure 41). Ranks (see Figure 37) decrease showing that the word is more and more important in the context of *mondialisation* even though it is less frequent.

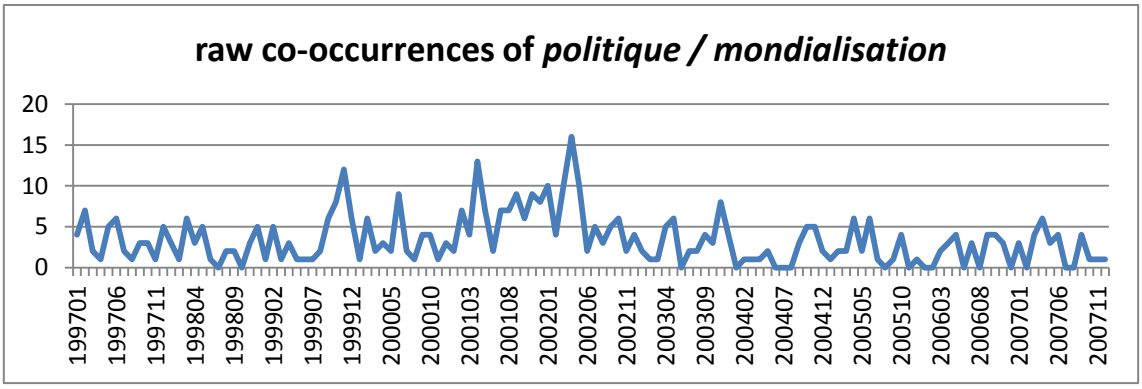


Figure 40 Raw co-occurrence of *politique* with *mondialisation*, in the corpus *Le Monde* (1997-2007) in stemmed version.

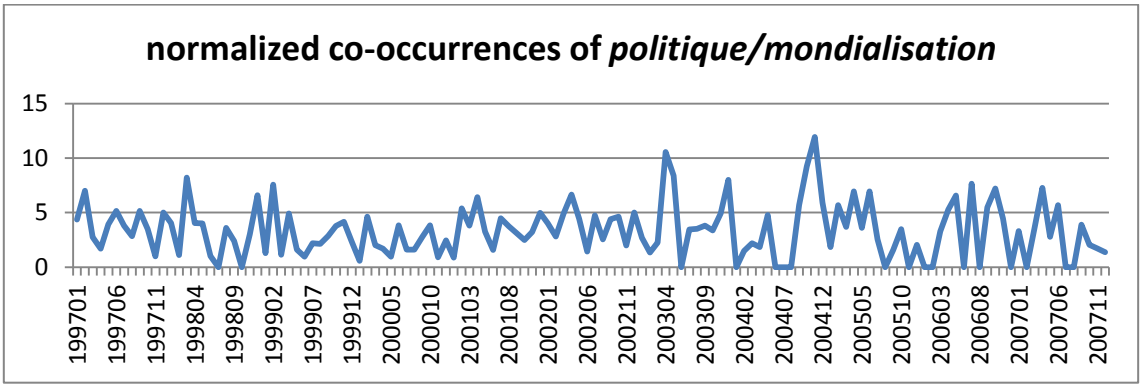


Figure 41 Normalized co-occurrence frequency for *politique* co-occurring with *mondialisation* (normalized in terms of frequency of the target word *mondialisation*) in the corpus *Le Monde* (1997-2007) in stemmed version.

The decreasing raw co-occurrence frequency is misguiding in the case of *politique*, since the term gains importance on the semantic plane. This appears in normalized co-occurrence frequencies and in the ranks.

The semantic shift of *mondialisation* is organized around these pivot words. Pivot words also point to a series of events, in which political opinion is turned against *mondialisation* as a concept.

3.2.3.4.1.8. Events

In phase two, from August 1999 to June 2002, *mondialisation*'s frequency increases to more than double. At the same time *contre* takes the first position in the co-occurrence network.

This time period is also strongly marked with national and international events related to globalization issues (see Figure 42). In November and December 1999, demonstrations took place in Seattle against the World Trade Organization (WTO) conference. In August the French activist José Bové led actions against the effects of globalization and was taken to court. In January 2001 the first World Social Forum took place in Porto Alegre, followed by the events of the 11th of September. In July 2001, demonstrations took place against the G8 in Genoa, and others followed the Doha WTO conference in November 2001. In 2002 the second Social Forum took place in Porto Alegre.

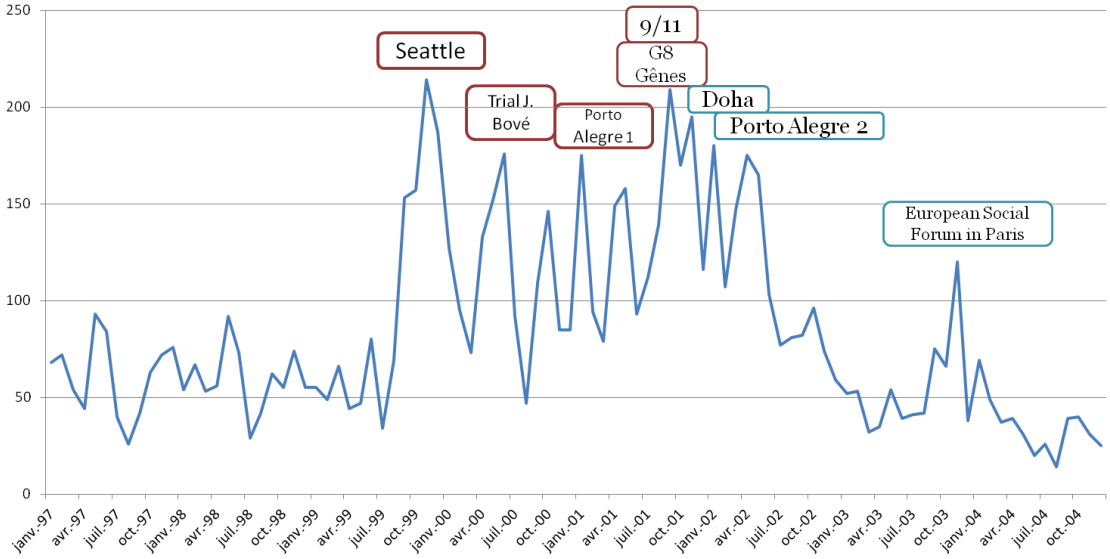


Figure 42 Frequency of *mondialisation* tagged with corresponding events in the corpus *Le Monde* (1997-2007) in stemmed version.

All these events mark the history of an ideology that becomes a political movement against a certain form of capitalism. It takes the name of *antimondialisme* ("antiglobalization" is the closest translation, lit. "anti-worldness") eventually renamed *altermondialisme*.

3.2.3.4.1.9. Morphological productivity

Globalization is perceived as a challenge for society. This challenge is also presented as a threat. The anti-globalization movement of was created against this threat. The movement calls itself a movement of *antimondialisation* ("antiglobalization"), a neologism entering the PR in 1997, and giving birth to the neologisms *antimondialiste*, *altermondialisme*, *altermondialisation* and *altermondialiste*, according to the following scheme:

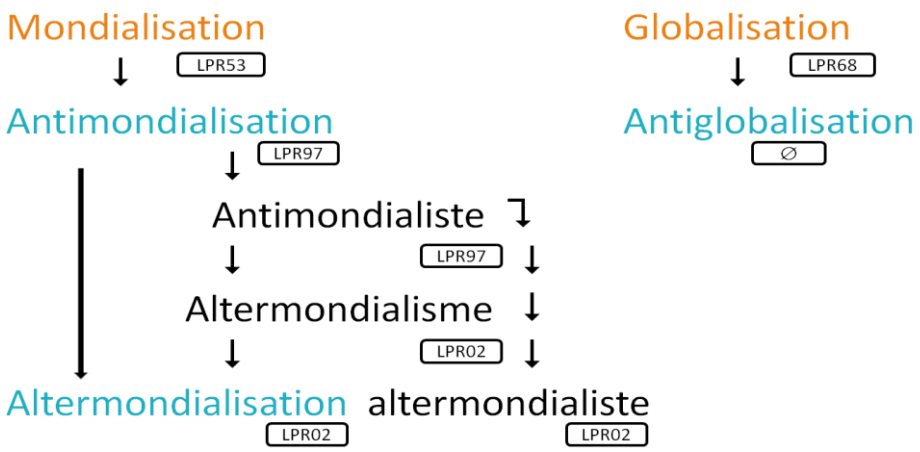


Figure 43 Classification of neologisms based on *mondialisation*, with entry dates in the PR.

Globalisation also gives birth to *antiglobalisation*, a word used in the corpus, but which does not enter the PR. There are three derivational neologisms built on the prefix *anti-* (idem) and *mondialisation*, entering the LPR in 1999 and two based on the element *alter-* , taken from Lat. "other", entering the PR in 2002. What is the difference in meaning between the constructions in *anti-* and *alter-* ? The definitions of *antimondialisation* and *altermondialisation* are the following:

> "Antimondialisation. n.f. et adj.inv.- 1997. De *anti-* et *mondialisation*. Mouvement de protestation qui s'oppose à la mondialisation, qui redoute ses conséquences économiques, sociales, écologiques. aussi altermondialisation, altermondialisme.
>
> –adj.inv. *Les militants antimondialisation*. antimondialiste.
>
> Altermondialisation. n.f. – 2002. De *antimondialisation*, d'après *altermondialisme*. Courant d'opinion qui propose un type de développement économique opposé au modèle libéral (mondialisation) plus soucieux du développement de l'homme et de la protection de l'environnement. altermondialisme."[177]

These two definitions are highly similar, and almost synonymic. However, there is a shift between the idea of a "movement of protestation against" globalization and an "opinion trend which offers" an alternative model. One can read between the lines that the *antimondialiste* movement , which was originally a movement of opposition (anti-), had to redefine its

[177] *Antimondialisation* : Protest movement opposed to globalization, which fears its economic, social and ecological consequences. *Altermondialisation*: Body of opinion which offers a type of economic development opposed to the liberal model (globalization) and is more concerned with the development of humankind and the protection of the environment.

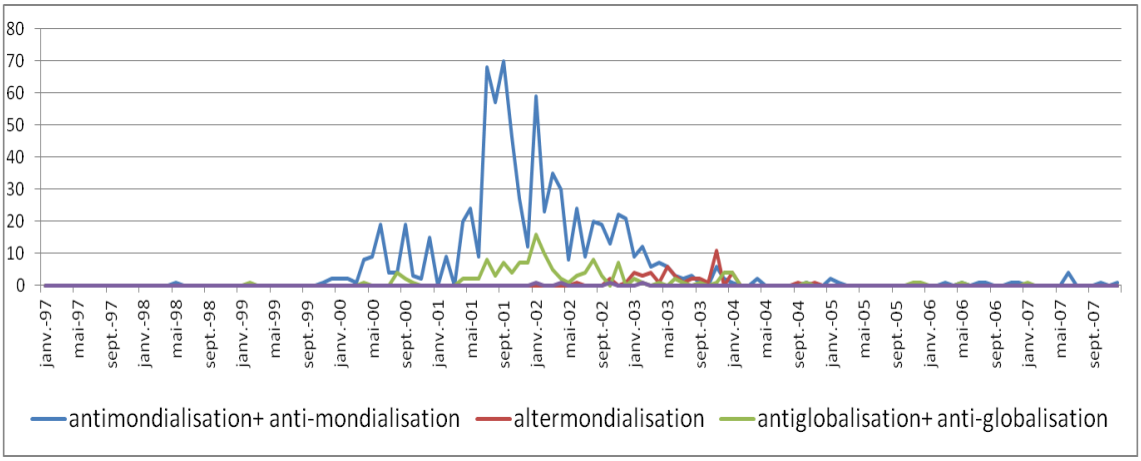


Figure 44 Raw frequencies for the neologisms based on *mondialisation* and *globalisation* in the corpus *Le Monde* (1997-2007) in stemmed version.

ideology and discourse as a political movement offering another vision (alter-), therefore becoming *altermondialiste*. In the corpus, these neologisms appear within phase 2, which seems to be a sort of 'catalyst'. Surprisingly the terms in alter-, used by the advocates of the movement, have very few occurrences, while terms in anti- have more in the second phase.

All terms show no occurrences before the second phase and after January 2004 (see Figure 44). *Antiglobalisation* (and *anti-globalisation*) does not enter the PR but has 132 occurrences in the corpus.

The question may be raised as to whether the creation of neologisms in anti- and alter- based on *mondialisation* have an impact on the definition of the source word. Indeed, these terms give *mondialisation* a clearly defined connotation, anchored in economic and political notions, and the idea of a type of society. The neologisms are not based on the restricted sense of *mondialisation*, defined as "becoming worldwide" and the idea of the "global village" as it is described by McLuhan (1994).

3.2.3.4.1.10. Measuring density and cohesion variability

Frequency variations as well as structural variations (in co-occurrence networks) were detected. These variations indicate that a change is taking place. The SA offers additional tools to evaluate the nature of this change with the system of cliques. The model generates associated terms and cliques for a word in each time chunk. From this data, the density and cohesion index is created. This index is based on the variation of the ratio between the number of cliques and the number of associated words for a head word over time.

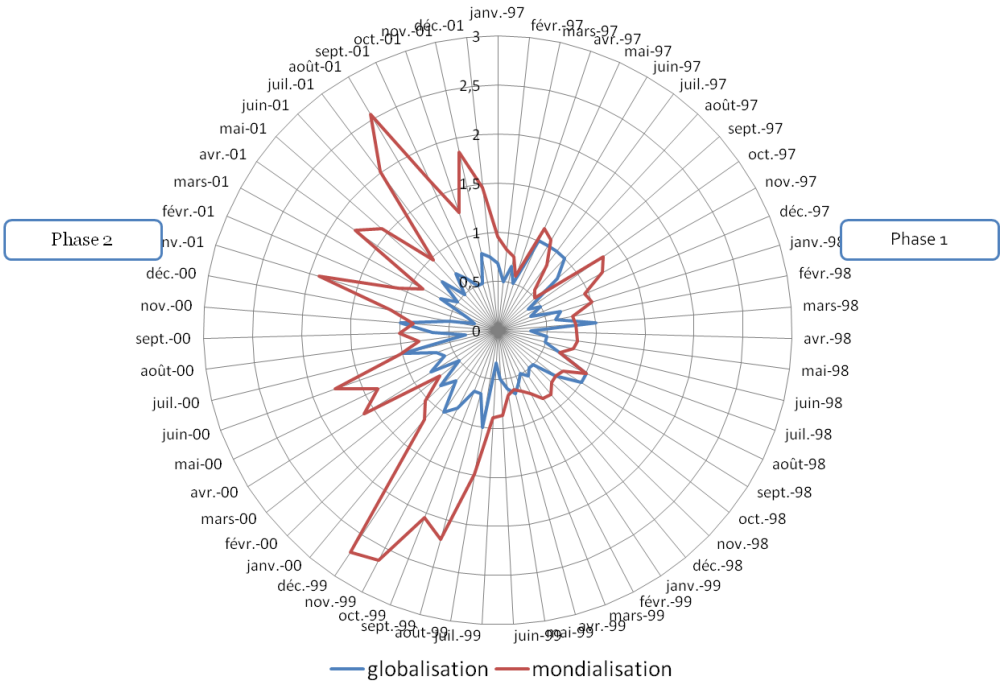


Figure 45 Representation of the density and cohesion variability, measured by the ratio cliques-terms for *mondialisation* and *globalisation* between January 1997 and December 2001.

Figure 45 shows a stability of the cliques-terms ratio for *globalisation*, since the ratio is comprised between 0.25 and 1. The ratio for *mondialisation* shows a variation comprised between 0.5 and 2.71. High variation is observed in phase two.

The proliferation of associated terms for *mondialisation* in this time period indicates an increase in density, which may be due either to the expansion of known contexts or to the addition of new contexts. The generation of numerous new cliques that mostly contain new associated terms show that it is due to an addition of new contexts. The high ratios observed at some points in time relate to the phases in which cohesion is in the process of restructuring. The density of links between associated terms increases. This shows that *mondialisation* is used in one or several new contexts in that period. For instance, the new context related to anti- and alter- emerges through the introduction of a new set of associated terms at the end of 1999. The model has to create new pivots (or "articulations") between cliques to integrate the new context in a continuous paradigm with the pre-existing ones. This phenomenon can be zoomed on by looking at cliques in detail. In August and September 1999 at the beginning of the first peak, the number of associated terms jumps from 38 to 73 and the number of cliques from 58 to 163. There are 20 new associated words and 90 new cliques in September. The new contexts of *mondialisation* are present in August and develop in September. As illustrated by the blue zones, in Figure 46 in August about half of the cliques are related to the general meaning of *mondialisation*, described above, and the other half is related to the activist farmer confederation, which called itself *altermondialiste*. In September, this theme

expands around the notion of collective citizenship (*citoyen*) and verbs related to the necessity to get involved in the defense of an idea.

