Méthodes de veille textométrique multilingue appliquées à des corpus de l'environnement et de l'énergie : " Restitution, prévision et anticipation d'événements par poly-résonances croisées "

Lionel Shen

UNIVERSITE SORBONNE NOUVELLE - PARIS 3

Ecole Doctorale 268 - LANGAGE ET LANGUES : DESCRIPTION, THEORISATION, TRANSMISSION

THÈSE DE DOCTORAT : Sciences du langage

Lionel SHEN

METHODES DE VEILLE TEXTOMETRIQUE MULTILINGUE APPLIQUEES A DES CORPUS DE L'ENVIRONNEMENT ET DE L'ENERGIE

« Restitution, prévision et anticipation d'événements par poly-résonances croisées »

Thèse dirigée par André SALEM

Soutenue le vendredi 21 octobre 2016

Jury

Annie BERTIN	Professeur, Université Paris 10 (Rapporteur)
André SALEM	Professeur émérite, Université Paris 3 (Directeur)
Monique SLODZIAN	Professeur émérite, INALCO (Présidente)
Pierre ZWEIGENBAUM	Directeur de recherche, LIMSI-CNRS (Rapporteur)

METHODES DE VEILLE TEXTOMETRIQUE MULTILINGUE APPLIQUEES A DES CORPUS DE L'ENVIRONNEMENT ET DE L'ENERGIE

« Restitution, prévision et anticipation d'événements par poly-résonances croisées »

Résumé

Cette thèse propose une série de méthodes de veille textométrique multilingue appliquées à des corpus thématiques. Pour constituer ce travail, deux types de corpus sont mobilisés : un corpus comparable et un corpus parallèle, composés de données textuelles extraites des discours de presse, ainsi que ceux des ONG. Les informations récupérées proviennent de trois mondes en trois langues différentes : français, anglais et chinois. La construction de ces deux corpus s'effectue autour de deux thèmes d'actualité ayant pour objet, l'environnement et l'énergie, avec une attention particulière sur trois notions : les énergies, le nucléaire et l'EPR. Après un bref rappel de l'état de l'art en intelligence économique, veille et textométrie, nous avons exposé les deux sujets retenus, les technicités morphosyntaxiques des trois langues dans les contextes nationaux et internationaux. Successivement, les caractéristiques globales, les convergences et les particularités de ces corpus ont été mises en évidence. Les dépouillements et les analyses qualitatives et quantitatives des résultats obtenus sont réalisés à l'aide des outils de la textométrie, notamment grâce aux analyses factorielles des correspondances, réseaux cooccurrentiels et poly-cooccurrentiels, spécificités du modèle hypergéométrique, segments répétés ou encore à la carte des sections. Ensuite, la veille bi-textuelle bilingue a été appliquée sur les trois mêmes concepts dans l'objectif de mettre en évidence les modes selon lesquels les corpus multilingues à caractère comparé et parallèle se complètent dans un processus de veille plurilingue, de restitution, de prévision et d'anticipation. Nous concluons notre recherche en proposant une méthode analytique par Objets-Traits-Entrées (OTE).

Mots-clés : textométrie, veille multilingue, opinions, corpus comparable, corpus parallèle, discours de presse, discours des ONG, fouille textuelle, cooccurrences, poly-cooccurrences, nucléaire, EPR, énergies, environnement.

Abstract

This thesis proposes a series of textometric multilingual information monitoring methods applied to thematic corpora (textometry is also called textual statistics or text data analysis). Two types of corpora are mobilized to create this work: a comparable corpus and a parallel corpus in which the textual data are extracted from the press and discourse of NGOs. The information source was retrieved from three countries in three different languages: English, French and Chinese. The two corpora were constructed on two topical issues concerning the environment and energy, with a focus on three concepts: energy, nuclear power and the EPR (European Pressurized Reactor or Evolutionary Power Reactor). After a brief review of the state of the art on business intelligence, information monitoring and textometry, we first set out the two chosen subjects – the environment and energy – and then the morphosyntactic features of the three languages in national and international contexts. The overall characteristics, similarities and peculiarities of these corpora are highlighted successively. The recounts and qualitative and quantitative analyses of the results were carried out using textometric tools, including factor analysis of correspondences, co-occurrences and polyco-occurrential networks, specificities of the hypergeometric model and repeated segments or map sections. Thereafter, bilingual bitextual information monitoring was applied to the same three concepts with the aim of elucidating how the comparable corpus and the parallel corpus can mutually help each other in a process of multilingual information monitoring, by restitution, forecasting and anticipation. We conclude our research by offering an analytical method called Objects-Features-Opening (OFO).

Keywords: textometry, multilingual information monitoring, opinions, comparable corpus, parallel corpus, media discourse, discourse of NGOs, text mining, co-occurrences, poly-cooccurrences, nuclear, EPR, energy, environment

Remerciements

J’adresse ma profonde reconnaissance à André Salem pour la confiance et la très grande liberté qu’il m’a accordées tout au long de ce travail. Ses encouragements, mais aussi ses critiques ont largement contribué à l’aboutissement de cette thèse.

Mes remerciements vont également aux membres du jury, Madame Annie Bertin, Madame Monique Slodzian et Monsieur Pierre Zweigenbaum.

Je remercie Raoul Berger et Kun JIN, pour leur soutien moral et leurs conseils techniques.

Je remercie chaleureusement la famille Guillot, Bernard Sargnon, Monique Suszylo, Patricia Poude, Jean-Michel Poude, ainsi que tous mes anciens professeurs et camarades de l’INaLCO, Paris 3 et Paris 10.

Merci à tous mes amis qui m’ont encouragé, donné des conseils et soutenu tout au long de ce travail.

Je remercie toute ma famille qui m’a toujours soutenu malgré la grande distance qui nous sépare.

Sommaire

Introduction générale

Le monde, qui utilise des centaines de langages depuis des millénaires, a formalisé les mots et les grammaires pour transcrire, enseigner et transmettre sur des supports, les savoirs, les faits et les pensées. Des hiéroglyphes aux idéogrammes, en passant par les alphabets, ces représentations diffusent ainsi l'image du monde à travers les époques, les évolutions, les mœurs et les courants de pensée. Cela représente aujourd'hui des centaines de milliards de mots dans des corpus différents, avec des occurrences variables. Il n'est pas possible à un être humain d'aborder par lui-même la masse des publications archivées ou en circulation.

Seul l'usage de l'informatique peut, à présent, dans le cadre de la mondialisation, permettre un balayage massif des séquences des corpus nécessaire à l'étude des occurrences et des usages des mots, au moins dans les langues essentielles diffusant le savoir, l'information et la communication entre les humains. L'utilité de ces recherches est étendue, allant des besoins sociaux, humains, scientifiques aux guerres économiques, en passant par les médias et les enjeux stratégiques des politiques. C'est la capacité à détecter, enregistrer, analyser et comprendre dans les meilleurs délais, qui va permettre aux différentes forces de pouvoirs d'anticiper les décisions et d'agir efficacement.

Cette force de veille, implantée de manière continue et basée sur des outils performants, élaborés et mis en œuvre par des chercheurs, des informaticiens, des stratèges, des économistes, sous l'autorité des décideurs... va donc construire les forces de demain, parfois à l'échelle de la planète.

Cette thèse prétend essayer d'apporter un éclairage sur ces besoins et sur les évolutions présentes ou à venir, sur la base de trois langues majeures, dans des domaines cruciaux de l'avenir du monde, et de décrire la conception et l'usage d'outils puissants en perpétuelle amélioration, afin de s'adapter tant aux besoins des utilisateurs qu'aux possibilités nouvelles des technologies.

La mise en réseaux des objets connectés au moyen des nouvelles technologies de l'information et de la communication permet l'accès à une gigantesque banque de données, d'où l'engouement autour du phénomène Mégadonnées (*Big Data*). D'une part, tous les milieux de la société, qu'ils soient de nature sociétale, gouvernementale, économique, politique, etc. contribuent à la génération et à la prolifération de ces données disponibles en ligne ; d'autre part, ces réseaux exploitent et transforment les valeurs de ces informations émises par les utilisateurs aux origines diverses et variées.

Sous l'impulsion des hautes technologies de nombreuses sciences, les fournisseurs de contenu se voient recommander d'indexer leurs bases avec des métadonnées et des taxinomies ou ontologies. Grâce à cela, les moteurs de recherche tels que *Google*, *Baidu*, *Yahoo*, *Bing* (*Microsoft*) étayés habituellement par les recherches de textes, évoluent progressivement vers le web sémantique.

«Nous ne cherchons jamais les choses, mais la recherche des choses.»
- *Pensées,* Blaise Pascal

La frénésie autour de ce phénomène métamorphose tous les services et les comportements de notre époque. La rapidité, l'accessibilité et la pertinence de ce nouvel univers de données engendrent des bénéfices prodigieux, notamment pour restituer, prévoir et anticiper des événements.

Sans retracer les définitions exhaustives de ces trois termes, nous en rappelons les significations essentielles :

- la restitution, c'est le rétablissement des informations dans leur état premier ou original,
- la prévision, c'est « *l'observation d'un ensemble de données qui permet d'envisager une situation future et d'entreprendre des actions pour y parer concrètement*»[1],
- l'anticipation, c'est la prévision du futur comme si un événement s'était déjà produit, et sa réalisation par des mesures concrètes dans l'objectif de maîtriser la situation.

Les avancées de ces outils de méga-information faciliteront :

- la gestion identitaire croisée des utilisateurs,
- l'accès aux informations issues du monde numérique (l'internet) avec celles du monde classique (radio, presse écrite, télévision, etc.),
- la confluence des informations opérationnelles et commerciales,
- l'agrégation des connaissances des usagers dans les systèmes d'information.