August 1999 : 38 termes / 58 cliques

1 : ADJ capable, ADJ humain, NOM mondialisation, VERpper priver
2 : ADJ capable, NOM mondialisation, VERinfi créer, VERpper priver
3 : ADJ commercial, ADJ paysan, NOM commerce, NOM construction, NOM mondialisation
4 : ADJ commercial, ADJ paysan, NOM confédération, NOM construction, NOM mondialisation
5 : ADJ commercial, NOM commerce, NOM construction, NOM institution, NOM mondialisation
6 : ADJ commercial, NOM commerce, NOM cycle, NOM mondialisation
7 : ADJ commercial, NOM commerce, NOM finance, NOM mondialisation
8 : ADJ commercial, NOM commerce, NOM institution, NOM mondialisation, NOM union
9 : ADJ commercial, NOM commerce, NOM mondialisation, NOM échange
10 : ADJ commercial, NOM confédération, NOM cycle, NOM mondialisation
11 : ADJ commercial, NOM confédération, NOM mondialisation, NOM porte-parole
12 : ADJ culturel, ADJ humain, NOM mondialisation
13 : ADJ culturel, ADJ militant, NOM confédération, NOM mondialisation
14 : ADJ culturel, ADJ militant, NOM institution, NOM mondialisation
15 : ADJ culturel, NOM commerce, NOM institution, NOM mondialisation
16 : ADJ culturel, NOM commerce, NOM mondialisation, NOM échange
17 : ADJ culturel, NOM diversité, NOM mondialisation, NOM émergence
18 : ADJ culturel, NOM effort, NOM mondialisation, VERinfi créer
19 : ADJ culturel, NOM libéralisme, NOM mondialisation
20 : ADJ humain, NOM danger, NOM mondialisation
21 : ADJ militant, ADJ paysan, NOM confédération, NOM construction, NOM mondialisation
22 : ADJ militant, ADJ paysan, NOM danger, NOM mondialisation
23 : ADJ militant, ADJ paysan, NOM débat, NOM mondialisation
24 : ADJ militant, NOM confédération, NOM mondialisation, NOM porte-parole
25 : ADJ militant, NOM construction, NOM faveur, NOM mondialisation
26 : ADJ militant, NOM construction, NOM institution, NOM mondialisation
27 : ADJ militant, NOM danger, NOM institution, NOM mondialisation
28 : ADJ militant, NOM débat, NOM mondialisation, NOM porte-parole
29 : ADJ puissant, NOM commerce, NOM institution, NOM mondialisation
30 : ADJ puissant, NOM confédération, NOM mondialisation

31 : NAM Europe, NAM Français, NOM mondialisation
32 : NAM Europe, NOM construction, NOM mondialisation
33 : NAM Europe, NOM mondialisation, NOM union
34 : NAM Europe, NOM mondialisation, NOM économie
35 : NAM Français, NOM effort, NOM mondialisation
36 : NAM Français, NOM mondialisation, NOM planète
37 : NOM citoyen, NOM construction, NOM mondialisation
38 : NOM citoyen, NOM libéralisme, NOM mondialisation
39 : NOM citoyen, NOM mondialisation, NOM porte-parole
40 : NOM commerce, NOM construction, NOM faveur, NOM mondialisation, NOM réponse
41 : NOM commerce, NOM faveur, NOM finance, NOM mondialisation
42 : NOM commerce, NOM finance, NOM mondialisation, NOM économie
43 : NOM commerce, NOM mondialisation, NOM réponse, NOM échange
44 : NOM construction, NOM institution, NOM mondialisation, VERpper priver
45 : NOM construction, NOM mondialisation, NOM planète
46 : NOM conséquence, NOM mondialisation, NOM échelle, NOM économie
47 : NOM débat, NOM mondialisation, NOM réponse
48 : NOM débat, NOM mondialisation, NOM union, VERinfi créer
49 : NOM effort, NOM faveur, NOM finance, NOM mondialisation
50 : NOM effort, NOM mondialisation, NOM profit
51 : NOM faveur, NOM mondialisation, NOM thèse
52 : NOM flux, NOM mondialisation
53 : NOM libéralisme, NOM mondialisation, NOM thèse
54 : NOM libéralisme, NOM mondialisation, NOM économie
55 : NOM mondialisation, NOM profit, NOM échelle
56 : NOM mondialisation, NOM profit, VERpper priver
57 : NOM mondialisation, NOM république
58 : NOM mondialisation, VERpres apparaître

September 1999: 73 termes / 163 cliques

1 : ADJ collectif, ADJ global, NOM citoyen, NOM mondialisation
2 : ADJ collectif, ADJ majeur, NOM citoyen, NOM mondialisation
3 : ADJ collectif, NOM citoyen, NOM confédération, NOM mondialisation, NOM solidarité
4 : ADJ collectif, NOM citoyen, NOM mondialisation, NOM régulation
5 : ADJ collectif, NOM citoyen, NOM mondialisation, VERinfi défendre
6 : ADJ collectif, NOM mondialisation, NOM régulation, VERpres rendre
7 : ADJ collectif, NOM mondialisation, NOM victime
8 : ADJ croissant, NOM marge, NOM mondialisation, NOM pression
9 : ADJ culturel, ADJ majeur, NOM changement, NOM mondialisation
10 : ADJ culturel, ADJ majeur, NOM institution, NOM mondialisation
11 : ADJ culturel, ADJ majeur, NOM mondialisation, VERpres compter
12 : ADJ culturel, NAM José, NOM exception, NOM mondialisation
13 : ADJ culturel, NOM changement, NOM contexte, NOM mondialisation
14 : ADJ culturel, NOM contexte, NOM développement, NOM mondialisation
15 : ADJ culturel, NOM différence, NOM identité, NOM mondialisation
16 : ADJ culturel, NOM développement, NOM institution, NOM mondialisation
17 : ADJ culturel, NOM développement, NOM mondialisation, VERpres compter
18 : ADJ culturel, NOM exception, NOM mondialisation, VERinfi défendre
19 : ADJ culturel, NOM identité, NOM logique, NOM mondialisation
20 : ADJ culturel, NOM institution, NOM logique, NOM mondialisation
21 : ADJ culturel, NOM institution, NOM mondialisation, VERinfi défendre
22 : ADJ culturel, NOM mondialisation, NOM victime
23 : ADJ démocratique, ADJ libéral, NOM mondialisation, VERinfi défendre
24 : ADJ démocratique, ADJ majeur, NOM citoyen, NOM mondialisation
25 : ADJ démocratique, ADJ majeur, NOM crise, NOM institution, NOM mondialisation
26 : ADJ démocratique, NAM Etats, NOM citoyen, NOM mondialisation, VERinfi défendre
27 : ADJ démocratique, NAM Etats, NOM mondialisation, NOM nécessité
28 : ADJ démocratique, NOM citoyen, NOM mondialisation, NOM mouvement
29 : ADJ démocratique, NOM citoyen, NOM mondialisation, NOM processus
30 : ADJ démocratique, NOM crise, NOM mondialisation, NOM mouvement
31 : ADJ démocratique, NOM crise, NOM mondialisation, NOM processus
32 : ADJ démocratique, NOM dialogue, NOM mondialisation, VERpper engager
33 : ADJ démocratique, NOM institution, NOM mondialisation, NOM nécessité
34 : ADJ démocratique, NOM institution, NOM mondialisation, VERinfi défendre
35 : ADJ démocratique, NOM mondialisation, NOM mouvement, VERpper engager
36 : ADJ démocratique, NOM mondialisation, NOM processus, VERpper engager
37 : ADJ global, NAM Etats, NOM citoyen, NOM mondialisation
38 : ADJ global, NAM Etats, NOM logique, NOM mondialisation
39 : ADJ global, NAM Etats, NOM mondialisation, NOM nécessité
40 : ADJ global, NAM José, NOM citoyen, NOM mondialisation
41 : ADJ global, NAM José, NOM exception, NOM mondialisation
42 : ADJ global, NOM contexte, NOM mondialisation
43 : ADJ international, NAM Etats, NOM concurrence, NOM mondialisation
44 : ADJ international, NAM Etats, NOM mondialisation, NOM nation
45 : ADJ international, NOM concurrence, NOM institution, NOM mondialisation, NOM économie
46 : ADJ international, NOM concurrence, NOM mondialisation, NOM pression
47 : ADJ international, NOM crise, NOM développement, NOM institution, NOM mondialisation, NOM économie
48 : ADJ international, NOM crise, NOM développement, NOM investissement, NOM mondialisation, NOM économie
49 : ADJ international, NOM développement, NOM investissement, NOM mondialisation, NOM rapport, NOM économie, VERpper publier
50 : ADJ international, NOM développement, NOM mondialisation, NOM nation, NOM rapport
51 : ADJ international, NOM mondialisation, NOM nation, NOM pression
52 : ADJ international, NOM mondialisation, NOM victime, VERpper publier
53 : ADJ inéluctable, NOM mondialisation
54 : ADJ libéral, NOM exception, NOM fondateur, NOM mondialisation
55 : ADJ libéral, NOM exception, NOM mondialisation, VERinfi défendre
56 : ADJ libéral, NOM identité, NOM libéralisme, NOM mondialisation
57 : ADJ libéral, NOM identité, NOM logique, NOM mondialisation
58 : ADJ libéral, NOM mondialisation, NOM régulation
59 : ADJ majeur, NOM changement, NOM concurrence, NOM mondialisation
60 : ADJ majeur, NOM citoyen, NOM mondialisation, NOM souveraineté

61 : ADJ majeur, NOM concurrence, NOM institution, NOM mondialisation
62 : ADJ majeur, NOM concurrence, NOM mondialisation, VERpres compter
63 : ADJ majeur, NOM défi, NOM maîtrise, NOM mondialisation
64 : ADJ majeur, NOM maîtrise, NOM mondialisation, NOM souveraineté
65 : ADJ paysan, NAM Bové, NAM José, NOM citoyen, NOM confédération, NOM fondateur, NOM mondialisation
66 : ADJ paysan, NAM Bové, NAM José, NOM citoyen, NOM confédération, NOM mondialisation, NOM solidarité
67 : ADJ paysan, NAM Bové, NAM José, NOM citoyen, NOM fondateur, NOM mondialisation, NOM mouvement
68 : ADJ paysan, NAM Bové, NAM José, NOM confédération, NOM construction, NOM mondialisation
69 : ADJ paysan, NAM Bové, NAM José, NOM confédération, NOM marge, NOM mondialisation
70 : ADJ rapide, NAM Asie, NOM mondialisation
71 : ADJ rapide, NOM concentration, NOM course, NOM mondialisation
72 : ADJ rapide, NOM course, NOM mondialisation, VERpper engager
73 : ADJ rapide, NOM mondialisation, NOM nécessité
74 : NAM Asie, NAM Etats, NOM mondialisation, NOM nation
75 : NAM Asie, NOM crise, NOM mondialisation
76 : NAM Etats, NOM citoyen, NOM mondialisation, NOM régulation
77 : NAM Etats, NOM citoyen, NOM mondialisation, NOM solidarité
78 : NAM Etats, NOM citoyen, NOM mondialisation, NOM souveraineté, VERinfi défendre
79 : NAM Etats, NOM concurrence, NOM logique, NOM mondialisation
80 : NAM Etats, NOM concurrence, NOM mondialisation, NOM nécessité, NOM régulation
81 : NAM Etats, NOM contrainte, NOM mondialisation, NOM solidarité
82 : NAM Etats, NOM instrument, NOM mondialisation, NOM régulation
83 : NAM Etats, NOM instrument, NOM mondialisation, VERinfi défendre
84 : NAM Etats, NOM mondialisation, NOM nation, NOM souveraineté, VERinfi défendre
85 : NAM Etats, NOM mondialisation, NOM nécessité, NOM régulation, VERpres rendre
86 : NAM Etats, NOM mondialisation, NOM nécessité, NOM solidarité
87 : NAM Etats, NOM mondialisation, NOM nécessité, NOM souveraineté
88 : NAM Etats, NOM mondialisation, NOM nécessité, VERpres affirmer, VERpres rendre
89 : NAM Jospin, NOM dialogue, NOM mondialisation, NOM île
90 : NAM Jospin, NOM mondialisation, NOM mouvement, NOM économie

91 : NAM Jospin, NOM mondialisation, NOM mouvement, VERpres affirmer
92 : NAM Jospin, NOM mondialisation, NOM pression
93 : NAM Jospin, NOM mondialisation, NOM rapport, NOM économie
94 : NAM Jospin, NOM mondialisation, NOM rapport, VERpres affirmer
95 : NAM Jospin, NOM mondialisation, NOM régulation, NOM économie
96 : NAM Jospin, NOM mondialisation, NOM régulation, VERinfi lutter
97 : NAM Jospin, NOM mondialisation, NOM réponse, NOM île
98 : NAM José, NOM confédération, NOM exception, NOM fondateur, NOM mondialisation
99 : NAM José, NOM dialogue, NOM mondialisation
100 : NOM changement, NOM concurrence, NOM contexte, NOM mondialisation, NOM économie
101 : NOM changement, NOM concurrence, NOM mondialisation, NOM nécessité
102 : NOM changement, NOM lien, NOM mondialisation
103 : NOM citoyen, NOM mondialisation, VERinfi répondre
104 : NOM concentration, NOM concurrence, NOM contexte, NOM mondialisation
105 : NOM concentration, NOM mondialisation, NOM processus
106 : NOM concurrence, NOM contexte, NOM libéralisation, NOM mondialisation
107 : NOM concurrence, NOM contexte, NOM mondialisation, NOM pression
108 : NOM concurrence, NOM contraire, NOM mondialisation, NOM régulation
109 : NOM concurrence, NOM différence, NOM mondialisation
110 : NOM concurrence, NOM institution, NOM logique, NOM mondialisation
111 : NOM concurrence, NOM institution, NOM mondialisation, NOM nécessité, VERinfi lutter
112 : NOM concurrence, NOM logique, NOM mondialisation, NOM pression
113 : NOM concurrence, NOM mondialisation, NOM nécessité, NOM régulation, VERinfi lutter
114 : NOM concurrence, NOM mondialisation, NOM régulation, NOM économie
115 : NOM constat, NOM mondialisation, NOM nécessité
116 : NOM construction, NOM institution, NOM mondialisation
117 : NOM construction, NOM mondialisation, NOM île
118 : NOM construction, NOM mondialisation, VERpper engager
119 : NOM contexte, NOM développement, NOM facteur, NOM mondialisation
120 : NOM contexte, NOM développement, NOM mondialisation, NOM économie

121 : NOM contexte, NOM facteur, NOM libéralisation, NOM mondialisation
122 : NOM contexte, NOM facteur, NOM maîtrise, NOM mondialisation, NOM souveraineté
123 : NOM contexte, NOM libéralisation, NOM marge, NOM mondialisation
124 : NOM contexte, NOM marge, NOM mondialisation, NOM pression
125 : NOM contrainte, NOM mondialisation, VERpres compter
126 : NOM contraire, NOM mondialisation, NOM régulation, VERpres rendre
127 : NOM course, NOM dialogue, NOM mondialisation, VERpper engager
128 : NOM crise, NOM différence, NOM identité, NOM mondialisation
129 : NOM crise, NOM développement, NOM mondialisation, NOM processus
130 : NOM crise, NOM identité, NOM mondialisation, NOM sentiment
131 : NOM crise, NOM mondialisation, NOM mouvement, NOM économie
132 : NOM crise, NOM mondialisation, NOM réponse
133 : NOM dialogue, NOM mondialisation, VERpres rendre
134 : NOM défi, NOM libéralisme, NOM mondialisation, VERinfi répondre
135 : NOM défi, NOM maîtrise, NOM mondialisation, NOM processus
136 : NOM développement, NOM institution, NOM mondialisation, NOM pauvreté
137 : NOM développement, NOM investissement, NOM mondialisation, NOM rapport, VERpper publier, VERpres affirmer
138 : NOM développement, NOM investissement, NOM mondialisation, VERpres affirmer, VERpres compter
139 : NOM fondateur, NOM mondialisation, NOM nation
140 : NOM identité, NOM libéralisme, NOM mondialisation, NOM solidarité
141 : NOM identité, NOM lien, NOM mondialisation, NOM solidarité
142 : NOM identité, NOM mondialisation, NOM nation
143 : NOM identité, NOM mondialisation, NOM nécessité, NOM sentiment
144 : NOM identité, NOM mondialisation, NOM nécessité, NOM solidarité
145 : NOM instrument, NOM mondialisation, VERinfi croire, VERinfi défendre
146 : NOM investissement, NOM mondialisation, VERpper engager
147 : NOM investissement, NOM mondialisation, VERpres affirmer, VERpres rendre
148 : NOM libéralisation, NOM mondialisation, VERinfi défendre
149 : NOM libéralisme, NOM mondialisation, NOM réponse, VERinfi répondre
150 : NOM lien, NOM mondialisation, VERpres rendre

151 : NOM marge, NOM mondialisation, NOM nécessité
152 : NOM modernité, NOM mondialisation
153 : NOM mondialisation, NOM nécessité, VERinfi opposer
154 : NOM mondialisation, NOM nécessité, VERinfi répondre
155 : NOM mondialisation, NOM pauvreté, NOM économiste
156 : NOM mondialisation, NOM sentiment, NOM île
157 : NOM mondialisation, NOM solidarité, NOM économiste
158 : NOM mondialisation, NOM souveraineté, NOM île
159 : NOM mondialisation, NOM thèse, NOM économiste
160 : NOM mondialisation, NOM thèse, VERinfi croire, VERinfi défendre
161 : NOM mondialisation, NOM victime, NOM île
162 : NOM mondialisation, NOM victime, VERinfi répondre
163 : NOM mondialisation, NOM économie, NOM économiste

Figure 46 Detail of cliques for *mondialisation* over 2 months: September 1999 and August 1999.

3.2.3.4.2. *Dynamic representations with the Semantic Atlas*

3.2.3.4.2.1. Methods and challenges

Maps were generated for the word *mondialisation* with the ACOM model described in Part.2. These maps were used to work experimentally on the dynamics of representations, and *mondialisation* was used as a test word. All the work in this section has been conducted collectively with Anne-Lyse Renon (graphic design and data visualization), Charlotte Franco (IT developing), Sylvain Lupone (engineer) and Sabine Ploux (head of the team)[178].

Maps for each time period are visualized via a java applet, which calls a program (in C) with four parameters. The parameters are coefficients of diffusion in space (see Ji, Ploux, and Wehrli 2003) that calculate the threshold of integration of contexonyms (co-occurrent words) and children of contexonyms. The program first generates a database of contexonyms and the coordinates for each clique. A series of maps is obtained for each time period. Interpolation allows for a dynamic linkage between the static maps.

3.2.3.4.2.2. Design and multidisciplinary issues

The team asked, as Renon (2010: 94) puts it :

"Est-il possible de qualifier par la dénomination du procédé graphique utilisé (dans le cas de la manipulation de deux formes géométriques) le « déplacement sémantique » existant entre les deux formes? "[179]

With this question the dialogue between linguistics, modeling and data visualization was opened. We asked what the relationship was between the changes in the shapes and the changes in meaning, and whether the first could help read the second.

---

[178] The following work on maps was conducted on the corpus covering 1997 to 2001. Subsequent parts of the corpus were acquired later in the course of this work, thanks to the scholarship given by the Region Rhône-Alpes ("Projet cible 2009"). Therefore, I have not replicated on the full corpus tests that demanded the input of the entire team who partially left the laboratory subsequently.

[179] Is it possible to define the "semantic shift" encoded in the passage from a shape to another via the type of graphic process that is used (when we manipulate two geometrical forms)?

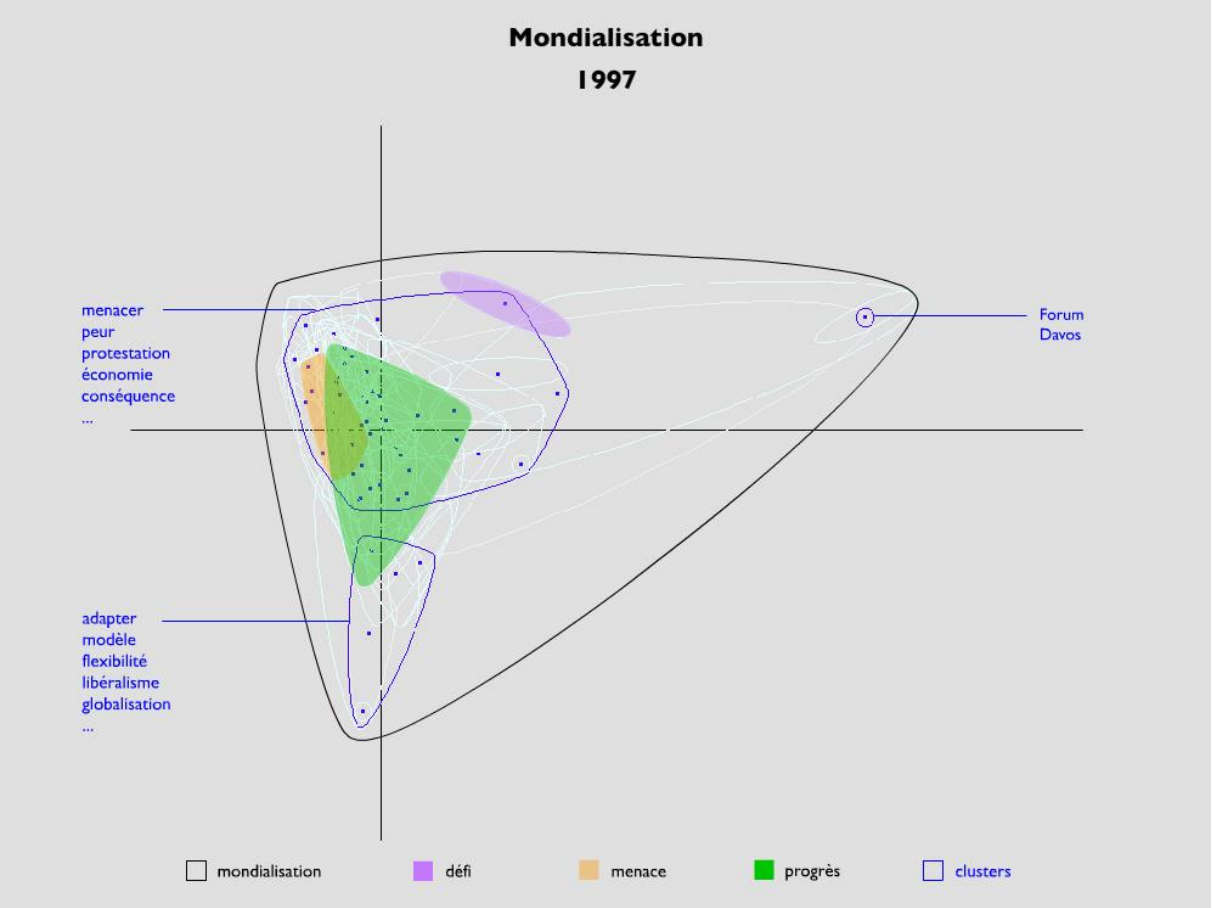
The team observed the structural features of shapes and their combinatorial relationship while operating the transition from two static images to animation. The topological relationship in the juxtaposition of two images was compared to the dynamic paradigm generated by the deployment of shapes. A.L Renon observed the variation in the positions of axes and in the envelope's size and position, and focused on the shape and surface characteristics generated by the geometrical model to understand the structural features at stake. From these observations, the question of the choice of tool to build a dynamic representation, among the tools offered by graphical treatment software, arose. Interpolation was proposed as a simple accessible tool in Graphic Design (in Illustrator or Flash for instance) to match two shapes generated in vector spaces. In the interpolation process, the surface and contour properties of envelopes were used in the mapping of two static maps. Interpolation also brought the question of the status of time and space in the animation process. Moreover, the introduction of interpolation issues from a graphic point of view later triggered further research questions about the types of interpolation available in mathematics (see Xia 2011).

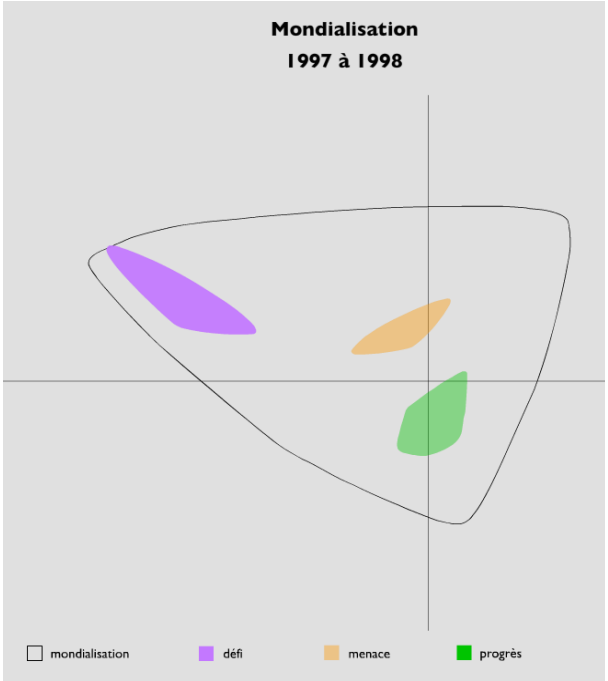
With this approach, a dynamic space was created. Its shape gives a detailed qualitative and quantitative access to the data. The interface and the animation make the reading of complex and heavy data easier, in order to facilitate a global vision of phenomena. As a first experimental step of modeling dynamics of change with the SA, the proposed visualization tried to include questions of modeling, simulation, and more generally issues of spatial representation of data, related to the expansion of digital tools and the computerized handling of data. However, the question of the relationship (if any) between shape variation and meaning variation remained unanswered. Rather, this question served as a common ground between the visualization and linguistic approaches.

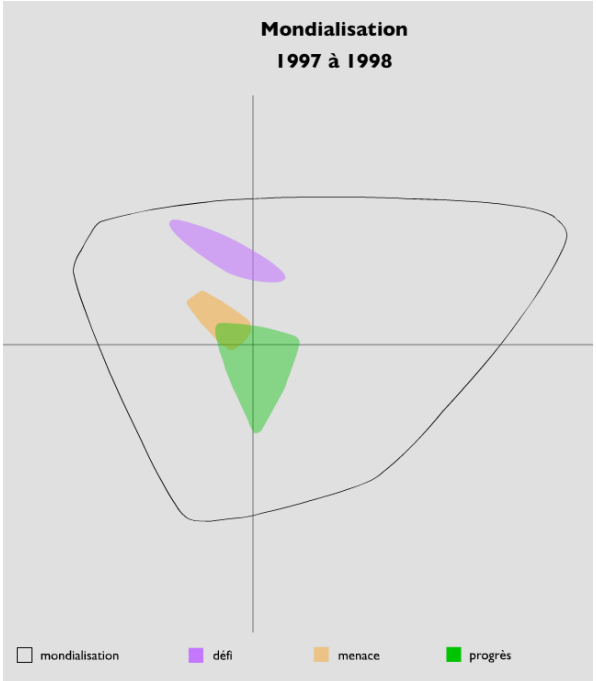
3.2.3.4.2.3. Dynamic maps

The obtained dynamic visualization is online at the following address: http://dico.isc.cnrs.fr/en/diachro.html. The images that compose the following Figure (all grouped under the heading Figure 47) are static snapshots of it, showing maps for *mondialisation* from 1997 to 2001. Key words have been selected manually: *défi* ("challenge"), *menace* ("threat"), *progrès* ("progress") and *alternatif* ("alternative"). These words were chosen since they are representative of the changing contexts of use. Their envelopes are colored to highlight word behavior in the maps. The roughest degree of

organization of contexonyms are the clusters, delineated in blue, and tagged with samples of representative words. Clusters carry whole, coherent themes. The maps show how *mondialisation*'s meaning is changing from general to specific, notably due to the creation of derivational neologisms in this period. The global envelope of the word also changes in shape and density, showing enrichment, and widening. The term widens but also specializes over a few years to include some of the ideas related to the *altermondialiste* taken in its wider definition. These ideas are based on the perception of the concept *mondialisation*, connoting it with fear when it is seen as a challenge and a threat. Conversely we have the positive idea of collective progress, in the spirit of the global village. Clusters sometimes group words that will be opposed in other maps, showing how the same sets of concepts feed opposite perceptions.







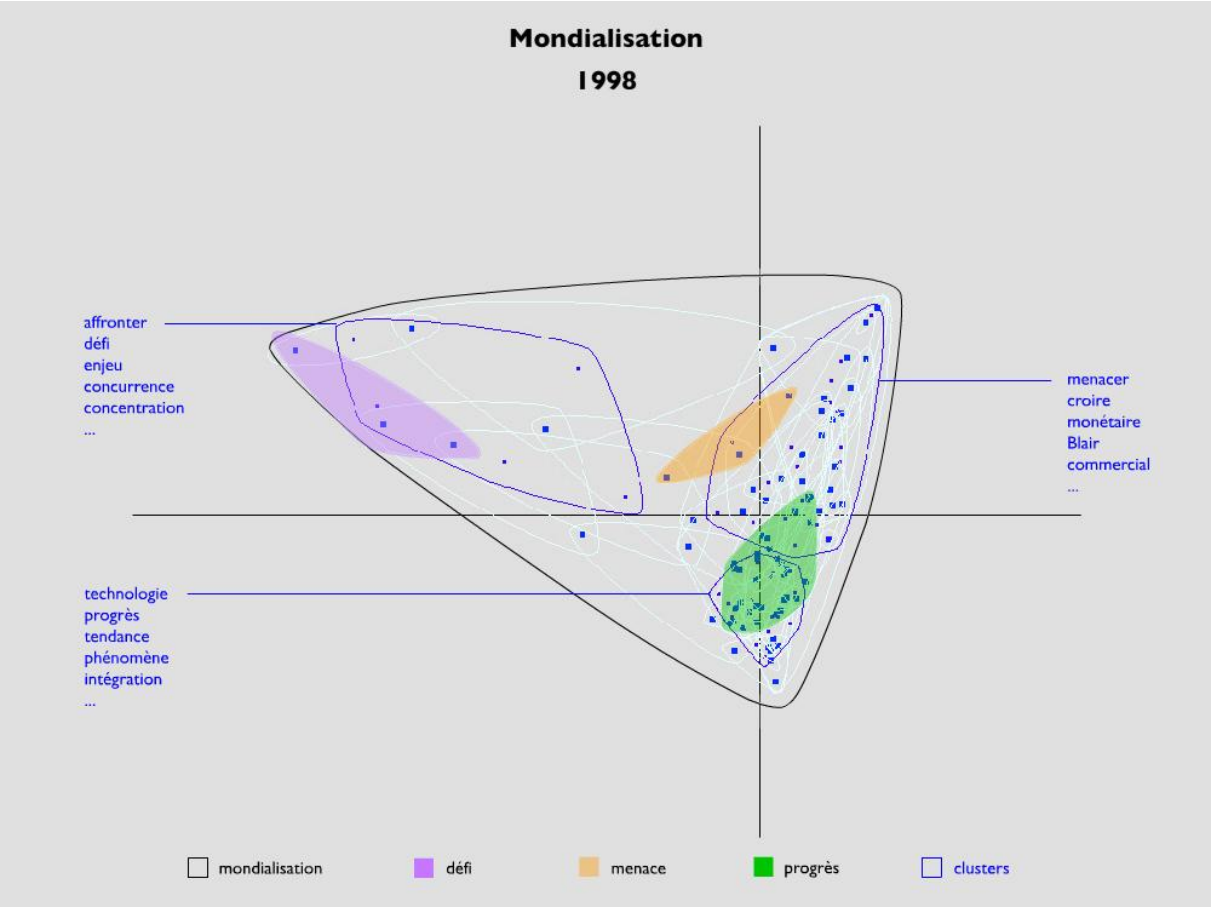
In 1997 the three thematic clusters are:

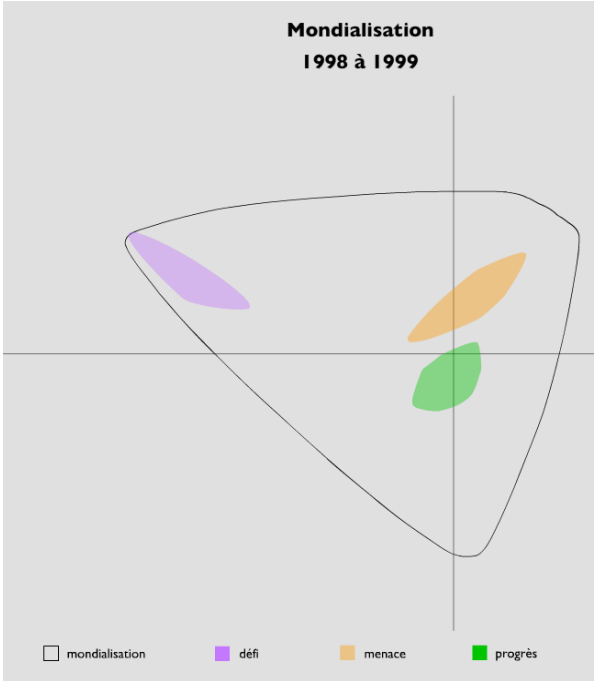
1. Menacer, peur, protestation économie, conséquence…
2. Adapter, modèle, flexibilité, libéralisme, globalisation…
3. Forum, Davos180

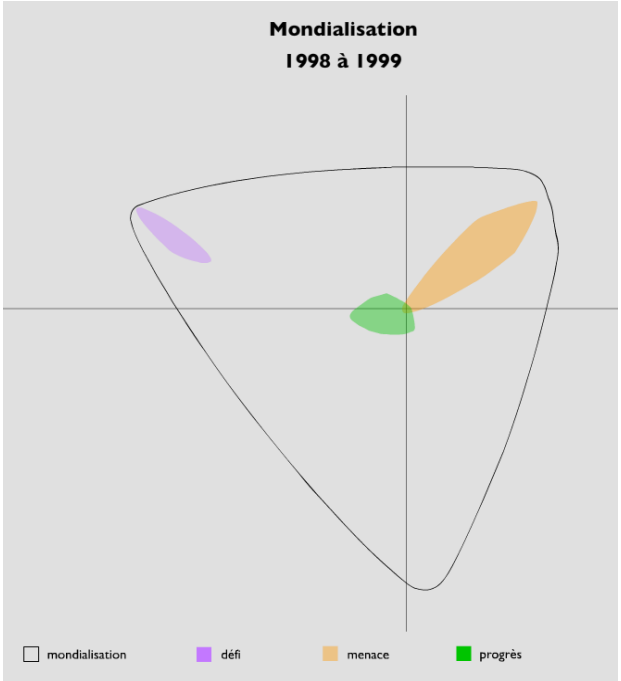
The envelopes of *progrès* and *défi* are in the same cluster. The envelopes of *progrès* and menace are superimposed and progrès is part of two clusters (1 and 2).The envelope of *progrès* is central, covering the origin of axes and has a significant size. This map shows how the contexts of use of *mondialisation* are structured around the notions of challenge (*défi*), threat (*menace*) and progress (*progrès*). The idea of adapting society to an economic model is mixed with words reflecting ideological positions about how to adapt. The third cluster, inclusive of the economic forum of Davos is also included in the envelopes of *planète* (“planet”) and *économie* (“economics”) in grey. On the map, the contextual network shows

[180] 1. Threat, fear, protest, economy, consequence...
2. Adapt, model, flexibility, liberalism, globalization ...
3. Forum, Davos

that *mondialisation* is perceived as a challenge that is seen either as a threat or as the trigger of progress. Several economic and political models are discussed in the corpus and several positions are defended. Between 1997 and 1998, the three envelopes of *défi*, *menace* and *progrès* separate to join three different clusters in 1998, almost resulting in two clusters based on the notion of threat and one on the notion of technological progress. The envelope of *progrès* diminishes whereas *défi* widens.