Toutefois, les défis demeurent colossaux. Les premiers résident dans la question du volume, de la vélocité, de la variété et de la véracité, défis soumis aux calculabilités computationnelles de ces masses titanesques de données. D'un côté, la sécurité et en particulier la confidentialité des données privées et publiques freinent leur diffusion ; de l'autre, les mégadonnées nécessitent une ouverture et une transparence constamment croissantes, et exigent une liberté d'expression et d'information. Ces deux tentacules s'articulent dans une sphère de guerre de l'information et sont le revers de la médaille qui mène les mégadonnées dans une dichotomie.

> « *Les langues sont un trésor et véhiculent autre chose que des mots. Leur fonction ne se limite pas au contact et à la communication. Elles constituent d'une part des marqueurs fondamentaux de l'identité, elles sont structurantes, d'autre part, de nos perspectives*. » (Serres, 1996)
>
> - Michel Serres

Dans un contexte de sociétés mondialisées, on peut parler de multilinguisme ou encore de plurilinguisme. La traduction devient alors un élément capital pour la communication entre les peuples. Une bonne traduction garantit la qualité de la transmission de toutes les informations. Cependant, devant la gageure que constitue le projet de réaliser une veille multilingue, peut-on utiliser simplement la traduction ?

Le multilinguisme dénote la coexistence des notions identiques et/ou similaires cryptées et véhiculées dans les différentes langues.

Dans l'optique de maîtriser les données plurilingues dans notre ère devenue de plus en plus numérique, nous proposons trois parcours afin de mieux cerner les notions multilingues :

1. parcours traitant les perceptions et connotations d'une langue à une autre,
2. parcours examinant les interprétations et appréhensions les unes des autres,
3. parcours développant les réactions provoquées par une langue par rapport aux autres.

Plus concrètement, il faut savoir comment repérer, déceler et valoriser les informations issues de multiples langues à l'intersection de ces trois parcours, et comment ces données plurilingues transitent dans les réseaux.

[1] Définition du TLFi (http://www.cnrtl.fr/definition/pr%C3%A9vision)

Introduction

Il faut commencer par une veille thématique multilingue de restitution, prévision, anticipation et décision. Aussi, nous nous reposons sur une herméneutique de textes thématiques constituée et examinée à partir de preuves scientifiques.

Définir la veille correspond à mettre en œuvre une chaîne d'activités analytiques et pratiques avec des objectifs précis qui permettent la fouille d'informations dans le passé, afin d'apporter de l'aide pour le présent mais aussi d'anticiper et de prévoir des actions pour le futur. Dans le cadre de notre recherche, nous allons exhumer un certain nombre d'événements portant des valeurs stratégiques autour des énergies et de l'environnement, plus particulièrement sur les réacteurs nucléaires EPR. Notre travail a mis à jour des informations sur la prévision de l'organisation technique et financière du projet EPR d'Hinkley Point en Angleterre, sujet d'actualité.

Si à ce jour, bon nombre d'acteurs déclarent participer à la veille, peu de métriques pragmatiques sont néanmoins mises en place. La restitution, la prévision et l'anticipation sont quasi-inexistantes en l'absence de véritables modèles prédictifs.

« *On ne peut se passer d'une méthode pour se mettre en quête de la vérité des choses.* »

- René Descartes

Pour ce faire, nous proposons un ensemble de méthodes organisationnelles, statistiques, informatiques, linguistiques et analytiques, pour atteindre ces objectifs multilingues, transdisciplinaires et trans-heuristiques[2].

Les méthodes de la textométrie multilingue deviennent désormais une aide indispensable pour les analyses de vastes corpus textuels rédigés dans des langues distinctes par une quantité croissante d'acteurs, médias informationnels, partis politiques, associations culturelles, blogs et forums. Parallèlement, la communauté scientifique qui se préoccupe du texte écrit, tente de maîtriser la profusion des normes de stockage de ces corpus de textes (*Unicode*, *Xml*, *Tei*, *Dublin Core*, *Olac*, etc.). Ces avancées structurent les tâches de veille technologique devenues partie intégrante de l'activité des entreprises modernes et innovantes, et s'élargissent à l'ensemble des acteurs de l'économie mondiale, où la préoccupation de recherche d'informations à caractère industriel et commercial, demeure cruciale pour le rayonnement et la compétitivité régissant leur avenir. Elles mettent en évidence la manière dont sont perçus et formulés les principaux concepts de chaque domaine dans l'approche de cultures et langues aussi bien proches qu'éloignées. Elles aboutissent à la constitution de ressources traductologiques et lexicographiques (dictionnairiques selon Monique Slodzian) adaptées à chaque secteur de recherche.

Le travail concerne les méthodes d'extraction, de sélection et d'intégration rationnelle dans des corpus de documents fournis par nos programmes et le développement de méthodes textométriques pour l'exploration de corpus hétérogènes (supports informationnels différents, textométrie multilingue, etc.). L'objectif est de simplifier les procédures de veille informationnelle. Nous utiliserons deux corpus, un comparable et un parallèle de textes français anglais et chinois, rassemblés sur le réseau internet au sujet de questions d'environnement et de l'énergie.

« *Pour atteindre la vérité, il faut une fois dans la vie se défaire de toutes les opinions qu'on a reçues, et reconstruire de nouveau tout le système de ses connaissances.* »

- René Descartes

[2] Qui sert à la découverte de différentes informations de manière transversale.

Dans le but de mettre en œuvre ces méthodes, le champ d'application retenu se rapporte à des sujets relatifs aux énergies et à l'environnement, deux sujets qui inquiètent les opinions publiques. Ce travail de doctorat est l'un des premiers à s'intéresser à la veille multilingue sur ces thèmes d'actualité à portée internationale.

Plus précisément, nous nous attachons à la statistique textuelle ou textométrie, trilingue, en français, anglais et chinois, appliquée aux corpus spécialisés de ces deux domaines disponibles sur la toile. Nous apportons et décrivons une série de méthodes pratiques et efficaces applicables à nombre de disciplines.

Le monde se trouve face à un véritable défi planétaire, d'une part, lutter contre le changement climatique et les conséquences dramatiques de la crise énergétique sans précédent qui ont fortement contaminé et dégradé l'environnement durant les dernières décennies, d'autre part, concevoir un nouveau modèle de développement économico-social défini, politiquement, désormais comme durable. En ce début du XXI^{e} siècle, certains pays dont la Chine, ont eu la perspicacité et l'intelligence d'opter pour de nouvelles orientations politico-énergétiques et de s'intéresser au progrès et au déploiement du nucléaire dans une période de transitions énergétiques et de remise en cause du choix des énergies, dans un climat de contestation nécessitant des argumentations convaincantes et rassurantes. Dans l'intention de saisir les différentes visions concernant ces deux thèmes, nous avons choisi d'analyser les opinions des trois mondes économiques : la France, les États-Unis et la Chine.

Dans un sujet aussi vaste que ces deux thèmes, nous consacrons notre veille trilingue à trois objets d'études, *énergie(s)*[3], *nucléaire* et *EPR*. Leurs relations sémantiques, hyperonymie, hyponymie holonymie et méronymie, les regroupent en structures ontologiques.

Cerner ces trois notions liées aux impacts politico-économiques par le biais des nouvelles technologies de la veille, nécessite à la fois des réponses proactives et des efforts concertés de la part de la communauté internationale. En dépit de sa confidentialité, la complexité et l'évolution de ces notions se détectent dans la presse. Peut-on appréhender la résonance textuelle de ces objets par la veille multilingue ? S'agissant d'un pays aussi étendu et aussi diversifié que la Chine, les études comparatives avec la France et les États-Unis peuvent paraître primordiales pour apporter la réponse.

Selon la classification du thésaurus classique, *énergie(s)* est un terme taxinomique, partie intégrante de l'environnement et du changement climatique, mais aussi un hyperonyme et un holonyme du *nucléaire*, alors qu'*EPR* reste un hyponyme des réacteurs de la troisième génération du *nucléaire*.

Cependant, ces trois concepts sont quantitativement et qualitativement instables dans une sémantique multilingue empirique. Nos travaux résolvent entre autres cette problématique.

Il devient alors possible d'expliquer comment ces trois notions se répercutent dans les trois langues, comment elles sont perçues, quelles sont leurs connotations et interprétations intrinsèques, quelles sont les actions événementielles révélées par la presse et le discours des ONG, et quelles sont leurs sémantiques ontologiques en tenant compte de leurs contextes linguistiques. La visualisation graphique de l'exploitation des données en facilite la compréhension.

Les constats et analyses à travers des calculs de cooccurrences et poly-cooccurrences des formes-pôles, comme 核能/hé néng/*énergie(s), energy*, *nucléaire, nuclear*, et *EPR* ont démontré de très nombreuses occurrences de la part des médias et des peuples du monde autour de l'événement planétaire de Fukushima. Ainsi, nous décidons de focaliser notre veille sur la période comprise de 2010 à 2012, dans l'objectif de retracer les opinions et comportements, avant, pendant et après la catastrophe qui a bouleversé notre vision du nucléaire, entrainant vraisemblablement un changement

[3] Tous les mots écrits en italique sont des formes issues soit de la littérature, soit des résultats de nos calculs et raisonnements.

politique et des impacts sur l'avenir de l'Atome. L'étude s'achève par une analyse continue du propos des ONG de 2012 à 2014.

Cette thèse s'organise en trois parties et se divise en sept chapitres.

La première partie est consacrée aux notions associées aux différents domaines de notre travail, en commençant par la description de l'état de l'art en matière de veille économique, d'intelligence économique et de textométrie. Nous rappelons ainsi le cadre conceptuel de notre veille textométrique multilingue.