In 1998, the three clusters are composed as follows:

1. Affronter, défi, enjeu, concurrence, concentration…
2. Menacer, croire, monétaire, Blair, commercial…
3. Technologie, progrès, tendance, phénomène…[181]

Menace and *progrès* both participate in two clusters (2 and 3) and are both close to the center, whereas *défi* is in the biggest cluster (1), away from the origin of axes. It results in a sort of

[181] 1.Confront, challenge, stakes, competition, concentration

2. Threaten, believe, monetary, Blair, commercial

3. Technology, progress, trend, phenomenon

triangle, in which, in front of *défi*, that connotes the major cluster, *menace* and *progrès* are positioned as two opposed contexts, or two possible perceptions of the challenge. *Menac*e is associated to the positions assumed by Tony Blair, perceived as threats, while *progrès* is associated to technological progress. Between 1998 and 1999 the size of the envelope of *menace* increases while those of *progrès* and *défi* diminish. *Menace* and *progrès* have opposed positions in space, on each side of the vertical axis even though they both belong to the main cluster in 1999.

In 1999, the three clusters as composed as follow:

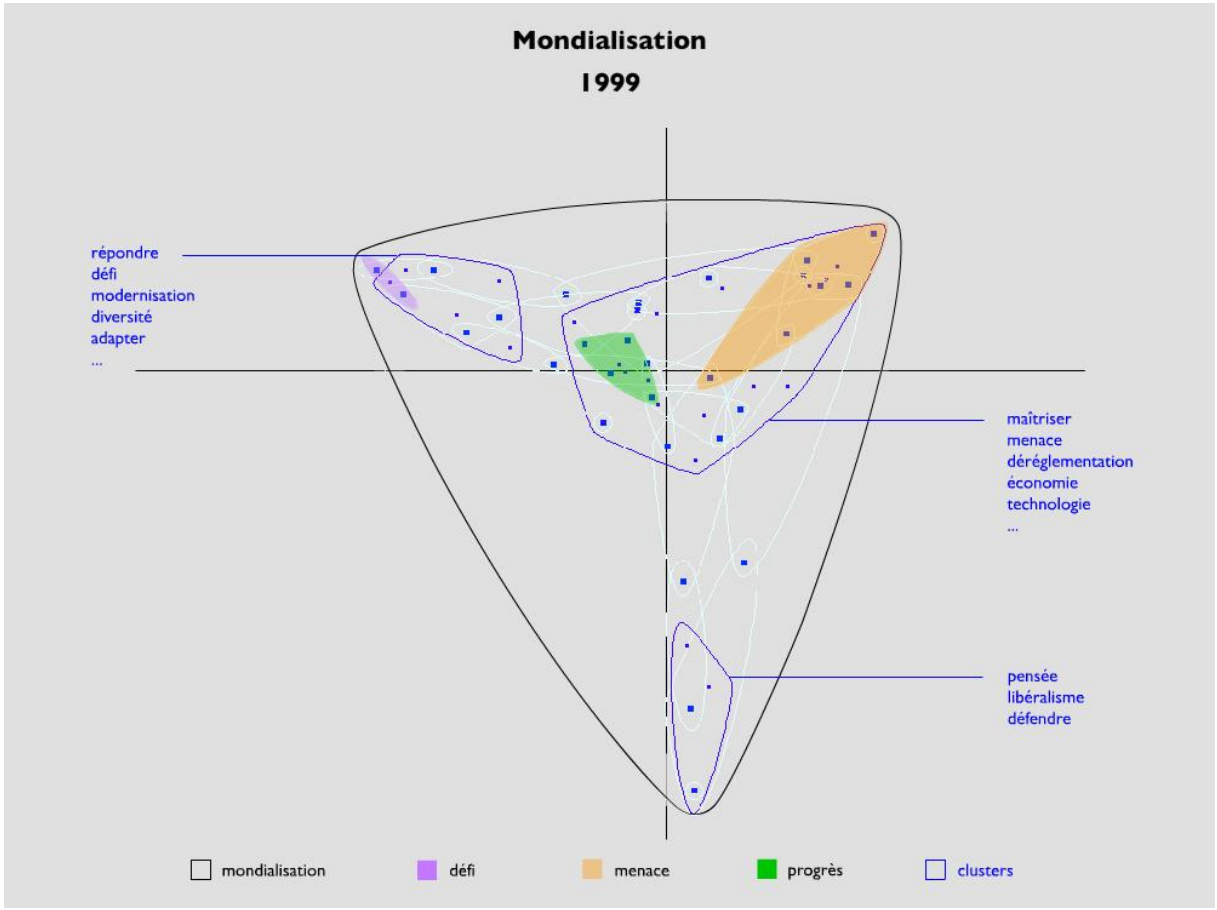
1. maîtriser, menace, asiatique, déréglementation, économie, technologie…
2. répondre, défi, modernisation, diversité, adapter…
3. pensée, libéralisme, défendre[182]

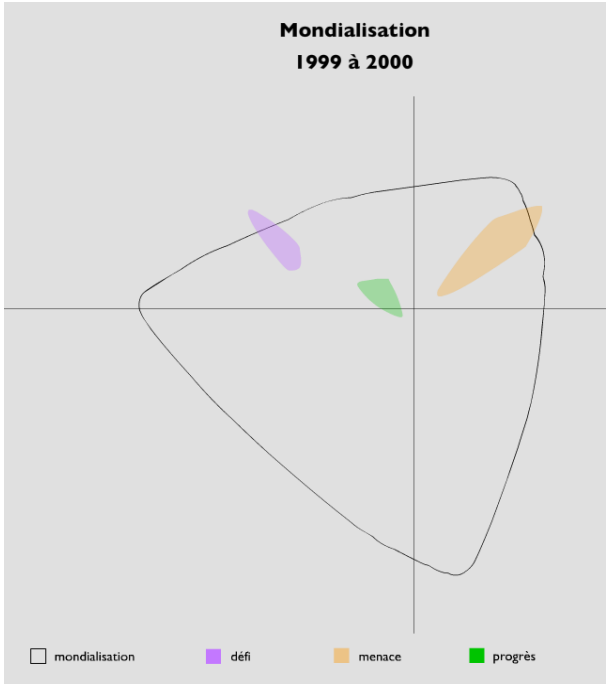
The first cluster mixes the idea of technological progress and threat of the Asian economy. The second one revolves around the idea that the challenge is to adapt to modernity. The third is structured around the defense of liberalism. The contexts of use are organized around the idea of confronting the challenge posed by the technological progress of Asia and the USA (in grey), a progress achieved via a certain form of liberalism.

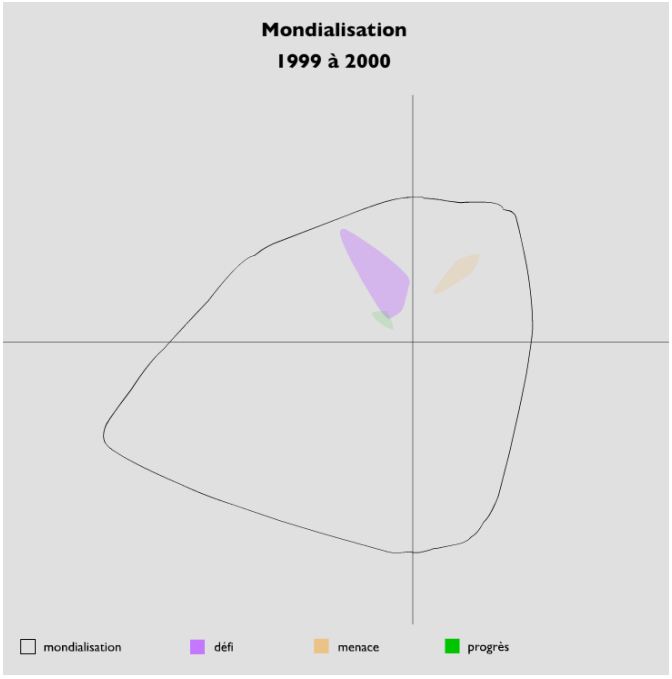
[182] 1. Control, threat, Asian, deregulation, economy, technology

2. answer, challenge, modernization, diversity, adapt

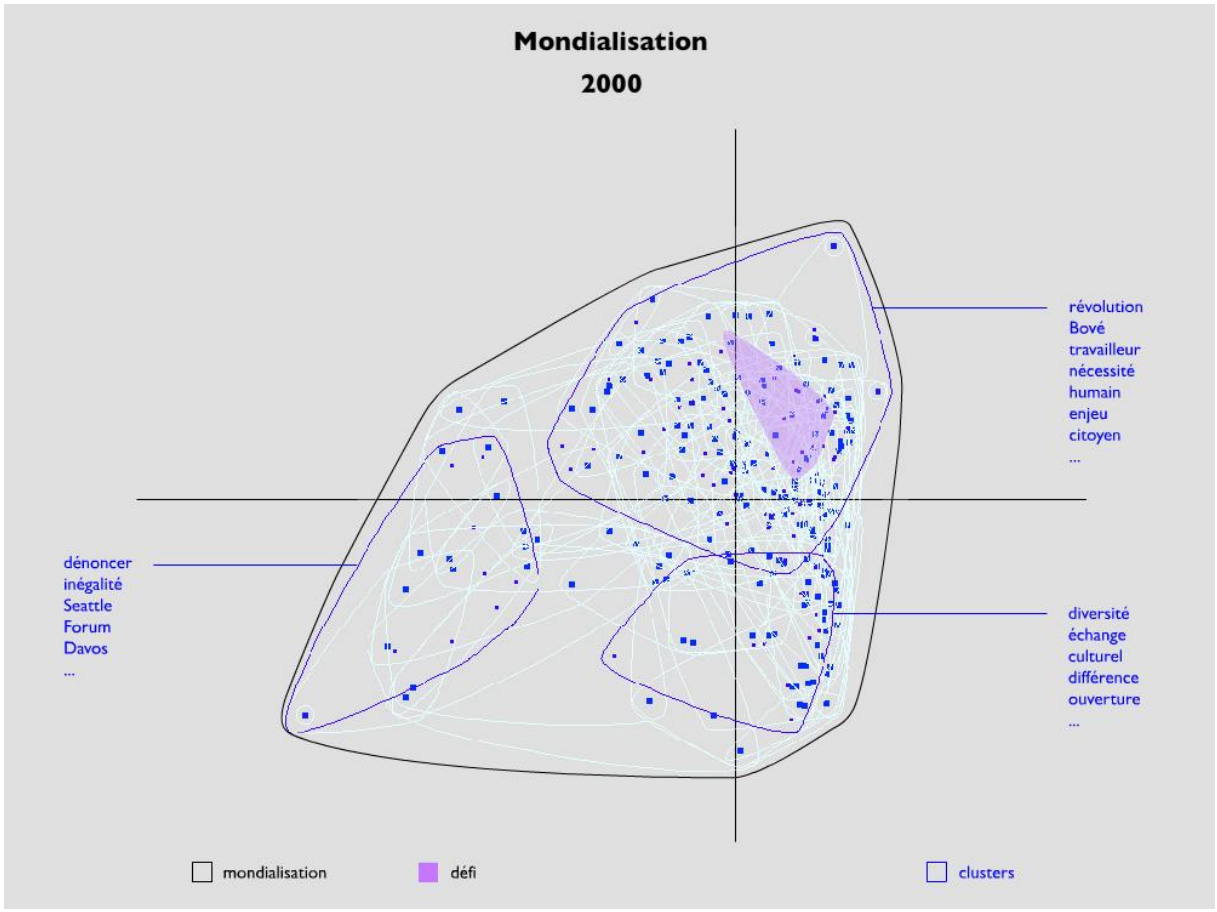
3. thought, liberalism, defend

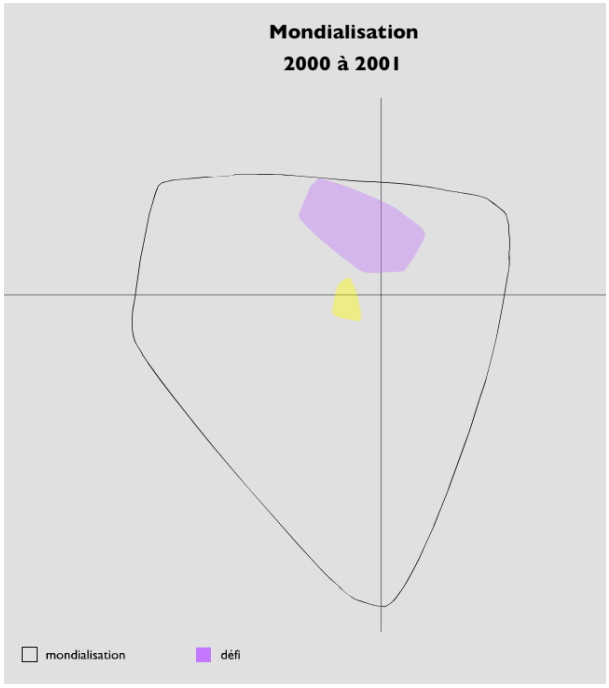

Mondialisation
1999
répondre
défi
modernisation
diversité
adapter
...
maîtriser
menace
déréglementation
économie
technologie
...
pensée
libéralisme
défendre
mondialisation
défi
menace
progrès
clusters

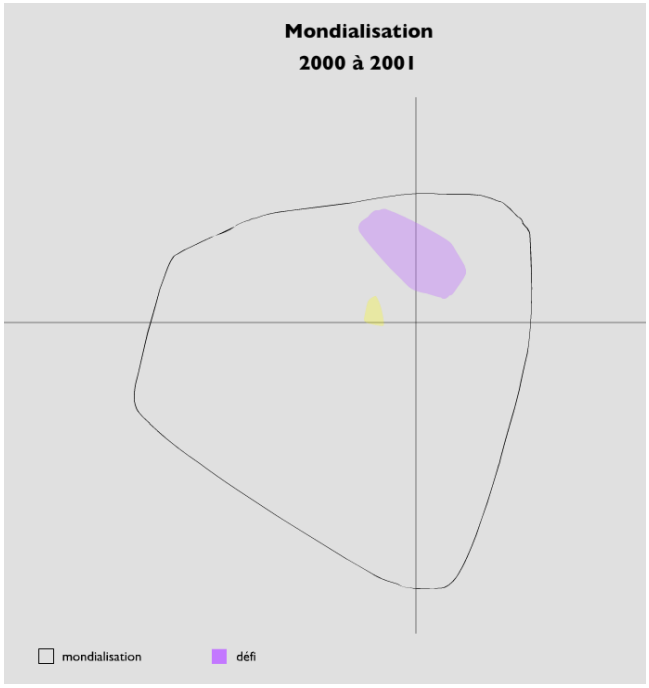

Mondialisation
1999 à 2000
mondialisation
défi
menace
progrès


Mondialisation
1999 à 2000
mondialisation
défi
menace
progrès

Between 1999 and 2000 the terms *menac*e and *progrès* disappear from the map while *défi* occupies a central position and increases in size







2000 is the map of change. Density explodes while the general envelope's shape changes drastically. It is precisely in 2000 that *antimondialisation* enters the corpus. Three important clusters appear:

1. Révolution, Bové, travailleur, nécessité, humain, enjeu, citoyen…
2. Dénoncer, inégalité, Seattle, Forum, Davos…
3. Diversité, échange, culturel, différence, ouverture…[183]

The two biggest clusters contain vocabulary related to anti-globalization and to events organized by its advocates or the establishment against whom they are acting. The third cluster possesses new ideological content with values such as openness and diversity. Between 2000 and 2001, *défi* remains central and keeps its size. The term *alternatif* comes into play in 2001 in the center and is linked with the two biggest clusters.

In 2001, the clusters are :

1. Alternative, message, construire, militant, écologique…
2. Conséquence, développement, économique, Seattle, citoyen…
3. Paysan, Bové, José, confédération[184]

The map is entirely articulated around anti-globalization. The biggest cluster is anchored in the context of the anti-globalization message; the second cluster gathers political and economic ideas that include fear, opposition, or thought while the third cluster is restricted to anti-globalization and one of its advocates.

[183] 1. Revolution, Bové, worker, necessity, human, issue (stakes), citizen
2. Denounce, inequality, Seattle, Forum, Davos
3. Diversity, exchange, cultural, difference, opening

[184] 1. Alternative, message, build, activist, ecological
2. Consequence, development, economic, Seattle, citizen
3. Farmer, Bové, José, confederation

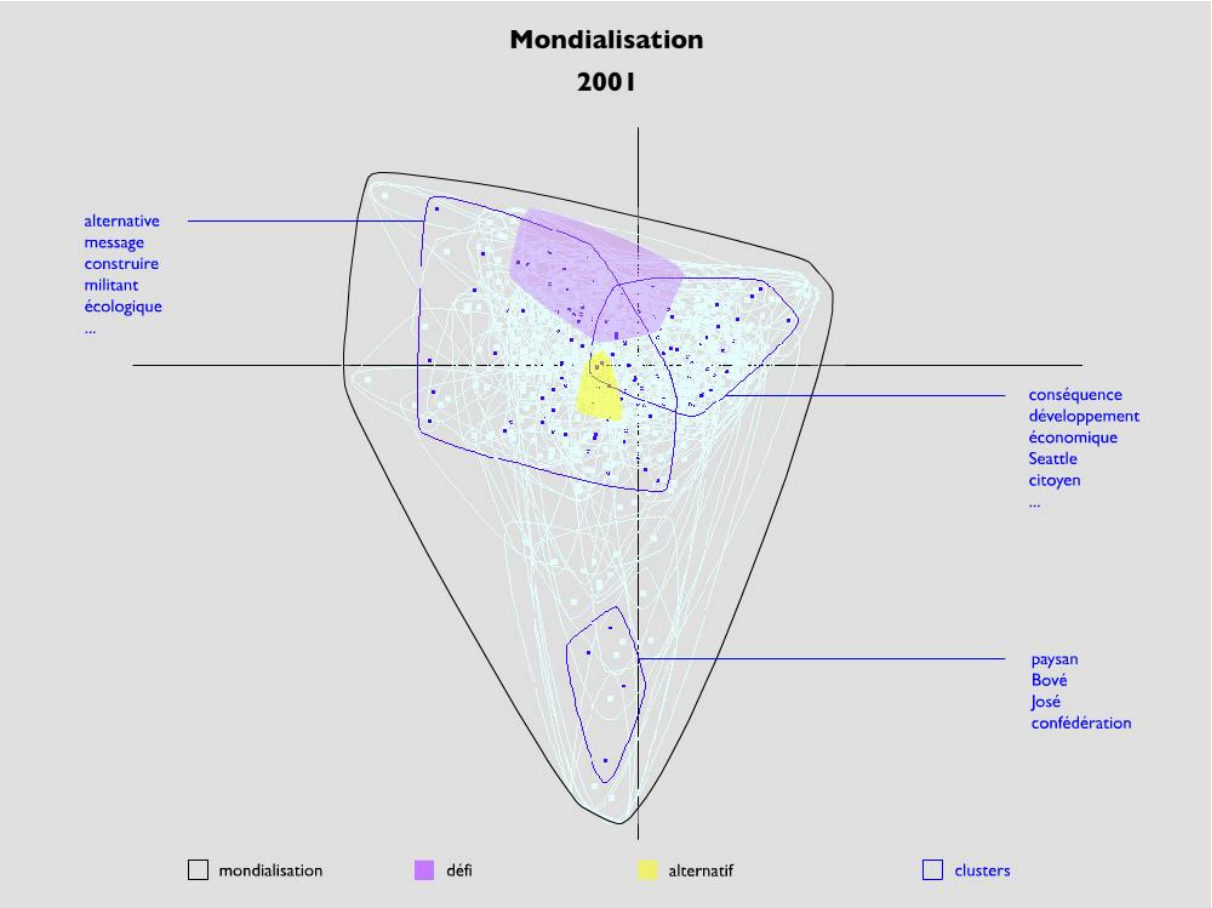


Figure 47 Snapshots from the dynamic visualization of *mondialisation* from 1997 to 2001

Therefore, between 1997 and 2001, a clear evolution is shown in the map: from a map with only a few references to anti-globalization in 1997 to one totally centered around it in 2001. The 2001 map is composed of alternative propositions at the ideological and political levels and mobilizing specific actors and events. The contexts of use of alter- and anti-*mondialisation* and their context of birth color those of *mondialisation* to the point of redefining its context map. *Mondialisation* specializes in a more political, economic and social use, in reaction to liberalism, while the word *globalisation* does not show such change.

3.2.3.4.3. ***Démondialisation***

The productivity of *mondialisation* has not said its last word yet. In November 2010 the motto of *démondialisation* appears in the press, associated with the politician Arnaud Montebourg, who was campaigning at that time. *Le Monde* called him "Arnaud Montebourg,

nouvel apôtre de la "démondialisation"[185] (Arnaud Montebourg new apostle of "deglobalization"). Montebourg is described as the new apostle of deglobalization in the same way that Bové was described as the spokesman or hero of the struggle against junk food (*malbouffe*). The concept of deglobalization is at the center of the French electoral campaign of 2012, with regards to the extreme left and some leftist parties. The concept is borrowed from a sociologist: Bello (2005). It is relayed in French in an influential book by Hessel and Morin (2011)[186] who transpose En. *deglobalization* into Fr. *démondialisation* (and not *déglobalisation*). In this book, numerous concepts give birth to derivations in dé- such as *dérobotisation* ("derobotisation") to refer to employees being treated as robots and needing to free themselves from this condition, *débureaucratisation* ("debureaucratisation") to refer to administrations, and the use of the attested verb *décloisonner* ("to decompartmentalize"), among others.

It seems that the new motto of some left wing politicians, repeated on a regular basis from November 2010 to today, has several "parents": a sociologist, two philosophers and a politician: Bello, Hessel and Morin and Montebourg. Moreover, away from the front scene media, several activist groups have been using prefixation in *dé-* before, as in the idea of *décroissance*. The *démondialisation* term shows a relatively stable use until today in the media,[187]and reinforces the hypothesis outlined in the study of *mondialisation.*

Mondialisation has therefore shifted from economics and finance, domains in which *globalisation* has specialized in turn, to politics. It then became a key word for a specific political movement and underwent pejoration in this context. This process gave birth to an opposition to globalization that produced neologisms built on *anti-* and *alter- mondialisation*. The positive idea of a world village in which everyone can communicate was overstepped by the negative idea that the globalized economy and system are a threat to people's well-being.

---

[185] Title of an article in 20 November 2010

[186] Used as a reference by the activists groups "les indignés".

[187] As an indication, 187 articles in Le Monde contain the term, from the 20th November to the 7th December 2012.

This process gave birth to the idea of "undoing" globalization, with the idea of *démondialisation* found in intellectual circles and later in mainstream politics.

The mechanisms at stake involve domain shifts, specialization, synonymic competition, (morphological) productivity and pejoration.

Chapter III.3 : DISCUSSION

These studies show how the interrelation of formal and semantic neology, morphological productivity and connotational drift in semantic change can be used to achieve a more holistic framework of detection and analysis of semantic change in corpus. They also show that the categories described by the literature are operational, but that they work in combinations. However, the proposed approach shows a certain weakness regarding the combinatorial aspect (at sentence level) which is managed by the model to some extent. Ideally, combinatorial patterns could benefit from more input and interest. The chosen indices are a good basis for semantic change analysis since they provide several levels of granularity in the outputs.

3.3.1. Levels of granularity

The indices are comparable to a set of lenses through which one can see the data. They provide information at different levels, in terms of granularity. Granularity is related to the level of detail, but also to the type of information. When granularity is independent from typology, and is rather a way of dividing meaning patterns into scalar layers of detail, the analysis can benefit from zooming in and out of these layers. Exploiting granularity is like collecting several snapshots of data with different amounts of detail and taken from different angles. The following indices and methods correspond to different levels of granularity:

1) Frequency alone (raw and normalized frequencies, regression coefficients on normalized frequencies, coefficients of variation) show general trends in use, of sociological and historical value, which may contain semantic and stylistic information.

2) The co-occurrence frequencies for a target word show its semantic network, providing information about semantic fields and contexts of use, at the lexicological and pragmatic levels.

3) Normalized frequencies of a co-occurrent word with the target word, ranks, and the density and cohesion index, show the structure of this network in time and show semantic information at a deeper level.

4) Maps obtained with ACOM, show three levels of granularity simultaneously: the clique, the word and the cluster, within a unique space. Trends, use, and structure are brought together. Interpolation gives a dynamic dimension to the maps.

Therefore the tool box seems satisfactory to grasp several layers of granularity.

3.3.2. Merging hypotheses from the literature

The mechanisms described in the literature are observed in combinations. The major mechanisms are:

-The transgression of lexicological and morphological rules (*malbouffe, mal-, bio-)*

-The imitation of attested mechanisms: for instance, confixation (*mal-*) and in general the imitation of word formation patterns of existing words in compounds

-Cross-linguistic phenomena (*crypto-*, *cyber-*, *bio-* for instance *bioplastique/bioplastic, crypto-catholic/ crypto-catholique…*)

-The presence of borrowings and calques, mostly from English (*globalisation, cybernétique…)*

-The shift in domains and addition of domains (*crypto)-*, via specialization (*mondialisation, globalization*) and by extension (*bouquet, bio-*)

-Synonymic competition (*mondialisation* with *globalisation. Biocarburant* with *biocombustible, biofuel, etc.)*

-Variation across linguistic communities and specialized linguistic communities (*bio-)*

-"paternity" processes (*malbouffe* with De Rosnay and Bové. *Mondialisation, antimondialisation, altermondialisation* with extreme left politicians and *démondialisation* with Bello, Hessel and Morin as well as Montebourg)

-Polygenesis (*bioplastique*)

-Ambiguity (*bio-)*

-Productivity of items involved in semantic change and connotational drift

(*Mondialisation* -> anti- (3), alter- (3), dé- (1)

Globalisation -> anti- (1)

Terroriste -> several compounds (51 among which *bio*- and *cyber*-))

-pejoration: *mondialisation*

The studies show a high number of idiosyncrasies, which seem to add to the creativity of other processes when they are part of sets (as in words based on *mal-*, on *terroriste*, and *bio-).* Specialized terminology also shows an extremely productive reservoir of neologisms. The rules of morphological word formation are generally more lax, maybe due to foreign influence (in this case English rules of word formation). Idiosyncrasies and neologisms show spelling variations and use of punctuation. New words seem to follow cyclical production and use; however, the corpora are too small to confirm this tendency statistically.

The behavior of *mondialisation* and *globalisation* also follows Bréal's (1899) definition of differentiation, since the two near synonyms diverge and specialize. They also correspond to what Blank's (1999) typology describes as sociocultural changes at the root of conceptual shift for existing words. Phenomena observed in *bio*- and especially the ambiguity between *bio-(*4) referring to the "organic" meaning and *bio-(*5) referring to ecological products, seem to confirm Blank's idea that speakers tend to integrate lexically isolated words in related classes, via popular etymology. Domain shifts, extensions and additions seem to operate at many levels: *bio*- adds two domains to its meaning (technology and ecology), *crypto*- adds religion, politics and social aspects to the domains of the natural sciences and politics, *mondialisation* becomes more political while *globalisation* specializes in the financial liberal domain. Register shifts also take place as *malbouffe* progressively exits the familiar register. It is also striking that new forms and new meanings, when they add to an already rich network, generally mimic other pre-existing combinatorial or semantic patterns (as seen with *malbouffe*, and in compounds).

3.3.3. Hypothesis of semantic plasticity

On the basis of the observation that some words have more flexible combinatorial ranges than others, the question of the impact of that flexibility in diachronic processes was raised. I formulate the hypothesis that, if a word already possesses a high degree of semantic *plasticity*[188] in synchrony (i.e., it appears in numerous contexts under numerous meanings and has a large panel of co-occurrent words), there are more chances that new meanings and combinations appear over time. The existing diversity in meaning is a substrate for new meanings to integrate the existing semantic structure of an item. That diversity is graspable in terms of semantic, combinatorial, polysemy and idiomatic features. For instance, with *bio-* , the fact that meaning (2) involves the relationship of biology with other sciences opens a huge panel of combinations in meaning. This panel is multiplied since *bio-* is an element of composition. In this sense, *bio-* has a high plasticity.

It is psychologically more acceptable for a speaker to accept the further enrichment of an item or word that is already rich in its current use. The fact that the word or item already appears in a multiplicity of contexts may facilitate its use in a new context or the extension/restriction of pre-existing lexical units and contexts and the mapping and transfer processes with other connected meanings. To observe how new contexts of use enter the existing structure, cliques are a very detailed tool. With *mondialisation*, I showed how the cliques provide insights about cohesion and density. As the new contexts emerge, they feed and change the existing network of associations, and progressively find a role in the meaning structure, leading to stabilization, after the divergence from the synonym has been overcome. Within that structure, a few associated words (co-occurrent words) operate as pivots.

3.3.4. The role of pivot words and concepts

The rhetorical figures involved in semantic change do not operate at the flipping of a coin but rather in slow gradual processes. Even a reversal figure, such as pejoration, shifts through intermediary stages of semantic drift. This happens, as I argue in this work, most often via a pivot concept and/or word. The example of *silly*, taken from the literature, shows a shift from "happy" to "stupid" via the notion of bliss. Blissful looking individuals may have been

[188] My coinage.

socially judged as idiots on the basis that there are notions in common in bliss and stupidity, notably looking happy whatever the situation, even if it should involve other emotions for the majority of individuals. This is echoed in the saying "blessed are the poor in spirit" (the Fr. "heureux sont les simples d'esprit" uses the equivalent term to "happy" in French). Steinmetz (2008: 207) mentions the pivot notions of "innocence" and "weakness" since *silly* went through an intermediary stage in meaning as "spiritually blessed, pious, holy, innocent, harmless" before it shifted to "deserving pity" and subsequently to "weak, feeble, insignificant" and finally "stupid, foolish." In the same way, the etymology of *eccentric*, shifted from astronomy in which it means "a circle in which the earth, sun, etc. deviates from its center" to referring to people "deviating from the usual character or practice; unconventional; whimsical; odd" (Steinmetz 2008: 59). Here the pivot notion of "deviation" shifts from literal to figurative, and therefore may metaphorically apply to people instead of orbits. The idea of "circle" however loses strength. I argue that in detecting semantic change, these pivot meanings are of great help. Pivot words and concepts and pivot networks of words and concepts may be part of the network of co-occurrences of the target word. As such, pivots provide information as to the reorganization of the semantic structure of a word's network. For *mondialisation*, the roles of *contre* was shown, for compounds in *bio-* the roles of *biologique* and *biotechnologie* were shown.

3.3.5. Plasticity, fluctuation and semantic change

Ultimately, fluctuation in synchrony and change in diachrony are linked, since the latter develops on the structures laid by the former. However, while computing networks in corpus, the variability that existed for a long time (fluctuation) and other types of variability are put on the same plane by the model. Plasticity and "natural" variation (fluctuation) are dealt with simultaneously with semantic change.

Measures derived from variation analysis (especially the coefficient of variation) have therefore no linguistic value if the purely automatic output is considered alone. With our current means, only speaker intuition and the human knowledge of language and the world may provide the necessary distinction between established richness and new richness. This step of the analysis could benefit from modeling the polysemy and idiomatic range of words, and evaluate their plasticity beforehand. Therefore, the detected variations could be analyzed in terms of how much they differ from "natural" variation (plasticity and fluctuation) for each

item. Measures of plasticity and fluctuation could therefore be good starting points in future models. Ideally, the combinatorial possibilities of words should be taken into account as well as their natural fluctuation in use.

3.3.6. Low frequency words

If frequency alone is not a sufficient indicator, it does contain some useful information. According to Blank (1999), low frequency is an indicator of complexity: the more frequent a word is, the less complex it becomes semantically. If complexity is understood as high plasticity and richness, then semantic change is likely to take root in words of low frequencies. Therefore, low frequencies are a good filter to detect neologisms, free-forms, creative uses and candidates to semantic change as was shown in case studies. Words of high or average frequencies may undergo deep changes too. However, they may do so on a bigger time scale. In this case, detecting them relies on analyzing gradual connotational drift as with *mondialisation.*

3.3.7. Interrelation of processes

It is my contention that the introduction of a novel element (a formal neology) in the system (language) implies a redistribution of semantic contents across lexical units and words that are related to it, via synonymy, co-occurrence, and semantic field or domain. Formal neology is easily detectable with frequency counts, and therefore frequency is not discarded as a useful index even if many studies have shown that frequency alone is not a reliable index to study semantic change. The neologism *malbouffe* showed to be related to a more general trend in *mal-*.The rules of word formation of *mal-* showed to be changing, along with a subtle meaning drift. The questions raised by compounding in *mal-* found an echo in other elements of composition, showing similar behaviors towards word formation rules, and ambiguity of the element of composition. The numerous neologisms in *bio-* showed that semantic change is taking place at the level of the element of composition and not only at the level of words. The neologisms in *anti-*, and *alter-* based on *mondialisation* showed an impact on *mondialisation*'s own context maps, connotation and retroaction. The case studies show that processes of formal and semantic neology are related, and that in general most processes of semantic change take place in combination and across different levels within language and outside language. Morphological, semantic and sociological patterns seem intertwined and

benefit from being dealt with in a single framework. At the heart of these interactions is the deployment of polysemy in time.

Conclusion & Perspectives

A mirror of society

For centuries, in the era of pen and paper, the pace of word meaning change seems to have been stable, slow and gradual. This assumption is widely shared by various researchers, even if no complete lexicostatistical study of word meaning change is available.[189] With the advent of the information society this pace is accelerating (and once again there is no clear measure yet to justify this assumption). Meaning change seems to happen quicker than it ever had since the Internet became widespread and accessible to more people. For a few decades, we have been witnessing the confrontation of the language system with mass communication. The state of the language system depends on the conditions in which it evolves. It adapts to its conditions. The lexicon of a language is organized in a complex semantic and associative system. Looking at how this system constantly reorganizes itself provides a mirror of civilization. Therefore, understanding semantic change is a precious tool to understand the evolution of the language-society paradigm. Major trends appear in the press corpus: they describe events, opinions, and debates that mark society. Is it bad to eat junk food? Are GMO's junk food? What is terrorism? Is globalization good or bad for society? What is the difference between natural, controlled and artificial life? These types of questions emerge across the corpus showing that semantic change in short diachrony in the information society today cannot be dealt with separately from the social reality generating them[190]. In this sense, this work is a small contribution to the history of words, concepts and ideas as well as a contribution to sociology.