La deuxième partie aborde de manière générale les thèmes de notre recherche, les énergies et l'environnement dans les trois pays choisis. Nous exposons ensuite les principaux aspects économico-culturels des trois pays, France, Etats-Unis et Chine pour lesquels certains médias font l'objet de notre étude veille textométrique multilingue. Le processus de récolte des données disponibles, relatives aux spécificités des trois langues, est ensuite décrit avec une attention particulière. Par la suite, nous construisons nos deux corpus, le comparable et le parallèle à partir des articles provenant des trois pays en vue de mettre en œuvre le processus de veille proprement dit.

Enfin, dans une troisième partie, les corpus constitués, ENRG_FR, ENRG_US, ENRG_CN, CLRG_CN et CLRG_EN font l'objet d'analyses textométriques. Nous nous intéressons aussi à l'emploi des formes 核能/hé néng/*énergie(s), energy*, *nucléaire, nuclear*, et *EPR* dans ces corpus.

La partie une est composée d'un unique chapitre (le chapitre 1), qui étudie le cadre conceptuel et la recherche de méthodes pour la textométrie multilingue. Il s'agit de rappeler les différentes acceptions des termes intelligence économique et veille multilingue et de voir en quoi celles-ci diffèrent. Nous décrivons les outils existants et proposons un processus à mettre en œuvre pour réaliser une veille multilingue.

La deuxième partie consiste à exposer les contextes transdisciplinaires de notre recherche, à savoir, les deux principaux thèmes : les énergies et l'environnement, les trois sphères de communication, la France, les États-Unis et la Chine, enfin les trois langues pratiquées majoritairement dans ces trois pays avec leurs spécificités culturelles et techniques, le français, l'anglais et le chinois. Par la suite, nous présentons nos deux types de corpus, corpus comparable et corpus parallèle, qui se déclinent en trois sous-corpus et un corpus en deux volets. Il est à noter que le corpus parallèle est constitué à partir d'un site ONG dans lequel les rédactions en anglais présentent à la fois les formes issues de l'anglais européen et celles de l'anglo-américain.

Le chapitre 2 est consacré à nos deux thèmes de travail, thèmes faisant aujourd'hui l'objet de nombreux débats passionnés.

Le chapitre 3 explore les trois mondes dans leurs contextes socio-économiques et cerne les langues des trois pays. Dans cette perspective, nous examinons d'abord les différences et les points communs de ces pays d'un point de vue géopolitique. Nous serons amenés à évoquer la situation de leur presse, et ainsi à justifier les choix de nos corpus. Nous verrons ensuite que les langues ont des systèmes linguistiques bien distincts, différences linguistiques se retrouvant aussi bien dans la phonologie que dans la syntaxe. En particulier, la langue chinoise se caractérise par un système graphique fruit d'une longue histoire, très différent du français et de l'anglais. Ce système complique ainsi le processus de segmentation de cette langue. Différents logiciels de segmentation du chinois existent cependant. Nous rappelons les fondements algorithmiques et statistiques de ces outils, nous les mettons ensuite en œuvre sur des textes de notre corpus, et en analysons les résultats.

Le chapitre 4 explicite le mode de constitution de nos deux corpus issus des cinq principales sources, qui sont :

- le journal *Le Monde*
- le journal *New York Times*
- le site internet *qq.com*
- le site internet *sina.com.cn*
- le *chinadialogue.net* (ONG)

La troisième partie se veut exploratoire, analytique, empirique et pragmatique. Les méthodes et outils textométriques (statistique textuelle), informatiques et analytiques seront appliqués sur les deux corpus afin de répondre à nos questions pré-exposées.

Le chapitre 5 applique la fouille textométrique à la fois d'un point de vue qualitatif et quantitatif aux sous-corpus français et anglais en tenant compte des critères de comparaison retenus. Des remarques analytiques permettant de mieux cibler les recherches multilingues sont proposées.

Quant au chapitre 6, celui-ci analyse seulement le sous-corpus chinois, traité à part, en raison de ses caractéristiques linguistiques, d'un point de vue qualitatif et quantitatif en tenant compte des critères de comparaison retenus.

Enfin le chapitre 7 aborde les textes du corpus parallèle bilingue anglais-chinois issus du site *chinadialogue.net*, en les étudiant et les comparant avec le corpus comparable. Nous terminons cette étude par une proposition d'une méthode analytique sémantique multilingue, appelée Objets-Traits-Entrées (OTE).

Partie 1 Etat de l'art : veille, intelligence économique, textométrie

« Les opérations de guerre de l'information se répartissent dans le domaine économique en 3 catégories :

1. *la Tromperie : (Désinformation, Manipulation, discrédit),*
2. *la Contre-information : (Identification des points faibles de l'adversaire, Exploitation de ses contradictions, frapper ses talons d'Achille, Utilisation de l'information vérifiable),*
3. *la Résonance : (Faire de l'agit-prop, Optimiser les caisses de résonances, Créer des réseaux d'influence, animer des forums de discussion…) »*

- Christian Harbulot, Les principes de la guerre de l'information, Doctrines (2001)

« L'intelligence est le croisement de l'information et de la stratégie. Le prisme est large. Il va du cycle du renseignement - dont la définition « officielle » de l'intelligence économique s'est inspirée - à la manipulation de la connaissance en passant par la désinformation. Dans tous les cas, l'information est au service d'une stratégie : en amont pour définir et comprendre son environnement pertinent, prévenir les risques, détecter les opportunités…; en aval pour décider, leurrer l'adversaire, le paralyser, … »

- Nicolas Moinet, *La guerre cognitive : l'arme de la connaissance* (2002)

« La justice de l'intelligence est la sagesse. Le sage n'est pas celui qui sait beaucoup de choses, mais celui qui voit leur juste mesure »

- Platon

Le but de ce chapitre est d'étudier le cadre conceptuel et la recherche de méthodes pour la textométrie multilingue. Il s'agira dans un premier temps de rappeler les différentes acceptions des termes intelligence économique et veille multilingue et de voir en quoi ils diffèrent. Nous étudierons ensuite plus en détail la veille multilingue, décrirons les outils existants et proposerons le processus à mettre en œuvre pour réaliser ce type de veille. Ce processus inclut en particulier la nécessité de mettre en place des équipes pluridisciplinaires :

- spécialistes des thèmes des corpus (ici les énergies, l'environnement et le nucléaire),
- statisticiens et informaticiens (pour définir et mettre en œuvre les méthodes statistiques et informatiques d'informations textuelles),
- linguistes.

Nous apporterons ensuite des rappels sur les notions de corpus, de corpus multilingue et d'alignement.

1. Cadre conceptuel de recherche et méthodes pour la textométrie multilingue

Les statistiques textuelles permettent d'aborder le texte sous l'angle de la fréquence des différentes unités qui le composent ; elles consistent à quantifier les mots et analyser les graphies, résultats obtenus par les comptages de ces derniers dans des corpus de textes afin de dégager du sens, tout en évitant de s'attacher trop rapidement à une sémantique ou à une grammaire de textes. Il s'agit là d'une rupture avec des méthodes qualitatives plus classiques. En travaillant sur la surface textuelle, cette méthode permet de distinguer les formes textuelles qui sont proportionnellement sur-employées ou sous-employées dans les différentes parties de l'ensemble du corpus. Vole alors en éclats la traditionnelle mais artificielle distinction entre le quantitatif et le qualitatif. En effet, les questions de sens n'interviennent qu'après cette confrontation des unités, les unes aux autres, lorsque l'analyste cherche à interpréter en quoi ces dernières sont sur ou sous-employées, peuvent être indice (Bonnafous et Tournier, 1995), ou en quoi elles sont prédictives de nouveaux comportements. L'objectif majeur de la statistique textuelle est alors de donner la parole à chacune de ces unités (occurrences/formes). Les fameux « points et individus » ne sont donc plus muets, ils parlent (Lebart et Salem, 1994).

Tenter d'observer le style propre à un auteur ou à une période ou encore comparer des comportements textuels dans des langues différentes sans les traduire ni les coder, sont autant de problématiques auxquelles l'analyse statistique apporte des solutions. Cette méthode s'adresse donc à ceux pour lesquels les recherches doivent décrire, comparer, classer, analyser des ensembles textuels. Il peut s'agir de genres très variés : littéraires, scientifiques, économiques, sociologiques (réponses aux questions ouvertes dans des enquêtes socio-économiques, entretiens divers en marketing, psychologie appliquée, pédagogie, médecine), ou encore de textes historiques, politiques.

Nous allons tenter de faire le point sur les apports de la statistique textuelle, en présentant des applications dans l'éventail des disciplines concernées. Les exemples qui suivent voudraient tout en montrant l'acquis de ce champ disciplinaire, témoigner de la richesse d'approches, de méthodologies mais aussi de domaines.

En exposant la problématique de cette recherche, cette partie est consacrée aux notions et définitions évoquées afin de définir le périmètre du travail, d'en établir un état de l'art pour y introduire les méthodes utilisées.

La révolution méthodologique survenue dans les années soixante-dix a mis à la disposition de larges communautés de chercheurs, dont celle des Sciences du langage, des méthodes permettant de soumettre les textes saisis sur support numérique à des procédures d'analyse informatisée dans le domaine du traitement automatique des langues, dont la *textométrie*. Elle s'appelle aussi logométrie ou statistique textuelle, c'est la forme actuelle de la lexicométrie (Lebart et Salem, 1994).

La textométrie, au carrefour de la linguistique, des statistiques et de l'informatique apporte des ensembles d'éléments langagiers classifiés, intrinsèquement liés ou pas, tant tangibles qu'abstraits, ensembles computationnels à manipuler au-delà de chacune de ces trois disciplines. La textométrie s'applique à l'ensemble des disciplines des sciences humaines et sociales et permet l'accès à toutes les informations textuelles.