---

[189] This is mostly due to the fact that semantic change is often dealt with as a subpart of linguistic evolution. In historical linguistics, stability is also referred to as *consistency*. Most theories about stability include much more than semantic consistency, and encompass syntax and morphology in cross-linguistic perspectives (see Nichols 2005, for instance).

[190] There is also a part of responsibility in the choices made by the media.

Towards a more holistic approach

If the mutability of language taken as a system is increasingly quick, it also generates areas of greater density in terms of contact, creativity and productivity as well as flexibility. The stories of target words shed light on some of these areas, in which a variety of semantic change phenomena participate and influence each other. It has been shown that most mechanisms of semantic change described by the literature are intertwined, and benefit from being dealt with together rather than separately. Confronting the approaches of several disciplines (linguistics, mathematics, NLP and data visualization) also proved to be relevant to deal with the complexity of semantic change in a more holistic perspective.

Most of the presented mechanisms of change have been described by the literature. However, the way these mechanisms are connected and interact is hardly described as a system in the literature. The case studies showed that different types of mechanisms benefit from being studied together. For instance, formal and semantic neology operate hand in hand, within the same word (*malbouffe*) or across several words (in *mondialisation* with *anti/alter-mondialisation*), synonymic competition operates with morphological productivity (*mondialisation* vs. *globalisation*), neology with morphological productivity, ambiguity and laxer morphological rules (*bio-).* No process is isolated.

Perspectives

The computational tools used in this study proved to be effective to browse large amounts of data without losing details. These tools can be understood as a basis for further research. The detected patterns too are a small step for further exploration and they open a perspective in modeling: patterns can be modeled to be fed back to corpora in a more detection-oriented perspective. Case studies were one of the ways of defining patterns and indices so that they can be later "fed back" to the whole corpus to search for matches. Modeled patterns, relying on statistics and geometry, could be used to detect all items that show a similar statistical (and maybe geometrical) behaviour over the whole corpus. They could also be used across different scales in other corpora of varying size and time coverage to test whether they could adapt.

However, the indices which are tested are relevant when combined with analysis and post-treatment. These indices too are to be taken as a perfectible tool box.

However, these tools need refinement and optimization. One of the promising paths that emerged is the combined use of ranks and coefficients of variation (in the vein of the tools created by Holz and Teresniak (2010), another is the calculation of variations within a co-occurrence network over time. For instance the coefficient of variation for a target word with its whole network or by pair with just one co-occurrent word was generated as well as regression coefficients on the same basis. Preliminary exploration was conducted on whole networks, and showed interesting avenues of research, however, at this stage the need for optimization and heavier language engineering emerged and could not be met. Specialists of computational semantics with a strong background in optimization could, however, apply these calculations at large scale. Pushing the calculation of rank variation within co-occurrence networks seems to be a promising path. Looking at the changes in the network of a target word, in terms of binary associations with its co-occurrent words evaluated by the regression coefficients, the coefficient of variation and the rank could provide a rich output, to be analyzed semantically by a linguist. Also, the density and cohesion variability index would benefit from large scale testing. Tentative tests were conducted but a full-fledged study of the index could substantially enhance the described tool box. Moreover, these tools could benefit from a probabilistic approach to add a layer of prediction in the prototype. Ideally, this tool box could be optimized and made available to a large audience.

The described tool box could serve as a basis for an open access platform, giving the freedom to users to combine tools. Depending on the type of data one wants to analyse, a different combination of filters may be necessary. In this vision, text can be compared to a landscape one takes a picture of. A color picture will show a certain amount of data. An infrared picture will bring out the thermal data, although it is monochromatic. Both pictures may help understand the landscape from complementary point of views. In the same way, filters and indices provide a series of snapshots of certain aspects of the data. Combining the different types of snapshots gives a complete view of the data.

The SA is already a platform used by numerous people. It is a good basis to build an open source user interface, in which users could upload their data and browse it with the tool box. Since the SA and ACOM already provide meaning maps, these maps could be enhanced with

a temporal dimension, providing dynamic graphs and maps to users, in the vein of those presented in Part III. With interpolation, we are putting those snapshots one after the other, to get from the picture to the video, while integrating the temporal dimension. In this perspective, anyone could visualize the evolution of a given word, idea or concept in any given text database.

The other perspective the model offers is its adaptability to dynamic corpora extracted from the Web. The progress achieved by Web crawlers in the past decade should make these tools available to a larger number of people in the near future.

These ideas encounter the limits of available calculation power. However, these limits are constantly reduced as computer power increases. I am confident that with the high speed development of technologies, there will be future frameworks to include the results of this type of research in collective paradigms such as distributed computing (in the spirit of DistributedDataMining[191] for instance), rather than centering heavy calculations in a few powerful dedicated servers that shall soon be overloaded.

Contribution

In the academic world, the contribution of this type of study impacts not only the knowledge involved in various disciplines but also a certain number of tools necessary to carry out research. These tools include databases, indexes and search engines for paper or electronic libraries, and other tools involved in knowledge organization and filing. For lexicologists, terminologists and lexicographers, overwhelmed by the amount of data to analyze, new tools are needed to be able to explore language on the Internet and decide which terms should enter a dictionary. Electronic dictionaries can benefit from semi-automatic updating. Diachronic distributional semantics are also extremely useful for translators as they provide information about the semantic distribution of meanings across languages and therefore help map concepts in several languages and follow the evolution of this mapping with precision. Currently, the SA is used in bilingual versions by translators. I believe that tools derived from it in diachrony also have this potential, as well as other online and electronic tools. In NLP in general,

[191] http://www.distributeddatamining.org/

updating is an issue. A lot of programs, databases and corpora are updated and annotated manually. In the same way, numerous software and search engines on the Internet now rely on constant updating. This creates confusion in that updating sometimes involves several individuals' subjectivity in structuring the knowledge. The need for standards has thus emerged. To create these standards, it is easier to define them while respecting as closely as possible the nature of the encoded object -here language- including its dynamics and relationship to human beings at the individual and collective level. Data mining at large and more particularly name-entity recognition and disambiguation processes could benefit from inputs from diachronic semantics.

Today, dynamic models of language are a necessity. Dynamics of language are at the heart of fundamental understandings about language, cognition, communication, sociology, history and information science. To be efficient, the tools we create have to include the dynamic aspect of language to adapt to the increasing speed of change. This piece of research is a small contribution to a complex issue. It shows that studies about the dynamics of meaning benefit from gathering insights from several disciplines.

References

Adelstein, A. 2007. “Unidad Léxica y Significado Especializado: Modelo De Representación a Partir Del Nombre Relacional Madre”. Doctoral Thesis, Universitat Pompeu Fabra.

Allan, K., and J.A. Robinson, ed. 2011. *Current Methods in Historical Semantics*. Mouton De Gruyter.

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).

Aronoff, M. 1976. *Word formation in generative grammar*. Linguistic inquiry monographs. Cambridge (Mass.): MIT Press.

Baayen, R.H. 1992. “Quantitative Aspects of Morphological Productivity.” *Yearbook of Morphology*: 109–150.

———. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baayen, R.H., and A. Renouf. 1996. “Chronicling the Times: Productive Lexical Innovations in an English Newspaper.” *Language*: 69–96.

Balbachan, F. 2006. “Killing Time: Metaphors and Their Implications in Lexicon and Grammar.” http://www.metaphorik.de/10/balbachan.pdf.

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.

Barthes, R. 1979. “Lecture in Inauguration of the Chair of Literary Semiology, Collège De France, January 7, 1977.” Translated by R. Howard. *October, the MIT Press* 8: 3–16.

Bastuji, J. 1974. “Aspects De La Néologie Sémantique.” Edited by J. Moeshler. *Langages* 8 (36): 6–19.

Bello, W. 2005. *Deglobalization: Ideas for a New World Economy*. Zed Books.

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

Berrendonner, A., M. Le Guern, and G. Puech. 1983. *Principes De Grammaire Polylectale*. Vol. 11. Presses universitaires de Lyon.

Binder, J. R., R. H. Desai, W. W. Graves, and L. L. Conant. 2009. “Where Is the Semantic System? A Critical Review and Meta-analysis of 120 Functional Neuroimaging Studies.” *Cerebral Cortex* 19 (12): 2767–2796.

Blackwell, S. 1993. “From Dirty Data to Clean Language.” *English Language Corpora: Design, Analysis and Exploitation*: 97–106.

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.

Blank, A., and P. Koch. 1999. “Introduction: Historical Semantics and Cognition.” In *Historical Semantics and Cognition*, edited by A. Blank and P. Koch, 1–16. Berlin/New York: 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.”Editions de Minuit.

Boussidan, A., S. Lupone, and S. Ploux. 2009. “La Malbouffe: Un Cas De Néologie Et De Glissement Sémantique Fulgurants.” In *“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, and 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.” Edited by J. F. Sablayrolles. *Cahiers De Lexicologie* 1 (100). Néologie Sémantique Et Analyse De Corpus: 117–136.

Boussidan, A., E. Sagi, and S. Ploux. 2009. “Phonaesthemic and Etymological Effects on the Distribution of Senses in Statistical Models of Semantics.” In *Proceedings of the Cognitive Science Conference Workshop on Distributional Semantics Beyond Concrete Concepts (DiSCo 2009)*, 35–40.

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

Buchi, E. 2000. “Le Point De Vue Onomasiologique En Étymologie. Réflexions Méthodologiques à Partir Du Roumain VREME Et TIMP.” *Revue De Linguistique Romane* 64 (255-56): 347–378.

Burgess, C., and K. Lund. 1997. “Modelling Parsing Constraints with High-dimensional Context Space.” *Language and Cognitive Processes* 12 (2): 177–210.

Cabré, M.T. 2006. “NEOROM, Réseau D’observatoires De La Néologie Des Langues Romanes.” *Neologica* 1: 115–118.

Cabré, M.T., M. Domènech, R. Estopà, J. Freixa, and E. Solé. 2003. “L’Observatoire De Néologie: Conception, Méthodologie, Résultats Et Nouveaux Travaux.” *L’innovation Lexicale*: 125–147.

Cabré, M.T., and L. de Yzaguirre. 1995. “Stratégie Pour La Détection Semiautomatique Des Néologismes De Presse.” *TTR: Traduction, Terminologie, Rédaction* 8 (2): 89–100.

Chesley, P. 2011a. “Linguistic, Cognitive, and Social Constraints on Lexical Entrenchment”. Doctoral Thesis, University of Minnesota.

———. 2011b. “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.

Collier, A. 1993. “Issues of Large-scale Collocational Analysis.” In *English Language Corpora: Design, Analysis and Exploitation, Papers from the 13th International Conference on English Language Research on Computerized Corpora*, 289–298.

Collier, A., M. Pacey, and A. Renouf. 1998. “Refining the Automatic Identification of Conceptual Relations in Large Scale Corpora.” In *Proceedings of the Sixth Workshop on Very Large Corpora, University of Montreal.*

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*. 2 vols. 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.
———. 1964. *Pour Une Sémantique Diachronique Structurale*. Centre de philologie et de littératures romanes de l'Université de Strasbourg.
Van de Cruys, T. 2010. "Mining for Meaning. The Extraction of Lexicosemantic Knowledge from Text". University of Groningen.
Crystal, D. 2005. "The Scope of Internet Linguistics." In *Proceedings of the American Association for the Advancement of Science Conference, Washington, DC, USA*, 17–21.
Darmesteter, A. 1887. *La Vie Des Mots*. Paris: Delagrave.
Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41 (6): 391–407.
Delacroix, H. 1924. *Le Langage Et La Pensée*. Librairie Félix Alcan. Paris.
Dozo, B.O. 2008. "Données biographiques et données relationnelles." *COnTEXTES* (n°3). La question biographique en littérature.
Dressler, W. U. 2005. "Word-formation in Natural Morphology." In *Handbook of Word-formation*, edited by Stekauer and Lieber, 267–284.
Drouin, P. 2003. "Term Extraction Using Non-technical Corpora as a Point of Leverage." *Terminology* 9 (1): 99–115.
Drouin, P., A. Paquin, and N. Ménard. 2006. "Extraction Semi-automatique Des Néologismes Dans La Terminologie Du Terrorisme." In *Proceedings of the 8th Conference "Journées Internationales d'Analyse Statistique Des Données Textuelles"(JADT 2006)*, 389–400.
Duchastel, J., F. Daoust, and D. Della Faille. 2004. "SATO-XML: Une Plateforme Internet Ouverte Pour L'analyse De Texte Assistée Par Ordinateur." *Le Poids Des Mots: Actes Des JADT 2004*: 353–363.
Durkheim, É. 1907. *Les Règles De La Méthode Sociologique*. Alcan.
Dury, P., and 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.
Eagleton, T. 1983. *Literary Theory: An Introduction*. Blackwell Publishers.
Eckardt, R, K. Von Heusinger, and C. Schwarze. 2003. "Historical Linguistics as a Transdisciplinary Field." In *Words in Time: Diachronic Semantics from Different Points of View*, edited by R. Eckardt, K. von Heusinger, and C. Schwarze, 1–33. 143. Mouton de Gruyter.
Fauconnier, G., and M. Turner. 2003. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books.
Fellbaum et.al, C. 1998. *WordNet: An Electronic Lexical Database*. MIT press Cambridge, MA.
Fernández-Domínguez, J. 2009. *Productivity in English Word-formation: An Approach to N+ n Compounding*. Vol. 341. Peter Lang Pub Inc.
Ferrand, M., D. Lequenne, V. Manneville, P. Jannot, and C. Lopez. 2009. "Apport De La Spatialisation Des Données En Analyse Multidimensionnelle Pour Évaluer L'impact

Des Activités Agricoles Sur La Teneur En Nitrates Des Eaux." *Revue MODULAD* 81 (39).
Firth, JR. 1957. "A Synopsis of Linguistic Theory." *Studies in Linguistic Analysis*.
Forston, B.W. 2005. "An Approach to Semantic Change." In *The Handbook of Historical Linguistics*, by Brian D Joseph and Richard D Janda, 648–666. Blackwell Handbooks in Linguistics. Malden (Mass.): Blackwell Publishers.
Frege, G. 1948. "Sense and Reference." *The Philosophical Review* 57 (3): 209–230.
Le Fur, D. 2008. *Dictionnaire des combinaisons de mots*. Les Usuels du Robert. Poche. Paris: Le Robert.
Furnas, GW, TK Landauer, LM Gomez, and ST Dumais. 1983. "Statistical Semantics: Analysis of the Potential Performance of Keyword Information Systems." In *Bell System Technical Journal, 62(6), 1753–1806.*
Garsou, M, ed. 1999. "Nouveaux Outils Pour La Néologie." *Terminologies Nouvelles* 20. Agence De La Francophonie Et Communauté Française De Belgique.
Gaume, B., K. Duvignau, L. Prévot, and Y. Desalle. 2008. "Toward a Cognitive Organization for Electronic Dictionaries, the Case for Semantic Proxemy." In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, 86–93.
Geeraerts, D. 1983. "Reclassifying Semantic Change." *Quaderni Di Semantica* 4: 217–240.
———. 1997. *Diachronic prototype semantics:a contribution to historical lexicology*. Oxford studies in lexicography and lexicology/ series ed. Richard Bailey, Noel Osselton and Gabriele Stein. Oxford: Clarendon Press.
———. 1999. "Diachronic Prototype Semantics.A Digest." In *Historical Semantics and Cognition*, by A. Blank and P. Koch, 91–107. Berlin/New York: Mouton de Gruyter.
Gérard, C. 2011. "Création Lexicale, Sens Et Textualité: Théories Et Analyses." *PhiN* 56: 1–30.
Gérard, C., and J. Kabatek. 2012. "La Néologie Sémantique En Questions." Edited by J.F. Sablayrolles. *Cahiers De Lexicologie* 1 (100). Néologie Sémantique Et Analyse De Corpus: 11–36.
Golub, G. H., F. T. Luk, and M. L. Overton. 1981. "A Block Lanczos Method for Computing the Singular Values and Corresponding Singular Vectors of a Matrix." *ACM Trans. Math. Softw.* 7 (2) (June): 149–169. http://doi.acm.org/10.1145/355945.355946.
"Google 'Semantic Search' Will 'Answer Questions' in a Shift Which Makes It More Like Bing." 2012. *Mail Online*, sec. Science. http://www.dailymail.co.uk/sciencetech/article-2115273/Google-Semantic-search-answer-questions-shift-makes-like-Bing.html.
Greimas, A.J. 1965. *Sémantique Structurale*. Larousse. Paris.
Grice, H. P. 1975. "Logic and Conversation." *1975*: 41–58.
Grzega, J., and M. Schöner. 2007. "English and General Historical Lexicology". Onomasiology Online Monographs Vol.1. Materials for Onomasiology Seminar. Katholische Universität Eichstätt-Ingolstadt, Germany. http://www.scribd.com/doc/57546044/12/Old-Words-in-New-Use-Semantic-Change.
Guilbert, L. 1975. *La créativité lexicale*. Langue et langage. Paris: Larousse.
Guillaume, A. 2010. "Diachronie Et Synchronie: Passerelles (étymo)logiques. La Dynamique Des Savoirs Millénaires." Edited by Duteil-Mougel Carine. *Texto! Textes Et Cultures* XV (2). http://www.revue-texto.net/index.php?id=2557.
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.