La fouille d'informations textuelles, appelée fouille de textes ou « *Text Mining* » en anglais, est un ensemble de modèles de traitements informatiques et statistiques appuyés sur les théories de la linguistique, permettant d'exploiter et de classer les données textuelles structurées ou non structurées.

Une deuxième période s'ouvre actuellement, qui voit la généralisation planétaire de la mise à disposition du public de textes constamment réactualisés, rédigés dans toutes les langues du monde par une quantité croissante d'organismes institutionnels divers, de médias informationnels, de partis politiques, d'associations culturelles diverses, voire d'individus s'exprimant de manière isolée (*blogs*), ou participant à des débats dans les réseaux sociaux *(Facebook, Twitter, etc.)*.

Après une longue période d'hésitations de la communauté scientifique en matière de conservation des textes écrits dans les différentes langues du monde, les systèmes informatiques sont en train de maîtriser cette profusion, tant pour ce qui concerne le stockage des corpus de textes écrits dans des langues différentes que pour ce qui concerne leurs méthodes de restitution et de gestion (*Unicode, XML, TEI, Dublin Core, OLAC, etc.*).

Ces nouvelles avancées permettent d'envisager sous un jour nouveau la plupart des tâches de *veille technologique* devenues partie intégrante de l'activité de recherche et de communication (Mercier et Charon, 2004) des acteurs économiques modernes.

1.1 Veille et intelligence économique, une concurrence sémantique

L'évolution fulgurante de la gestion technologique de l'information et du traitement automatique des langues, fortement influencée par la mondialisation, conduit les différents partenaires, qu'ils soient privés ou étatiques, à se doter de véritables stratégies pour mettre en place les diverses activités de veille, activités devenues indispensables dans le contexte politico-économique actuel et ainsi à s'organiser efficacement afin de prendre les bonnes orientations et les décisions adéquates (Lesca, 1986-1990, 1989, 1992, 1995, 1996 ; Lesca et Schuler, 1998).

Les nouvelles technologies de l'information et de la communication (NTIC) fournissent des outils de plus en plus puissants et performants, technologies devant être adaptées et mises au service des différents partenaires (Libaert et Westphalen, 1999) afin de leur assurer la meilleure compétitivité face à une concurrence de plus en plus rude (Lesca et Raymond, 1993 ; Lesca et Caron, 1995 ; Lesca et Chokron, 2000).

Les paragraphes suivants font appel à la méthode analytique QQOQCCP[4] et tentent de cerner par l'approche scientifique, dite hypothético-déductive[5], les notions de la veille versus l'intelligence économique, la veille multilingue, le traitement automatique des langues et la textométrie dans nos sociétés mondialisées.

1.1.1 La veille stratégique

Selon la norme Afnor XP X 50-053, « Prestations de veille et prestations de mise en place d'un système de veille », la veille est décrite comme une *« activité continue et en grande partie itérative visant à une surveillance active de l'environnement scientifique, technologique, juridique, commercial, socio-politique etc., pour en anticiper les évolutions ».*

La veille est désormais une pratique courante devenue incontournable pour tous les acteurs économiques. La veille a fait l'objet de nombreuses autres définitions, comme par exemple la suivante où la veille est une *« activité de surveillance permanente de l'environnement interne d'une organisation qui doit permettre un repérage de signaux et de signes révélateurs de changements importants »* (Guidère, 2008).

Mais cette veille présente un aspect stratégique pour les sociétés.

> *« La fonction de veille est désormais considérée comme stratégique parce qu'elle permet à une entreprise, à une organisation ou à une institution de se mettre à l'écoute de son environnement mondialisé pour prendre les décisions adéquates et agir de façon ciblée pour la réalisation de ses objectifs »* (Guidère, 2008).

L'expression veille stratégique est *« une expression générique qui englobe plusieurs types de veilles spécifiques telles que la veille technologique, la veille concurrentielle, la veille commerciale, etc. »* (Lesca, 1997). Une entreprise n'est pas forcément dans la nécessité de mettre en œuvre toutes ces veilles spécifiques[6].

Cette expression reste dominante par son utilisation voire abusée pour remplacer tout type de veille dans le langage courant. Cependant, quelques chercheurs ont tenté de mener des travaux d'éclaircissement sur le sujet (Frion, 2004, 2012).

La veille stratégique se rattache donc au management stratégique (Martinet et Petit, 1982) dans le sens où il consiste à mobiliser, combiner et engager des ressources à des fins d'efficience, d'efficacité et de réduction d'incertitudes afin de créer des opportunités, de détecter suffisamment tôt des menaces et de réduire l'incertitude des dirigeants (Koenig, 1990, 1996).

Afin de bien cerner le concept de veille stratégique et les spécificités de ce système d'information stratégique (Marmuse, 1992 ; Merland, Binot et al, 2005), rappelons la définition de Lesca : *« la veille stratégique est le processus informationnel volontariste par lequel l'entreprise se met à l'écoute anticipative (ou prospective) des signaux précoces de son environnement socio-économique dans le but créatif d'ouvrir des opportunités et de réduire les risques liés à son incertitude »* (Lesca, 1994).
Finalement, l'objectif de la veille stratégique est *« de permettre d'agir très vite et au bon moment. Les anglo-saxons utilisent les expressions Environmental Scanning et Competitive Intelligence pour désigner des concepts très voisins »* (Lesca, 1997).

[4] « Qui ? Quoi ? Où ? Quand ? Comment ? Combien ? Pourquoi ? », du latin : « Quis, Quid, Ubi, Quibus auxiliis, Cur, Quomodo, Quando » : 7 questions énoncées par Quintilien, rhéteur latin, lors d'une plaidoirie, qui définissent, entre autres, les « circonstances » d'une situation (en latin : « circum-stare » désigne « ce qui se tient autour de »).
[5] Voir le glossaire.
[6] Se reporter à l'annexe A, section A.1, La veille.

De façon imagée, la veille stratégique de l'entreprise peut être comparée au radar d'un navire (Aguilar, 1967), puisqu'elle vise à anticiper des événements avant qu'il ne soit trop tard pour pouvoir agir. Cependant, à la différence du radar, la veille stratégique est caractérisée par son aspect interprétatif, voire constructiviste, et permet à l'entreprise de détecter des signes annonciateurs de changements et de se préparer à leur venue (Lesca, 2008).

La mise en place d'une veille *cf. supra* dans toutes les organisations économiques privées ou étatiques serait difficile selon Lesca : *« La problématique de l'utilité est sans cesse soulevée par les responsables d'entreprise. Elle englobe plusieurs questions telles que : utilité pour quoi faire ? Utilité pour qui, au sein de l'entreprise ? Utilité de quoi : du concept même de veille stratégique ou d'un dispositif spécial pour la veille stratégique qui permettrait de passer du concept à l'action ? »* (Lesca, 1997).
Comme Vauban en son temps disait : *« Celui qui ne veut rien faire trouve toujours de bonnes raisons pour ne rien faire »*.

Le processus de veille stratégique peut *« fonctionner selon deux modes distincts mais non exclusifs : le mode commande et le mode alerte »* (Lesca, 1997).

1.1.2 L'intelligence économique

Quant à la notion d'I.E, en France, elle est « *le fruit d'une réflexion atypique* » développée de façon empirique dès la fin des années 1980, grâce notamment à l'action conjuguée de personnalités issues de milieux très variés (universitaires, fonctionnaires, représentants du monde de l'entreprise, de la défense nationale, etc.) *« ... en marge de l'institution et du monde de l'entreprise»* (Harbulot, 2004). Son développement s'est véritablement accéléré en 1994 à partir du rapport « Intelligence économique et stratégie des entreprises », une œuvre collective du Commissariat général du Plan, organisme disparu en 2006. Ce rapport en donne la définition suivante : *« L'intelligence économique peut être définie comme l'ensemble des actions coordonnées de recherche, de traitement et de distribution en vue de son exploitation, de l'information utile aux acteurs économiques »* (Martre, 1994). Par la suite, afin de renforcer cette notion dans le paysage français, d'autres structures gouvernementales ont été également créées, par exemple, le Secrétariat général de la défense nationale (Juillet, 2004).

L'approche de ce concept s'est focalisée sur les objectifs des entreprises (Mousnier, 2005) (innovation, mieux vendre, fabrication de produits de bonne qualité, etc.) et a été influencée par les écoles de gestion américaines (Harbulot et Baumard, 1997). Mais selon l'époque, le contexte géopolitique, les approches seront différentes. Comme par exemple, les écrits de l'ingénieur allemand Herzog rédigés au moment de la première guerre mondiale, énumèrent les moyens d'action à mettre en œuvre pour préserver les intérêts de l'Allemagne vis à vis des autres belligérants, *« ...on suivra toutes les inventions et perfectionnements techniques réalisés à l'étranger, pour les porter à la connaissance de ces industriels allemands qu'ils peuvent intéresser »* (Herzog, 1915).

Une des premières définitions anglo-saxonnes de l'intelligence économique moderne date de 1967 (Wilensky, 1967). L'intelligence économique est alors définie comme l'activité de production de connaissance servant les buts économiques et stratégiques d'une organisation, recueillie et produite dans un contexte légal et à partir de sources ouvertes. De nombreuses définitions ont suivi, définitions évoluant selon les contextes politiques, économiques et mondiaux[7].

[7] Se reporter à l'annexe A, section A.2 : panorama de l'I.E en France : date et définitions

Pour un acteur économique, l'I.E peut se résumer de la manière suivante : *«… ensemble des moyens qui, organisés en système de management par la connaissance, produit de l'information utile à la prise de décision dans une perspective de performance et de création de valeur pour toutes les parties prenantes… »*[8].

1.1.3 Veille stratégique ou intelligence économique

La veille et l'intelligence économique (Boizard, 2005), deux pratiques en symbiose, permettent de repérer, collecter, traiter, stocker et organiser les informations et les événements de façon à ce que les acteurs économiques et tous leurs partenaires puissent mieux définir les perspectives stratégiquement optimales.

Qu'est-ce que la veille stratégique ? C'est le radar de l'entreprise ! Qu'est-ce que l'intelligence économique ? C'est fournir la bonne information, au bon moment, à la bonne personne pour lui permettre de prendre la bonne décision, de bien agir, et idéalement de faire évoluer son environnement dans le bon sens. Dans quel but ? Celui d'ouvrir des fenêtres d'opportunités et de réduire les risques liés à l'incertitude (Hermel, 2010).

> La veille est de *« ... mettre en place des systèmes de collecte et de validation de l'information fait appel à des connaissances simples et des moyens techniques à la portée organisationnelle, intellectuelle et financière de toute entreprise quelle que soit sa taille. L'intelligence économique se veut plus active et cherche à influencer son environnement. Elle met en jeu un nombre important de concepts pas toujours faciles à appréhender pour le profane. Elle implique aussi un grand nombre d'acteurs et a donc un coût »* (Boizard, 2005).
>
> *« La veille, par son flot d'information et sa surveillance systématique provoque un engorgement des neurones, retarde la décision car les veilleurs attendent la meilleure information possible avant de livrer leur synthèse. Enfin, elle engendre un phénomène d'accoutumance alors que l'intelligence économique est au contraire une situation ponctuelle de réactivité à une stratégie envisagée, entraînant une recherche intensive mais limitée dans le temps d'information stratégique… La veille répond à un besoin d'information alors que l'intelligence économique répond au besoin de décision »* (Frion, 2012).

La veille comme l'I.E produisent des masses d'information, informations qui n'ont pas toutes le même degré d'accessibilité, ni la même provenance[9]. En effet, la veille est de recueillir un maximum d'informations alors que l'I.E correspond à une prise de décision (Etienne et al, 2003).

Dans les deux cas, l'information est la *« matière première »* (Etienne et al, 2003) et devient primordiale pour l'ensemble des acteurs. L'information se présente sous diverses formes et son origine est multiple (informations blanches ou grises et des moyens légaux, voire éthiques)[10], mais cette information peut être ouverte ou fermée : « *l'information fermée représente près de 10% de l'information globale utile à l'intelligence économique* » (Besson et Possin, 2001, 2002).

Les entreprises doivent développer leurs capacités à s'adapter à leur environnement face à la concurrence mondiale. L'information est précieuse et se doit d'être traitée en permanence, mais cette information doit être maîtrisée et exploitée à bon escient dans l'intérêt de l'entreprise. L'I.E couplée à

[8] Définition retenue par Association Française pour le Développement de l'I.E (AFDI.E) dans son ouvrage « Modèle d'I.E » (*Economica* - sept.2004).
[9] Se reporter à l'annexe A, section A.3 : Sources formelles et informelles
[10] Se reporter à l'annexe A, section A.3

la veille doivent permettre aux sociétés d'anticiper et d'avoir une meilleure vision sur leur environnement (Jakobiak, 2004 ; Coutenceau, Barbara et al, 2009 ; Coutenceau, Barbara, Chapuis-Thuault et al, 2014).

Dans la pratique, le pilotage assisté par l'I.E s'effectue en trois axes suivants : la veille, la protection et la décision d'agir et le pilotage de l'action. Concrètement,

> *« il s'agit de construire un plan de veille pour identifier les signaux faibles porteurs d'opportunité ou de menace en regard de sa stratégie afin d'agir en exploitant les informations recueillies et conforter ses positions sans oublier de protéger ce qui est important. La recherche de renseignements vient confirmer l'état final recherché, vérifier les orientations prises et le choix de la tactique.*
>
> *Les objectifs d'un pilotage par l'intelligence économique sont :*
> - *une meilleure capacité d'anticipation;*
> - *une vision objective de son environnement »* (Coutenceau, Barbara et al, 2009 : 9-10).

1.1.3.1 Comparatif veille et intelligence économique

Le QQOQCCP (« Qui ? Quoi ? Où ? Quand ? Comment ? Combien ? Pourquoi ? »), sigle résumant une méthode empirique de questionnement, *« un véritable couteau Suisse »* (Delengaigne et al, 2011), aide à toute démarche d'analyse informationnelle impliquant en effet une phase préalable de questionnement systématique et exhaustif. *«Elle est notamment utilisée dans les démarches qualité, pour s'accorder sur un problème à traiter ou pour définir le périmètre et les objectifs d'un projet à lancer»* (Delengaigne et al, 2011). Sa qualité conditionne celle de l'analyse proprement dite.

Les interrogations les plus importantes sont le comment ? et le pourquoi ?, suscitant des difficultés dans les réponses, mais le pourquoi du comment demeure souvent le plus difficile. Or, lors de la consultation des informations sur la toile, ces deux questions ne sont pas souvent traitées en profondeur.

Dans la plupart du temps, le début d'un travail de recherche scientifique est de cerner une situation, un problème voire un processus ou établir un état des lieux, cette méthode a pour objectif de recenser, prélever et mettre en évidence les données pertinentes dans un domaine donné. Sa simplicité, son caractère logique et systématique contribuent efficacement à la structuration des informations recueillies et à la restitution des résultats analytiques. Le tableau 1.1 ci-dessous fournit un comparatif entre la veille et l'intelligence économique.

Tableau 1.1 Comparatif veille et intelligence économique

QQOQCCP[11]	Veille[12]	Intelligence Economique
Qui	Tous les acteurs économiques (entreprises, institutions étatiques), (Lesca, 1997)	Dirigeants des acteurs économiques (Gouvernances privées et étatiques), (Lesca, 1997)
Quoi	Collecter des informations afin de maîtriser l'environnement socio-économique en vue de la conduite de la politique stratégique de l'établissement. Bien connaître son environnement, bien connaître ses partenaires de coopération et la politique scientifique et d'innovation de leur gouvernement, permet de faire des choix de coopération plus efficaces et plus fructueux.	Une démarche afin d'anticiper, de connaître les autres, de ne pas se laisser surprendre, de développer des stratégies à l'international et de définir la stratégie (alliance, fusion, concurrence et influence) dans son environnement politique, économique, social et technologique et obtenir de l'information stratégique afin de prendre de bonnes décisions stratégiques.
Où	Tous les lieux et/ou services au sein des acteurs économiques.	Lieux ou services au sein de la direction générale (le directoire) des acteurs économiques.
Quand	En continuum ou période dans son intégralité maximale.	A partir des années 1990 et en continu.
Comment	S'assurer de l'adhésion des acteurs, déterminer le périmètre de la veille en fonction des objectifs, mettre en place des plans de veille, recherche des sources à partir des brevets, des normes, des colloques, etc. et d'outils de veille.	Définir le besoin, les bonnes pratiques, établir une charte I.E et la déployer à tous les niveaux. Des indicateurs de suivis qualitatifs sur les forces et les faiblesses de l'entreprise, l'identification des risques évités ainsi que le sentiment d'être mieux informé, moins surpris, vous permettront d'évaluer l'impact de la stratégie d'I.E. Calculer le retour sur investissement *(Return on Investments, ROI)* est une étape importante qui doit impérativement être faite avant de se lancer dans la réalisation de projet.
Combien	Coût humain et coût matériel : - Par exemple, en France, un quart des PME interrogées déclarent disposer d'une personne ou d'une équipe dédiée à la veille[13]. Aussi, l'Etat encourage à la mutualisation des moyens. - Les outils de veille sont un investissement qui peut être coûteux.	Coût humain et coût matériel peuvent être onéreux, les solutions logicielles sont de plus difficiles à mettre en place. Le coût de prestations en I.E peut être comparé au coût d'une « assurance » (exemple de prestations de « savoir-faire » : conseil, transfert de méthodes, des formations, prestations d'expertise[14]).
Pourquoi	La veille stratégique permet d'avoir une bonne connaissance de son environnement socio-économique, et notamment : d'identifier les opportunités de développements technologiques, de suivre les évolutions des politiques publiques et des contextes économiques et internationaux.	Mondialisation, concurrence, multiplication des informations et vitesse de propagation, crise économique.

1.1.3.2 Veille stratégique : une activité du domaine de l'intelligence économique ?

L'intelligence économique va bien plus loin que la veille stratégique, mais il y a une interaction entre les deux concepts (Jakobiak, 2004).

A titre de comparaison, *« la veille stratégique serait les yeux et les oreilles d'un être humain »* (Martinet et Marty, 1995), alors que l'intelligence économique représente le cerveau. La première se limite à capter et à ressentir l'environnement dans lequel l'entreprise évolue et à faire parvenir au cerveau les informations ; la seconde, analyse les informations utiles qui permettront de faire réagir le corps de manière adéquate, c'est-à-dire l'entreprise toute entière.

L'intelligence économique se charge de toute la partie intellectuelle du traitement de l'information et des prises de décisions, mais plutôt que d'être fixée sur l'environnement externe, elle se concentre sur

[11] Démarche d'analyse impliquant au préalable un « questionnement systématique et exhaustif » afin de collecter les données nécessaires et suffisantes pour dresser l'état des lieux et rendre compte d'une situation, d'un problème, d'un processus.
[12] Source : Guide de l'intelligence économique pour la recherche, D2I.E, fiche 1 : veille stratégique.
[13] Etudes menées dans plusieurs régions (Bretagne, Lorraine, Sarthe) par les CCI de France, URL : http://www.intelligence-economique.gouv.fr/dossiers-thematiques/veille-strategique (consulté le 03/02/2015).
[14] http://www.acrie.fr/index.php/Coût.html (consulté le 03/02/2015)

l'entreprise elle-même, son champ d'attention dépend immédiatement du bon fonctionnement de la firme à laquelle elle appartient.