Habert, B., A. Nazarenko, and A. Salem. 1997. *Les linguistiques de corpus*. U. Linguistique. Paris: A. Colin.
Halle, M., and A. Marantz. 1993. “Distributed Morphology and the Pieces of Inflection.” In *Current Studies in Linguistics*, edited by K. Hale and J. Keyser. Vol. 24. 20. Cambridge (Mass.).
Harris, Z.S. 1954. “Distributional Structure.” *Word*.
Hessel, S., and E. Morin. 2011. *Le Chemin De L’espérance*. Fayard.
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*.
Hilpert, M., and S.T. Gries. 2009. “Assessing Frequency Changes in Multistage Diachronic Corpora: Applications for Historical Corpus Linguistics and the Study of Language Acquisition.” *Literary and Linguistic Computing* 24 (4): 385–401.
Hjelmslev, L. 1959. “Pour Une Sémantique Structurale.” *Essais Linguistiques*: 96–112.
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.
Huguet, E. 1934. *L’évolution Du Sens Des Mots Depuis Le XVIe Siècle*. Etudes De Philologie Française. Paris: Librairie E.Droz.
Jackendoff, R.S. 1985. *Semantics and Cognition*. The MIT Press.
Janicijevic, T., and D. Walker. 1997. “NeoloSearch: Automatic Detection of Neologisms in French Internet Documents.” In *Proceedings of ACH-ALLC’97*, 93–94.
Jespersen, O. 1922. *Language, Its Nature, Development, and Origin*. H. Holt.
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.
Ji, H., B. Lemaire, H. Choo, and S. Ploux. 2008. “Testing the Cognitive Relevance of a Geometric Model on a Word Association Task: A Comparison of Humans, ACOM, and LSA.” *Behavior Research Methods* 40 (4): 926.
Ji, H., and S. Ploux. 2003. “Automatic Contexonym Organizing Model.” In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, 622–627.
Ji, H., S. Ploux, and E. Wehrli. 2003. “Lexical Knowledge Representation with Contexonyms.” In *Proceedings, Machine Translation Summit IX*.
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.
Kilgarriff, A., S. Reddy, J. Pomikálek, and A. Pvs. 2010. “A Corpus Factory for Many Languages.” In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. Malta.
Klemperer, V. 1975. *LTI, la langue du IIIe Reich: carnets d’un philologue*. Translated by E. Guillot. 1 vols. Agora. Paris: Pocket.
Kloumann, I.M., C.M. Danforth, K.D. Harris, C.A. Bliss, and P. S. Dodds. 2012. “Positivity of the English Language.” *PLoS ONE* 7 (1).
Kosko, B. 1993. *Fuzzy Thinking: The New Science of Fuzzy Logic*. New York: Hyperion.

L'Homme, M.C., C. Bodson, and R. S. Valente. 1999. "Recherche Terminographique Semi-automatisée En Veille Terminologique: Expérimentation Dans Le Domaine Médical." *Terminologies Nouvelles* (20): 25–36.
Lafon, P. 1980. "Sur La Variabilité De La Fréquence Des Formes Dans Un Corpus." *Mots* 1 (1): 127–165.
Lakoff, G. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago University Press.
———. 1993. "The Contemporary Theory of Metaphor." *Metaphor and Thought* 2: 202–251.
Lakoff, G., and M. Johnson. 1980. *Metaphors We Live By*. Vol. 111. Chicago London.
Landauer, T.K., and S.T. Dumais. 1997. "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge." *Psychological Review* 104 (2): 211.
Landauer, T.K., D.S. McNamara, S.E. Dennis, and W.E. Kintsch. 2007. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates Publishers.
Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford University Press.
Lebart, L, and A Salem. 1994. *Statistique Textuelle*. Dunod.
Lebart, L., M. Piron, and J.F. Steiner. 2003. *La Sémiométrie*. Dunod. Paris.
Leblanc, J.M., and P. Fiala. 2004. "Autour Du Je Présidentiel." In *Le Poids Des Mots: Actes Des 7es Journées Internationales D'analyse Statistique Des Données Textuelles*. Vol. 2. Louvain-la-Neuve.
Lemaire, B. 2008. "Limites De La Lemmatisation Pour L'extraction De Significations." *9e Journées Internationales d'Analyse Statistique Des Données Textuelles* Lyon, France.
Lenci, A., ed. 2008. In *ESSLLI 2008 Workshop on Distributional Lexical Semantics*.
Lewes, G. H. 1878. *The Physical Basis of Mind*. Kessinger Publishing.
Lin, D., and P. Pantel. 2001. "DIRT-discovery of Inference Rules from Text." *Knowledge Discovery and Data Mining*.
Littré, E. 1888. *Comment Les Mots Changent De Sens*. C. Delagrave.
Locke, J. 1959. *An Essay Concerning Human Understanding: Collated and Annotated, with Prolegomena, Biographical, Critical, and Historical by Alexander Campbell Fraser*. Dover Publications.
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.
Marchand, H. 1966. *The Categories and Types of Present-day English Word-formation: A Synchronic-diachronic Approach*. 13. University of Alabama Press.
Martinet, A. 1964. *Économie Des Changements Phonétiques: Traité De Phonologie Diachronique*. Vol. 10. A. Francke.
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.
McCallum, A.K. 2002. "Mallet: A Machine Learning for Language Toolkit." *Http://mallet.cs.umass.edu*.
McLuhan, M. 1994. *Understanding Media: The Extensions of Man*. MIT press.
Meillet, A. 1906. "Comment Les Mots Changent De Sens." *Linguistique Historique Et Linguistique Générale*: 230–271.

Mejri, S. 2006. “La Reconnaissance Automatique Des Néologismes De Sens.” In *Mots, Termes Et Contextes, Actes Des Septièmes Journées Scientifiques Du Réseau LTT*, edited by D. Blampain, P. Thoiron, and M. Van Campenhoudt,, 545–557. Paris: Editions des Archives Contemporaines et Agence universitaire de la francophonie.

Milner, J.C. 1978. *L’amour De La Langue*. Vol. 2. Éditions du Seuil.

Milroy, J. 1992. *Linguistic Variation and Change*. Oxford: Blackwell.

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

Nazar, R. 2011a. “Neología Semántica: Un Enfoque Desde La Lingüística Cuantitativa.” *Seminario IULAterm*. http://www.iula.upf.edu/materials/111214nazar.pdf.

———. 2011b. “A quantitative approach to concept analysis”. Universitat Pompeu Fabra. http://tdx.cat/handle/10803/7516.

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

———. 1999a. “Synedcoque as a cognitive and communicative strategy.” In *Historical semantics and cognition*, edited by A. Blank and Peter Koch, 197–214. Cognitive linguistics research. Berlin: Mouton de Gruyter.

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

Newman, J. 2012. “Google Talks Up Big Search Changes | Techland | TIME.com.” *Time*, March 15. http://techland.time.com/2012/03/15/google-talks-up-big-search-changes/.

Nichols, J. 2005. “Diversiy and Satbility in Language.” In *The handbook of historical linguistics*, by Brian D Joseph and Richard D Janda, 283–310. Blackwell handbooks in linguistics. Malden (Mass.): Blackwell Publishers.

Oakley, T. 1998. “Conceptual Blending, Narrative Discourse, and Rhetoric.”

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

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.

Patel, M., J.A. Bullinaria, and J.P. Levy. 1998. “Extracting Semantic Representations from Large Text Corpora.” In *Proceedings of the 4th Neural Computation and Psychology Workshop*, 199–212.

Paul, H. 1880. “Prinzipien Der Sprachgeschichte, 1880.” *Manfred Lurkner, W\örterbuch Der Symbolik, S* 478.

Peirsman, Y., D. Geeraerts, and D. Speelman. 2010. “The Automatic Identification of Lexical Variation Between Language Varieties.” *Natural Language Engineering* 16 (4): 469–490.

Peirsman, Y., K. Heylen, and D. Geeraerts. 2008. “Size Matters: Tight and Loose Context De Finitions in English Word Space Models.” In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, 34–41.

Peirsman, Y., and D. Speelman. 2009. “Word Space Models of Lexical Variation.” In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 9–16.

Picton, A. 2009. “Diachronie En Langue De Spécialité. Définition D’une Méthode Linguistique Outillée Pour Repérer L’évolution Des Connaissances En Corpus. Un Exemple Appliqué Au Domaine Spatial.” Université Toulouse Le Mirail. http://tel.archives-ouvertes.fr/tel-00429061/fr/.

Plag, I. 1999. *Morphological productivity: structural constraints in English derivation*. Topics in English linguistics. Berlin; New York: Mouton de Gruyter.

Ploux, S. 1997. “Modélisation Et Traitement Informatique De La Synonymie.” *Linguisticae Investigationes, Revue Internationale De Linguistique Française Et De Linguistique Générale* 21 (1): 1–28.

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), Valletta, Malta*.

Ploux, S., and B. Victorri. 1998. “Construction D’espaces Sémantiques à L’aide De Dictionnaires De Synonymes.” *Traitement Automatique Des Langues* 39 (1): 161–182.

Posner, R. 1997. *The Romance Languages*. Cambridge: Cambridge University Press.

Pustejovsky, J. 1995. *The generative lexicon*. Cambridge (Mass.); London: MIT Press.

Rastier, F. 1987. *Sémantique Interprétative*. Presses universitaires de France.

———. 1999. “Cognitive Semantics and Diachronic Semantics.” In *Historical Semantics and Cognition*, by Andreas Blank and Peter Koch, 109–144. Berlin/New York: Mouton de Gruyter.

———. 2000. “De la sémantique cognitive à la sémantique diachronique: les valeurs et l’évolution des classes lexicales.” In *Théories contemporaines du changement sémantique*, edited by François Jacques, 9:135–164. Mémoires de la Société de Linguistique de Paris. Leuven: Peeters.

———. 2001. *Sémantique Et Recherches Cognitives*. Presses universitaires de France.

Rastier, F., and M. Valette. 2009. “De La Polysémie à La Néosémie.” *Langue Française*.

Reisig, K. 1839. “Semasiologie Oder Bedeutungslehre.” In *Professor Karl Reisigs Vorlesungen Über Lateinische Sprachwissenschaft*, by Friedrich Haase. Leipzig: Lenhold.

Rémi-Giraud, S. 2000. “Schémas notionnels et significations lexicales: l’exemple du mot AIR (apparence, manière d’être) aux XVIIe et XXe siècles.” In *Théories contemporaines du changement sémantique*, edited by Société de linguistique de Paris, 31–58. Leuven; Paris: Peeters.

Rendgen, S. 2012. *Information Graphics*. Edited by J Wiedemann. Taschen.

Renon, A.L. to be published. “‘Graphic Design’ and ‘Objectivity’, a Study About Meta-atlases.” In *Knowledge in Architecture and Graphic Design: What Relation?* Ecole Supérieure d’Art et Design, Grenoble - Valence: B42.

———. 2010. “Design as a Verb”. Mémoire de recherche, Valence: Ecole supérieur d’Art et Design, Genoble.

Renouf, A. 1993a. “A Word in Time: First Findings from the Investigation of Dynamic Text’.” *English Language Corpora: Design, Analysis and Exploitation, Rodopi, Amsterdam*: 279–288.

———. 1993b. “Sticking to the Text: a Corpus Linguist’s View of Language.” In *Aslib Proceedings*, 45:131–131.

———. 1993c. “What the Linguist Has to Say to the Information Scientist.” *Journal of Document and Text Management* 1 (2): 173–190.

———. 1993d. “Making Sense of Text: Automated Approaches to Meaning Extraction.” In *International Online Information Meeting*, 77–86.

———. 1996. “The ACRONYM Project: Discovering the Textual Thesaurus.” *Language and Computers* 16: 171–188.

———. 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, C. 2012. “Vers Un Traitement Automatique De La Néosémie: Approche Textuelle Et Statistique.” Thèse de doctorat, Université de Lorraine.

Rissanen, M., M. Rissanen, M. Kytö, and S. Wright. 1994. “The Helsinki Corpus of English Texts.” In *Corpora Across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora*, 73–79.

Robert, S. 2008. “Words and Their Meanings.” In *From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations*, by M. Vanhove, John Benjamins Pub Co, 106:55–92.

Roche, S, and L Bowker. 1999. “Cenit: Système De Détection Semi-automatique Des Néologismes.” Edited by M Garsou. *Terminologies Nouvelles* 20. Agence De La Francophonie Et Communauté Française De Belgique: 12–15.

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.

Rosch, E. 1973. “Natural Categories.” *Cognitive Psychology* 4 (3): 328–350.

———. 1978. “Principles of Categorization.” *Fuzzy Grammar: a Reader*: 91–108.

Rosell, M. 2009. *Text Clustering Exploration: Swedish Text Representation and Clustering Results Unraveled*. Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan.

Rosell, M., M. Hassel, and V. Kann. 2009. “Global Evaluation of Random Indexing Through Swedish Word Clustering Compared to the People’s Dictionary of Synonyms.”

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

Sablayrolles, J. F., ed. 2012. “Cahiers De Lexicologie N° 100: Néologie Sémantique Et Analyse De Corpus”. Laboratoire Lexiques, Dictionnaires, Informatique (LDI, université Paris 13 – université de Cergy-Pontoise – CNRS).

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

———. 2000. *La néologie en français contemporain: examen du concept et analyse de productions néologiques récentes*. 1 vols. Lexica. Paris: H. Champion.

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.

———. 2011. “Tracing Semantic Change with Latent Semantic Analysis.” In *Current Methods in Historical Semantics*, edited by Kathryn Allan and Justyna A. Robinson, 73:161–183.

Sahlgren, M. 2005. “An Introduction to Random Indexing.” In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*. Vol. 5.

———. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-dimensional Vector Spaces*. Department of Linguistics, Stockholm University.

Salton, G., A. Wong, and C.S. Yang. 1975. “A Vector Space Model for Automatic Indexing.” *Communications of the ACM* 18 (11): 613–620.

De Saussure, F. 1916. *Cours De Linguistique Générale*. Vol. 1. Otto Harrassowitz Verlag.

Schütze, H. 1993. “Word Space.” In *Advances in Neural Information Processing Systems 5*.

———. 1996. *Ambiguity in Language Learning: Computational and Cognitive Models*. Cambridge University.

Selkirk, E. 1982. *The Syntax of Words*. Cambridge: MIT Press.
Shapiro, M. 1991. *The Sense of Change: Language as History*. Bloomington: Indiana University Press.
Smith, Adam. 1982. *The theory of moral sentiments*. Edited by David Daiches Raphael and Alexander Lyon Macfie. 1 vols. Indianapolis, Ind.: Liberty Fund.
Sperber, Dan, and Deirdre Wilson. 1995. *Relevance: communication and cognition*. Oxford; Malden (Mass.): Blackwell.
Sperber, Hans. 1923. *Einführung in die Bedeutungs Lehre*. Bonn; Leipzig: Kurt Schroeder.
Steinmetz, S. 2008. *Semantic Antics: How and Why Words Change Meaning*. Random House Reference.
Stern, G. 1931. *Meaning and Change of Meaning; with Special Reference to the English Language.* Bloomington: Indiana University Press.
Sweetser, E. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: Cambridge University Press.
Taylor, B. 1999. "Cognitive semantics and structural semantics." In *Historical semantics and cognition*, edited by A. Blank and Peter Koch, 17–48. Cognitive linguistics research. Berlin: Mouton de Gruyter.
Thagard, P. 2009. "Why Cognitive Science Needs Philosophy and Vice Versa." Edited by W.D. Gray. *Topics in Cognitive Science. Journal of the Cognitive Science Society.* 1 (2). Philosophy in and Philosophy of Cognitive Science (April): 237–254.
Tilby, M. 2009. "Neologism: A Linguistic and Literary Obsession in Early Nineteenth-Century France." *The Modern Language Review* 104 (3): 676–695.
Tournier, J. 1985. *Introduction Descriptive à La Lexicogénétique De L'anglais Contemporain*. Champion Books.
———. 1991. *Précis De Lexicologie Anglaise*. Nathan.
Traugott, E. C. 1989. "On the Rise of Epistemic Meanings in English: An Example of Subjectification in Semantic Change." *Language*: 31–55.
Traugott, Elizabeth Closs, and Richard B Dasher. 2002. *Regularity in semantic change*. Cambridge studies in linguistics. Cambridge: Cambridge University Press.
Turner, M. 1996. *The Literary Mind: The Origins of Language and Thought*. New York: Oxford University Press.
Turney, and M.L. Littman. 2003. "Measuring Praise and Criticism: Inference of Semantic Orientation from Association." In *ACM Transactions on Information Systems (TOIS)*.
Turney, P.D., M.L. Littman, J. Bigham, and V. Shnayder. 2003. "Combining Independent Modules in Lexical Multiple-choice Problems." *Recent Advances in Natural Language Processing III: Selected Papers from RANLP* 2003: 101–110.
Turney, P.D., and P. Pantel. 2010. "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* 37 (1): 141–188.
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.
———. 1972. *Semantics*. Edited by Thomas Albert Sebeok. Current trends in linguistics. The Hague: Mouton.
Ussishkin, A. 2005. "A Fixed Prosodic Theory of Nonconcatenative Templaticmorphology." *Natural Language & Linguistic Theory* 23 (1): 169–218.
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)*, 257–262.
Weaver, W. 1955. "Translation." *Machine Translation of Languages* 14: 15–23.