Dans tous les cas, l'information cernée par la veille est celle pouvant s'avérer opportune à la mise au point de la stratégie compétitive de l'entreprise au sein de son environnement. Quant à l'intelligence économique, celle-ci est beaucoup plus ambitieuse puisqu'elle cherche à influencer l'environnement, elle est une extension de la veille stratégique qui elle-même est un des fondements de l'intelligence économique. *« La différence essentielle entre veille stratégique et intelligence économique réside dans la modification de l'environnement que vise cette dernière »* (Larivet, 2001).

Mais les avis des experts divergent quelque peu. Pour Baumard, *« La veille n'est qu'un outil alors que l'intelligence est un système complet »* (Baumard, 1991).

En conclusion,

> *« la démarche d'intelligence économique est un préalable à la démarche de veille »* (Frion, 2002). *Mais d'autres chercheurs « pensent que l'intelligence économique est plus exigeante et plus complète que la veille, car elle met en œuvre infiniment plus de compétences et de réseaux que celle-ci. La veille est le système le plus technique, donc apparemment le plus facile à mettre en place. Cependant, rien n'indique qu'il faille développer une veille puis l'étendre à un système d'intelligence économique ou faire le contraire »* (Boizard, 2005).

L'I.E et la veille, une kyrielle d'amalgames, ne cessent de s'affronter et se noyer dans cette guerre sémantique. Pour chacune de ces écoles de pensées, l'une se veut supérieure, et l'autre ne se laisse difficilement régenter.

Les constats et comparaisons qui précèdent, tendent à éclaircir comme suite : la veille est constituée d'une multitude de processus de traitements pragmatiques. Quant à l'I.E, une pléiade de conceptions intellectuelles sert, entre autres, en amont à orchestrer la veille (ou cerner les QQOQCCP de la veille), et en aval, à apporter des réflexions décisives macro et microscopiques, une fois que le travail de veille commence à produire ses fruits.

Ce couple de concepts agissant en osmose, octroie des atouts absolus aux acteurs économiques face aux concurrences mondialisées. Les deux concepts synergiques font donc appel actuellement, par le biais de processus informatiques automatisés, d'une part aux connaissances économiques, et d'autre part aux techniques de traitement de l'information, sur la base de l'observation du contenu textuel disséminé dans des documents, rapports, journaux, blogs, Twitter, courriers électroniques, etc. (Fillias, 2005, 2007).

Toutes ces nouvelles avancées s'ouvrent à une communauté linguistique plus étendue, puis à l'ensemble des acteurs de l'économie mondiale, à la recherche d'informations relatives aux objectifs particuliers en matière de recherche et de développement. Elles permettent en outre de mettre en évidence la manière dont sont perçus et formulés, dans différentes langues et cultures, les principaux concepts de chaque domaine considéré. Elles offrent, entre autres, la possibilité de constituer des ressources traductologiques adaptées à chaque secteur de recherche.

1.2 Veille multilingue, une approche mondialisée

Dans un contexte social et économique qui s'achemine vers une mondialisation, il s'avère primordial pour communiquer et pour échanger de connaître et pratiquer plusieurs langues afin d'évoluer vers une veille multilingue et non plus de rester dans un milieu de veille monolingue, c'est-à-dire de se limiter à son propre environnement.

Grâce à Internet, les échanges sont devenus rapides et surtout les informations sont diffusées à travers tous les pays, ce qui a pour conséquences d'interpréter les informations dans de multiples langues.

> *« Le monde contemporain se caractérise par la conjonction de trois phénomènes majeurs : la mondialisation et l'avènement des nouvelles technologies de l'information et de la communication (NTICs), l'émergence d'un monde multipolaire en raison de la montée en puissance des BRICS, l'entrée dans un « monde post-américain »* (Zakaria, 2008).

> *« Ce triple processus fait de la diversité linguistique un enjeu central de la mondialisation et rend obsolète le modèle dominant du tout-anglais. Ce n'est pas sans poser un problème méthodologique fondamental, qui ne saurait être résolu que dans un cadre résolument pluridisciplinaire »* (Oustinoff, 2013).

L'anglais est resté longtemps la langue unique des échanges depuis l'époque où l'empire britannique s'est étendu un peu partout dans le monde en multipliant ses colonies et bien sûr en utilisant sa langue dans tous les domaines, qu'ils soient sociaux, administratifs ou économiques. Connaître l'anglais était suffisant pour effectuer une veille. La modernisation se matérialise par des progrès dans les domaines scientifiques et technologiques, tandis que la sphère sociale est confrontée à des phénomènes comme le plurilinguisme[15] ou encore le multilinguisme[16], la globalisation[17] et la mondialisation[18] où la concurrence ne cesse de renforcer la nécessité de la veille dans les entreprises (Levet, 2008).

Plurilinguisme ou multilinguisme

> *« Nous convenons dans ce qui suit de désigner par plurilinguisme l'usage de plusieurs langues par un même individu. Cette notion se distingue de celle de multilinguisme qui signifie la coexistence de plusieurs langues au sein d'un groupe social. Une société plurilingue est composée majoritairement d'individus capables de s'exprimer à divers niveaux de compétence en plusieurs langues, c'est-à-dire d'individus multilingues ou plurilingues, alors qu'une société multilingue peut être majoritairement formée d'individus monolingues ignorant la langue de l'autre. »*[19] (La Charte européenne du plurilinguisme, mise à jour : 22 juillet 2015)

[15] État d'un individu ou d'une communauté qui utilise concurremment plusieurs langues selon le type de communication; situation qui en résulte. http://www.cnrtl.fr/definition/plurilinguisme (consulté le 03/02/2015)
[16] État d'un individu ou d'une communauté linguistique qui utilise concurremment trois langues différentes ou davantage. http://www.cnrtl.fr/definition/multilinguisme (consulté le 03/02/2015)
[17] Fait de percevoir, de concevoir quelque chose ou quelqu'un comme un tout. http://www.cnrtl.fr/definition/globalisation (consulté le 03/02/2015)
[18] Action, fait de donner une dimension mondiale à quelque chose. http://www.cnrtl.fr/definition/mondialisation (consulté le 03/02/2015)
[19] Source : http://www.observatoireplurilinguisme.eu/index.php?option=com_content&view=article&id=9441:la-charte-europ%C3%A9enne-du-plurilinguisme-version-interm%C3%A9diaire&catid=177778327&Itemid=178380924&lang=fr (consulté le 03/02/2015)

Il est important de distinguer ces deux termes, le plurilinguisme se rapporte à un individu parlant plusieurs langues, alors que le multilinguisme désigne la coexistence de langues différentes dans une sphère de communication. L'Union européenne a été confrontée dès sa création au plurilinguisme du fait des particularismes qui la composent. Elle a encouragé tous ses citoyens à parler leur langue maternelle, mais aussi d'autres langues. C'est une des premières sources de réussite pour une Union durable.

1.2.1 Information et communication multilingues

1.2.1.1 Une communication multilingue

Dans ce contexte de sociétés mondialisées, le Conseil de l'Europe apporte une définition de la communication multilingue :

> *« [...] l'approche plurilingue met l'accent sur le fait qu'au fur et à mesure que l'expérience langagière d'un individu dans son contexte culturel s'étend de la langue familiale à celle du groupe social, puis à celle d'autres groupes[...], il ne classe pas ces langues et ces cultures dans des compartiments séparés, mais construit plutôt une compétence communicative à laquelle contribuent toute connaissance et toute expérience des langues et dans laquelle les langues sont en corrélation et interagissent »* (Conseil de l'Europe, 2005 : 11).

Le Conseil de l'Europe privilégie le terme plurilingue, c'est-à-dire, le fait de pouvoir s'exprimer dans plusieurs langues.

1.2.1.2 Une société d'informations

A notre ère, d'une part, les innovations scientifiques et technologiques ne cessent de transformer profondément les infrastructures sociales, d'autre part, les pratiques et les mœurs sont soumises régulièrement à des adaptations rudes.

Jadis, Isaac Newton trouve les lois du mouvement[20] et apporte une contribution décisive pour les sciences de la dynamique (en 1687), la machine à vapeur de James Watt voit le jour (brevet déposé en 1769), André-Marie Ampère (en 1822), Michael Faraday (en 1830) et James Maxwell (en 1860) établissent les postulats de base de l'électromagnétisme[21], Ludwig Boltzmann (en 1872)[22] théorise la thermodynamique, avec leurs conséquences révolutionnaires, puis l'électricité est arrivée, et la lampe électrique à incandescence de Thomas Edison bouleverse la nuit du temps. Au milieu du XIX^{e} siècle, la révolution industrielle et la démocratisation de l'instruction poussent la presse écrite (Pellaton et Delobbe, 2006) à devenir un vecteur d'information atteignant un large public.

Dans la société de l'information d'aujourd'hui, de la relativité d'Albert Einstein en passant par la physique quantique de Max Planck, avec leurs applications avancées dans l'énergie nucléaire et les nanotechnologies, l'informatique, le web, et les télécommunications sont devenus de nouveaux modes de partage de l'information et des connaissances. L'information du XXI^{e} siècle est une information des connaissances partagées. Aussi, *« la maîtrise de l'information est considérée comme un nouveau paradigme »* (Guidère, 2008).

[20] « *Philosophiae naturalis principia mathematica* » publié à Londres en 1687. La traduction française fut publiée à Paris en 1756, sous le titre « Principes mathématiques de philosophie naturelle rédigée » par Émilie du Châtelet.
[21] La découverte au XIX^{e} siècle par Oersted, Ampère et Faraday de l'existence d'effets magnétiques de l'électricité a conduit progressivement à envisager que les forces « électrique » et « magnétique » puissent être unifiées, et Maxwell propose en 1860 une théorie générale de l'électromagnétisme classique, qui pose les fondements de la théorie moderne.
[22] Le théorème H — prononcer théorème êta — est un théorème démontré par Boltzmann en 1872 dans le cadre de la théorie cinétique des gaz.

1.2.2 Enjeux de langues et traductions

Le multilinguisme et/ou le plurilinguisme, c'est promouvoir, protéger et préserver toutes les langues et cultures du monde, alors que la traduction serait un élément primordial pour le multilinguisme dans la construction de liens entre les différents peuples. La bonne traduction garantit la qualité de la transmission de tous les savoirs, quelles que soient leurs formes, politiques, économiques, sociales, culturelles, religieuses.

« La langue n'est pas seulement une donnée essentielle de la culture, c'est aussi un moyen d'accès aux manifestations de la culture. [...]. Les différentes cultures (nationales, régionales, sociales), auxquelles quelqu'un a accédé ne coexistent pas simplement côte à côte dans sa compétence culturelle. Elles se comparent, s'opposent et interagissent activement pour produire une compétence pluriculturelle enrichie et intégrée dont la compétence plurilingue est l'une des composantes, elle-même interagissant avec d'autres composantes » (Conseil de l'Europe, 2005 : 12).

L'Histoire, les cultures, les sciences, en résumé, le patrimoine de l'humanité se racontent, se communiquent, se diffusent et se traduisent depuis la nuit des temps. La traduction et le multilinguisme établissent les communications transdisciplinaires, parfois stratégiques, véhiculées par l'intelligence de l'Homme, tout en conservant les identités culturelles de chacun.

> *« Depuis quinze ou vingt ans, nous disposons d'une dizaine d'ouvrages en français, en russe, en tchèque, en anglais voire en portugais, qui offrent une bonne base de réflexion systématique – on peut même dire scientifique – sur cette opération intellectuelle et linguistique, laquelle paraît banale aux profanes, et mystérieuse au contraire aux savants : la traduction. Cette situation assez nouvelle nous change des périodes antérieures où la traduction n'avait donné lieu qu'à des notes, à des observations éparses, à des conseils empiriques, à des méditations de type artisanal.»* (Margot et Mounin, 1990 : 9).

1.2.3 La veille multilingue

Comme la définit Mathieu Guidère*, « la veille multilingue désigne l'activité de suivi informationnel effectuée parallèlement en deux ou plusieurs langues concernant un sujet spécifique ou un domaine particulier. Elle englobe plusieurs types de veilles spécifiques telles que la veille média, la veille juridique et réglementaire, la veille scientifique et technologique, la veille sanitaire et médicale, la veille économique et concurrentielle ou encore la veille géopolitique et sécuritaire »* (Guidère, 2008).

Si à ce jour, bon nombre d'acteurs déclarent participer à la veille, peu de métriques pragmatiques sont néanmoins mises en place. La restitution, la prévision et l'anticipation sont quasi-inexistantes en l'absence de véritables modèles prédictifs. L'un des objectifs principaux de notre travail est de créer des méthodes efficaces et facilement applicables, méthodes permettant de fouiller les textes, d'identifier les informations clés et de s'approprier un mode opératoire pour la veille multilingue.

1.2.4 Terminologie multilingue de la veille et de l'intelligence économique

1.2.4.1 En français

Les discussions supra attestent qu'en français l'Intelligence Compétitive serait plus clairvoyante que l'intelligence économique, dénomination sous l'influence de la langue anglaise. Tandis que la veille reste un terme aussi générique que ses fonctions de détections et de repérages d'informations multi-catégorielles.

Qu'en est-il de cette terminologie en anglais et en chinois ?

1.2.4.2 En anglais

Les anglo-saxons, depuis de nombreuses années, ont modifié leurs structures de décision face à la masse d'informations en développant, notamment une approche transversale de l'information autour de trois principes fédérateurs : le *business intelligence*, la *competitive intelligence* (Gilad et Gilad, 1998) et le *knowledge management* (Harbulot, 2001, 2004).

La veille se traduit de manière générale par information monitoring et la veille stratégique se veut *strategic monitoring* ou *monitoring* and *evaluation in strategic management*.

La notion d'intelligence économique varie selon les pays anglophones, elle évoque le *« renseignement »* (MacMurray, 2012 : 30), alors qu'en français, elle est perçue

> *«...comme un concept ambigu, interprété tantôt comme une méthode d'espionnage économique – c'est le côté « barbouzerie » d'officine – tantôt comme une méthode classique au service des seules entreprises : veille commerciale, veille juridique, veille concurrentielle...»* (Carayon, 2003, 2004).

Les anglo-saxons s'orientent vers une nouvelle dénomination plus claire et moins ambiguë, la *Competitive Intelligence*, notion regroupant

> *« l'ensemble des techniques de gestion des sources ouvertes pour s'assurer un avantage concurrentiel pérenne sur les concurrents et les futurs compétiteurs ... »* (Harbulot, 2000).

La notion de Business Intelligence était souvent associée à l'intelligence économique par le passé. Le BI (Business Intelligence)

> *« systématise le recours à la méthodologie du cycle du renseignement pour optimiser l'approche du client ... »* (Harbulot, 2004).

Avec cette nouvelle dénomination de la *competitive intelligence*, l'*Economic Intelligence* retrouve son sens initial sans trop d'ambigüité (Jenster et Solberg Soilen, 2009).

1.2.4.3 En chinois [23]

La veille en chinois se traduit généralement par le terme 监测/jiān cè, dont le 监/jiān veut dire : superviser, inspecter ou veiller, et le sens du caractère 测/cè est sonder, mesurer ou arpenter. Par exemple, la notion de la veille d'opinion en chinois serait le 舆情监测/yú qíng jiǎn cè, dont le sens littéralement des quatre caractères est *opinion, renseignement, veiller, sonder*. Cette traduction impose fortement une connotation « militaire ou militariste » du terme.

En mandarin de la Chine continentale, la notion d'intelligence, 智能/zhì néng/intelligence a été traduite de façon littérale comme intelligence au sens QI (Quotient Intellectuel, l'intelligence de l'Homme), puis au sens renseignement, 情报/qíng bào/renseignement (possédant souvent une connotation militaire), grâce à des connaissances plus profondes du terme. Toutefois, les deux traductions coexistent toujours, plus particulièrement, la traduction de *business intelligence*, au sens 商业智能/shāng yè zhì néng /intelligence de l'Homme pour les *business*, domine une bonne partie de sources textuelles pour les professionnels des sciences économiques et de la gestion sur le Web[24].

1.3 Méthode et stratégies de veille multilingue

Inspiré des théories de la veille et de l'intelligence économique, nous proposons une méthode de veille multilingue dont le processus est exposé dans la figure 1.1 ci-dessous.

Par souci du coût et de la rentabilité du travail, la veille multilingue se veut toujours thématique, des amorçages par l'intelligence économique s'avèrent donc indispensables pour tout type de veille multilingue.

[23] Pour des facilités de lecture, chaque notion en chinois est présentée de la manière suivante :
caractères chinois /pinyin/ traduction en français. Par exemple : 汉字/hàn zì/sinogramme ou idéogramme.
[24] http://wiki.mbalib.com/wiki/%E5%95%86%E4%B8%9A%E6%99%BA%E8%83%BD
(consulté le 13/01/2015)