Webster, F. 2006. *Theories of the Information Society*. Taylor & Francis.
Wilkins, D. 1996. “Natural Tendencies of Semantic Change and the Search for Cognates.” In *The Comparative Method Reviewed,*, edited by Durie, Mark and Malcolm Ross, 236–304. Oxford University Press.
Wilson, M. 1988. “MRC Psycholinguistic Database: Machine-usable Dictionary, Version 2.00.” *Behavior Research Methods* 20 (1): 6–10.
Wittgenstein, L. 1953. *Philosophical Investigations*. Blackwell Publishing.
Xia, Meng. 2011. “Construction D’un Modèle Dynamique Diachronique Pour l’Atlas Sémantique”. Rapport de stage de première année pour le diplôme d’ingénieur de l’école centrale de Lyon. Réalisé au L2C2 sous la direction de Sabine PLoux.
Zeldin, T. 1994. *An Intimate History of Humanity*. Vintage.

APPENDICES

APPENDIX 1 : INNOVATIVE WORDS IN mAL-

ADJ.

mal-accueillants
mal-administrés
malaimable
mal-aimables
mal-aimant
mal-aimé
malaimé
malaimée
malaisant
malaisantes
mal-aisé
malaisé
malapposition
malaventures
malavisé
mal-baisée
malbouffe
malcamera
malcommode
mal-comprenants
malconnue
malconnues
malcoordonnée
mal-défendus
mal-disants
maléficiantes
maléficieuse
mal-endettés
malformatif
malformatifs
malformative
malformatives
mal-foutus
mal-gouvernés
malgracieux
mal-né
mal-nés
mal-nommée
mal-nommés
mal-nourri
mal-nutris
mal-orthographiants
mal-parlants
malparti
malpayés
malpensant
mal-pensant
malpensante
malpensants
malpratique
malprotégée
maltraitables
maltraitante
maltraitantes
maltraitants
mal-vêtus

N.

malacquis
mal-administration
maladministration
mal-aimants
mal-aimé
malaimée
malaisant
mal-amour
mal-appris
mal-assis
malassis
malaventure
mal-baisées
mal-bête
mal-boire
malbonheur
malbonheur
mal-bouffe
malbouffe
malboulot
mal-Bové
mal-chaussés
mal-classé
mal-classés
malcoiffée
mal-comprenants
mal-compris
mal-comptants
mal-construction
mal-croyants
maldéveloppement
mal-développement
mal-développements
mal-développernent
maldiction
mal-dire
mal-eBay
mal-élus
mal-emploi
mal-emplois
malendettement
malfait
malfonctionnement
mal-fondé
malformatif
malformatifs
malformation
malformative
malformatives
malformé
malformées
malformés
mal-généré
malgérance
malgouvernance
mal-gouvernance
mal-gouvernement
malinfo
mal-informés
mal-inscrits
malintention
mal-intention
malinvestissement
mal-lettrés
mal-logé
mallogement
mal-logement
mal-logements
mal-loti
mal-lotis
mal-lunées
mal-marié
malmenage
mal-mesure
mal-mesure
mal-modernes
mal-monde
mal-mort
malmort
mal-morts
mal-mutants
malnourris
malnuit
malnutri
malnutrie
malnutries
malnutris
malotis
mal-partis
malpensance
mal-pensance
mal-pensant
malpensants
mal-pense
mal-penser
mal-perceuses
mal-perçu
malpère
malposition
mal-protégés
mal-régulation
malsapé
malséance
mal-sentants
maltaiteur
maltélé
maltournée
maltraitement
malvie
mal-vie
mal-vivre
malvoyance
malvoyant
malvoyante
malvoyantes
malvoyants

APPENDIX 2: Excerpt of an article taken from the corpus *Le Monde,* January 1999 :

« Pour la seule année 1998, la « veille néologique » de Larousse, assurée par la lexicographe Hélène Houssemaine- Florent, a noté dans notre journal 2 194 néologismes ! Entendez par là des expressions qui ne se trouvent pas dans les dictionnaires. Ce chiffre peut paraître énorme, mais la base de données est formelle, fournissant pour chaque mot la date de son emploi et la phrase dans laquelle il figurait. Un classement en dix-neuf rubriques indique, par exemple, que la vie quotidienne, les loisirs et le sport ont donné lieu, en 1998, à 230 nouveautés. Les sciences et techniques, prises globalement, ainsi que la culture, sont grandes productrices de néologismes. C'est moins vrai pour l'économie (66 recensions), et beaucoup moins pour les religions (11), la police et l'armée (5). Parmi les mots nouveaux apparus en 1998 dans Le Monde, on relève des adverbes, plus ou moins heureux : tartuffement, capitalistiquement, improbablement, illimitablement... La féminisation des noms a conduit à écrire littératrice, docteure, rapporteure, amatrice ou metteuse en scène. Dans l'euphorie du Mondial, le football a inspiré les plumes : footeux, footophile, footocratie et même footballistoïde...

Des mots nouveaux sont composés avec des préfixes à la mode, comme auto (autodénigrement, autogénocide, autoputsch, autocongratulation, autofiction, ou encore - de manière plus obscure - autopathographie). Eco - comme d'ailleurs euro -, se met à toutes les sauces : écotourisme, écoguerrier, écotaxe, écoconseiller... L'évolution des techniques fait écrire biofibre, bionique, biométrique, mais aussi biojeu, bioterrorisme, biovigilance, bioprospection. Très prisé également dans Le Monde, le préfixe dé, qui semble illustrer un délitement général (déliaison, déprotection, décivilisation, déspectacularisation, désintermédiation). Quant à cyber, il n'a sans doute pas dit son dernier mot , après cybercitoyen, cybercriminalitéet cybernétisation. »

My translation:

"Just for the year 1998, Larousse's "neology watch", conducted by the lexicographer Hélène Houssemaine- Florent, noted 2 194 neologisms in our newspaper. Understand by this, expressions that cannot be found in dictionaries. This figure may seem huge, but the database is formal and provides for each word its date of use and the sentence in which it appeared. A classification in 19 sections shows for example that everyday life, hobbies and sports gave birth to 230 new items in 1998. Sciences and techniques, taken as a whole, as well as culture, are great neologisms producers. It is less so for economy (66 listed) and even less so for religions (11), the police and the army (5). Among new words in 1998 in Le Monde, there are more or less fortunate adverbs: *tartuffement, capitalistiquement, improbablement, illimitablement...* The feminization of words led to write *littératrice, docteure, rapporteure, amatrice* and *metteuse en scène.* With the euphoria of the world cup, football inspired writers: *footeux, footophile, footocratie* and even *footballistoïde...*

New words are composed with fashionable prefixes, such as auto (*autodénigrement, autogénocide, autoputsch, autocongratulation, autofiction,* or even –in a more obscure way- *autopathographie*). Eco –in the same way as euro- is adapted to any purpose: *écotourisme, écoguerrier, écotaxe, écoconseiller...* The evolution of techniques makes people write *biofibre, bionique, biométrique*, as well as *biojeu, bioterrorisme, biovigilance, bioprospection.* The prefix dé is also very popular in Le Monde, and seems to illustrate a general splintering *(déliaison, déprotection, décivilisation, déspectacularisation, désintermédiation).* As for cyber, it has not yet said its last word, after *cybercitoyen, cybercriminalitéet cybernétisation."*

APPENDIX 3 Words un cyber in middle and high frequency ranges in the corpus *Le Monde*

total frequency range <3 and < 50		
unattested		attested
cyberachat	cyberdissidents	cybernétiques
cyberadresses	cyberéconomie	cybernaute
cyber-amis	cybercinéma	cyberguerre
cyberattaque	cyber-comm	cyber-café
cyberbanque	cybercentres	cyber-espace
cyberboutique	cybercentre	cyberespaces
cyberbus	cybercommunes	cyber-criminalité
cybercasinos	cybercash	
cybercité	cyberconflits	
cybercitoyen	cyberlicence	
cyberclient	cyberthéâtre	
cybercriminels	cyberconsommation	
cyberdissident	cyberdépendance	
cyber-dissident	cyberbase	
cyberdistributeurs	cyberspace	
cyberentreprises	cybermétrie	
cyberentretiens	cyberdeck	
cyberflash	cyberpapy	
cyberflics	cybermarchés	
cyberhosto	cybercâble	
cyberjournalisme	cybermarchand	
cyberjournaliste	cyberpolice	
cyberlecteurs	cyberposte	
cyberlibrairie	cybermarché	
cybermaison	cybertribunal	
cyber-militant	cyberspatial	
cybermoine	cyberport	
cybernation	cyberacheteur	
cyberpapys	cyberpublicité	
cyberpirates	cyberconsommateur	
cyberpromotions	cybercommerce	
cyberpub	cybersurveillance	
cyberpunks	cyberdémocratie	
cybersalon	cyberacheteurs	
cybersexe	cybercampagne	
cyber-sexe	cyberjeunes	

cybersquat	cyberterrorisme
cybersquatters	cyberjournalistes
cyberterroriste	cyberattaques
cybercrime	cyberouest
	cyberparties

total frequency range > 50	
unattested	attested
cybermarchands	cyberespace
cyberpunk	cyber
	cybercafé
	cybernétique
	cybercriminalité
	cyberculture
	cybermonde

APPENDIX 4: All innovative words in *bio-* in Le Monde. Recently attested words are in red.

Bio- as "life"/Life sciences/biology			
NOM_bio-accumulable	ADJ_biovalidité	NOM_biofibre	NOM_bio-minéralisation
ADJ_bio-accumulable	NOM_bio-vigilance	NOM_biofilm	NOM_bio-monitor
NOM_bioaccumulables	NOM_bio-astronomie	NOM_biofiltration	NOM_biomoraliste
NOM_bioaccumulation	NOM_bioastronomie	NOM_biofiltre	NOM_bionanotechnologie
ADJ_bioaccumulatif	NOM_biophotonique	NOM_bio-fonctionnelle	NOM_bionicus
ADJ_bioactif	ADJ_biophotonique	NOM_bio-game	NOM_bionien
NOM_bioactif	NOM_bioscience (1982)	NOM_biogénéticien	NOM_bio-organique
ADJ_biocentrique	NOM_bioterrorisme	NOM_bio-génétique	NOM_bioparc
ADJ_bioclinique	NOM_bio-terrorisme	NOM_biogéochimie	NOM_biopatrimoine
NOM_bioclinique	NOM_bio-acousticien	NOM_bio-géochimie	NOM_biopharmaceutique
ADJ_biodégradé	NOM_bioagresseurs	NOM_biogéochimiste	NOM_bio-pharmaceutique
NOM_bio-diesel (1992)	NOM_bio-amélioré	NOM_biogéographe	NOM_biopharmacie
ADJ_biodisponible	NOM_bio-art	NOM_bio-géographie (attested without a hyphen)	NOM_bio-pharmacie
ADJ_bioécologique	NOM_bioArt	NOM_biogéographie-écologie	NOM_biopharmacologie
ADJ_bio-écologique	NOM_bioartiste	NOM_biogéoscience	NOM_biophobie
ADJ_bioéconomique	NOM_biobanque	NOM_biogestionnaire	NOM_biophore
ADJ_bioéthique	NOM_biobricolage	NOM_bio-hacker	NOM_bio-physico-chimie
ADJ_bio-éthique	NOM_Bio-Bug	NOM_biohistoire	NOM_biophysiologiste
NOM_bioéthique	NOM_biocapitale	NOM_biohistorien	NOM_biopigment
NOM_bio-éthique	NOM_biocapteur	NOM_bio-imagerie	NOM_biopiratage
NOM_bioéthique-biovigilance	NOM_bio-capteur	NOM_bio-immunothérapie	NOM_biopirate
ADJ_biofiltrant	NOM_bioinformaticien	NOM_bio-indicateur	NOM_biopiraterie
ADJ_biogéochimique	NOM_bio-informaticien	NOM_bio-industrie	NOM_bio-piraterie
ADJ_biogéologique	NOM_biochrome	ADJ_bio-industriel	NOM_biopolitologue
ADJ_biohydrogène	NOM_biochronologie	NOM_bio-industriel	NOM_biopompe
ADJ_bioinformatique	NOM_bioclimaticien	NOM_bioremédiation	NOM_biopouvoir
ADJ_bio-informatique 1995	NOM_bio-colonialisme	NOM_bio-informaticien	NOM_bio-pouvoir
NOM_bioinformatique	NOM_biocombinat	NOM_bio-ingénierie	NOM_bioprospecteur
NOM_bio-informatique	NOM_bio-communicant	NOM_bio-inorganique	NOM_bioprospection
ADJ_biomarin	NOM_biocompatibilité (1980)	NOM_bio-inspiration	NOM_bioprotecteur
NOM_biomécanicien	NOM_bioconcentration	NOM_bio-inspiré	NOM_bioréacteur (1982)
ADJ_biomécanicien	NOM_biocontamination	NOM_bio-invasion	NOM_biorésorbable
ADJ_biomédico-généalogico-génétique	NOM_biocontrôle	NOM_biojeu	NOM_biorobotique
ADJ_biométéorologique	NOM_biocontrôleur	NOM_bio-jeu	NOM_bioroïde

ADJ_biomilitaire	NOM_bioconversion	NOM_bio-junior	NOM_biorythme (1972)
ADJ_biomimétique	NOM_biocratie	NOM_bio-légitimité	NOM_biosenseur
ADJ_biomorphe	NOM_bio-criminologue	NOM_biolistique	NOM_biosilicone
ADJ_biopathologique	NOM_bio-cristallographie	NOM_biologisation	NOM_biosonar
ADJ_biopharmaceutique	NOM_biodéfense	NOM_biologues	NOM_biospécificité
ADJ_biophilosophique	NOM_bio-délire	NOM_biomachine	NOM_biosynthèse
ADJ_bioplastique	NOM_biodesign (1987)	NOM_bio-marin	NOM_biosystématique
ADJ_biopolitique	NOM_bio-design	NOM_biomarqueur	NOM_biosystème
NOM_biopolitique	NOM_biodétecteur	NOM_biostatisticien	NOM_biotechnicien
NOM_bio-politique	NOM_biodisponibilité	NOM_biostatisticiens	NOM_bioterroriste (1998)
NOM_biopuce (1997)	NOM_bioéconomie	ADJ_biomarqueur	NOM_bioterrorisme (1998
NOM_biocatalyse (1979)	NOM_bioénergie (1975)	NOM_biomatériau (1982)	NOM_bio-terroriste
NOM_biocéramique	NOM_bio-énergie	NOM_biomatériel	NOM_bio-test
ADJ_biosensoriel	NOM_bioénergiste	NOM_biomatricien	NOM_biothèque
ADJ_biosimilaire	NOM_bioenvironnement	NOM_biomédicament	NOM_bio-transformation
NOM_biosimilaire	NOM_bioéquivalence	NOM_bio-métal	NOM_biotransformés
ADJ_biosocial	NOM_bioéquivalent	NOM_biométéorologie	NOM_biotypage
ADJ_biosomatique	NOM_bio-esthéticien	NOM_biométrisation	NOM_bio-zoologique
ADJ_biosonique	NOM_bioéthicien	NOM_biomimesis	
ADJ_biostatistique	NOM_bio-ethnique	NOM_biomimétique	
NOM_biostatistique	NOM_bio-évolutionniste	NOM_biomimétisme	
ADJ_biotoxicologiques	NOM_biofeedback	NOM_biominéralisation	

Bio- as "organic"	
ADJ_biocosmétique	NOM_biodéchets
NOM_biocosmétique	NOM_bio-dermique
ADJ_bioéquitable	NOM_biodynamie (1976)
NOM_bio-équitable	NOM_biodynamique
ADJ_bio-équitable	NOM_bio-génération
ADJ_biotoniques	NOM_bio-ménagère
NOM_bioanalyse	NOM_bio-paranoïa
NOM_bio-attitude	ADJ_biorelaxant
NOM_bio-business	NOM_bioséchage
NOM_bio-conquistadors	NOM_bioservice
NOM_bio-cons	NOM_bioproduits
NOM_bioculture	ADJ_biomagique

Bio- as “ecological”: made from organic substance/biodegradable	
ADJ_biocombustible	NOM_bio-polyester
NOM_biocombustible	NOM_biopolymère
NOM_biodiesel (1992 under the spelling bio-diesel) ADJ_biodiesel	NOM_bio-polymère
	NOM_bioprocédé
NOM_bio-carburant (1977)	NOM_bioraffinerie
NOM_bio-combustible	NOM_bio-raffinerie
NOM_bioessence	NOM_bioressource
NOM_bio-essence	NOM_biopesticide (1988 Canada.)
NOM_bioéthanol (1987)	ADJ_biopesticide
NOM_bio-éthanol	NOM_bio-plastique
NOM_biofioul	NOM_bioclimatisation
NOM_biofuel	NOM_bio-écologie
NOM_biogaz	NOM_bio-réfugiés
NOM_bio-gaz	NOM_bio-séquestration
NOM_biokérosène	NOM_biotraitement
NOM_biolubrifiant	NOM_bioconstruction
NOM_biométhane	

Bio- > biotechnologie
NOM_biotech
NOM_bio-tech
NOM_bio-techno
NOM_biotechno
NOM_biotechnocratie
NOM_biotechnologie (1980)
NOM_bio-technologie
NOM_biotechnologiste
NOM_biotechno-militaire
NOM_biotech-santé
NOM_bioindustrie (av. 1979)
NOM_bionumérique
NOM_biopunks
NOM_bio-punk

Idiosyncrasies
NOM_biotennistique
NOM_bio-lieu
NOM_biomerde
NOM_bio-tennis
NOM_bio-sun-sport

> biosphère	> biographie
ADJ_biosphériques	ADJ_biocinématographique
NOM_biosphériens	NOM_bio-auto-graphie
	NOM_bio-filmographie