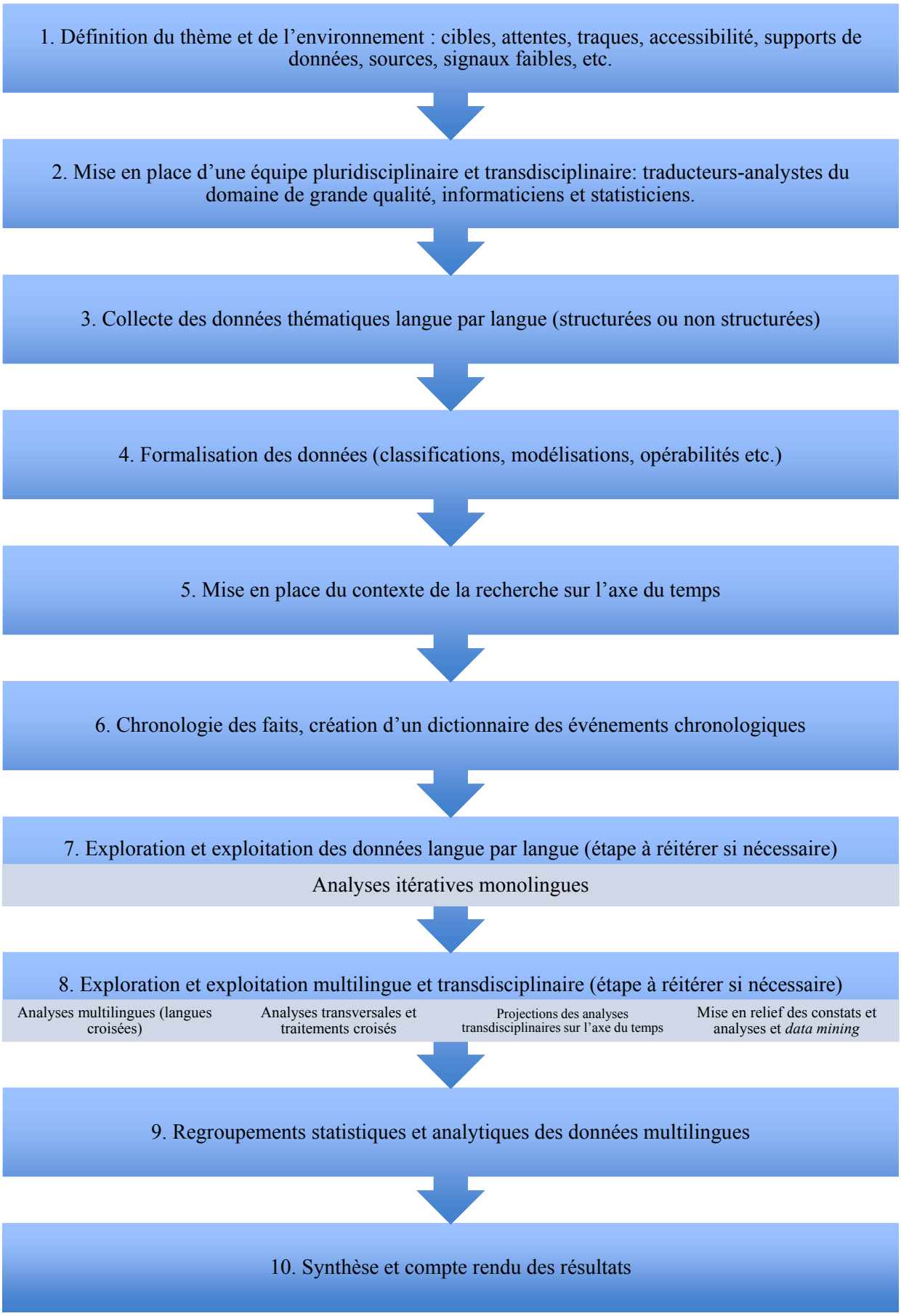


Figure 1.1 Schéma du processus de la veille multilingue

1.4 Statistique textuelle, un outil puissant et efficace

D'une part, avec les évolutions des NTIC, ainsi que les nouveautés fonctionnelles des outils et plates-formes de veille,

> *« le marché de l'information d'aujourd'hui n'est plus celui d'hier ». Par exemple, le « web sémantique offre de nouvelles opportunités pour nombre d'acteurs »,* et *« l'émergence de l'open data qui ouvre de nouvelles perspectives dans le domaine de l'information décisionnelle en apportant au veilleur des champs de données jusqu'alors inexplorés et dont le croisement doit produire des services d'information innovants, comme les interfaces cartographiques de data visualisation »* (Bour et al, 2012).

D'autre part, avec l'arrivée des mégadonnées (Bollier, 2010), *« l'enjeu organisationnel devient dès lors l'optimisation de la circulation de cette information distribuée »* (Bour et al, 2012), la redéfinition imminente des domaines tels que la capture, le stockage, la recherche, le partage, l'analyse et la visualisation des données serait la nouvelle priorité de la prochaine décennie et l'un des plus grands défis informatiques en matière de recherche et développement.

Aujourd'hui, les données disponibles sont quantitativement volumineuses et qualitativement disparates. Afin de détecter l'information pertinente et de prendre des décisions adéquates, des outils automatisés doivent être mis en place pour extraire, collecter, trier, classer et analyser ces masses.

Dans le cadre de ce travail, notre collecte d'informations s'orientera plus particulièrement vers deux thèmes : énergies et environnement. Cette recherche s'effectuera principalement à partir d'articles de journaux disponibles sur le Web, articles provenant de trois pays, France, États-Unis et Chine.

La veille et l'intelligence économique sont soumises dès lors à un volume de données colossales et non structurées produites entre autres par les médias informationnels. Par conséquent, la mise en circulation de nouveaux outils dédiés aux traitements de l'information devient incontournable.

La fouille et la veille d'information par la statistique textuelle, la textométrie[25] introduit pour un veilleur de nouvelles dimensions attestant l'authenticité des traces textuelles et contextuelles (sens premier du terme), telles que la cartographie, la cooccurrence et la poly-cooccurrence, en matière de comparaisons analytiques.

1.4.1 Veille textométrique : une école du TAL statistique

Commençons par une définition du TAL : *« Le traitement automatique des langues est constitué des méthodes et des programmes qui prennent pour données des productions langagières, quand ces méthodes et ces programmes tiennent compte des spécificités des langues humaines »* (Cori, 2004).

Les utilisations statistiques du TAL (Muller, 1967, 1968, 1973, 1975, 1977) reposent sur des méthodes stochastiques, probabilistes ou simplement statistiques pour résoudre les problèmes linguistiques et paralinguistiques impliqués. Ces méthodes pratiquées comportent souvent l'utilisation de corpus et d'outils de formalisation comme le traitement de la parole, la traduction automatique, la compréhension automatique des textes, la génération automatique de textes, la gestion électronique de l'information et des documents existants (GEIDE), la linguistique de corpus (Habert, Nazarenko et Salem, 1997 ; Teubert, 2009) robuste, la correction automatique orthographique, etc. Ces méthodes statistiques comportent toutes les approches quantitatives (McEnery et Wilson, 2004) du traitement

[25] Rappelons que selon le domaine d'utilisation, la statistique textuelle se nomme également : statistique linguistique, statistique lexicale, lexicométrie et textométrie (MacMurray, 2012 : 60).

linguistique automatisé, incluant la modélisation (Wang, Hodges et Tang, 2003), la théorie de l'information, et l'algèbre linéaire. L'apprentissage automatique et la fouille de données ou *data mining* (Naïm et Bazsalicza, 2001) sont deux principaux domaines d'applications du TAL statistique, impliquant l'apprentissage à partir des données venant de l'intelligence artificielle.

« Le TAL au sens large (traitement automatique des langues, ingéniérie des langues, linguistique computationnelle) voit sinon converger du moins se rencontrer et se mêler des communautés naguère clairement séparées : intelligence artificielle (IA), traitement de la parole (speech processing, speech synthesis), recherche d'information (information retrieval), extraction d'information (information extraction) » (Habert et Zweigenbaum, 2002).

1.4.2 Analyse textuelle et analyse du discours

L'expression *analyse du discours* est née en France, et *« c'est en avril 1968 au premier colloque de lexicologie politique de Saint-Cloud que le linguiste français Jean Dubois ouvrait la perspective de l'analyse du discours à partir d'un article du linguiste américain Z.Harris »*. L'expression peut donner lieu à de multiples approches ainsi que le montre la citation suivante :

> *Dans ma propre perspective (Maingueneau, 1995), l'analyse du discours est seulement une des disciplines des études de discours : rhétorique, sociolinguistique, psychologie discursive, analyse des conversations, etc. Chacune de ces disciplines est gouvernée par un intérêt spécifique. L'intérêt de l'analyse du discours est d'appréhender le discours comme articulation de textes et de lieux sociaux. Son objet n'est ni l'organisation textuelle ni la situation de communication, mais ce qui les noue à travers un certain dispositif d'énonciation. La notion de « lieu social » ne doit pas être prise dans un sens trop immédiat : ce lieu peut être une position dans un champ symbolique (politique, religieux…). En conséquence, l'analyse du discours accorde un rôle clé aux genres de discours, qui ne sont pas considérés comme des types de textes, dans une perspective taxinomique, mais comme des dispositifs de communication, de nature à la fois sociale et linguistique.* (Maingueneau, 2012)

Ces analyses utilisent des approches qualitatives et quantitatives. Elles cherchent le plus souvent à qualifier les éléments du texte à l'aide de catégories et à les quantifier en analysant la répartition statistique des éléments du texte. Une telle analyse a par exemple été très inspirée par les travaux de Jean-Paul Benzécri[26] et fut utilisée dès les années 1960 sur les textes littéraires, ainsi l'exemple d'une étude du poème « les Chats » de Charles Baudelaire par Jakobson et Lévi-Strauss (Jakobson et Lévi-Strauss, 1962).

Soulignons que l'analyse du discours renvoie, de manière générale, à la pragmatique selon la définition du TLFi : *Qui étudie le langage du point de vue de la relation entre les signes et leurs usagers*[27].

Par sa nature interdisciplinaire, la statistique textuelle (Salem, 1993) est étroitement liée au TAL (Cori et Léon, 2002 : 22-43). La veille textométrique (MacMurray, 2012 : 123) est un ensemble de méthodes utilisant la statistique textuelle et la linguistique pour effectuer la fouille de textes dans un objectif de veille ou d'I.E.

[26] Jean-Paul Benzécri : fondateur de l'école française d'analyse de données dans les années 1960-1990, a développé des outils statistiques, notamment l'analyse factorielle des correspondances qui permet de traiter de grandes masses de données afin de visualiser et hiérarchiser l'information.
[27] http://www.cnrtl.fr/definition/pragmatique (consulté le 23/01/2016)

L'analyse des données textuelles se décompose en deux grandes familles : la catégorisation et la fouille de textes. La catégorisation, comme son nom l'indique permet de classer un document dans une ou plusieurs thématique(s), voire analyser les tendances ou comportements des utilisateurs (Feldman et Dagan, 1995 : 115). Elle est basée sur le traitement automatique du langage naturel. La classification des documents est pilotée par une taxinomie, définissant les thèmes de classification ainsi que les mots, groupes de mots ou règles linguistiques permettant d'associer cette classification à ce document. A l'évidence, les documents traités doivent être numérisés, un document pouvant être un mail, un pdf, une page html, etc. Nous classons donc les documents sur un a priori, la taxinomie[28], issu de l'expérience propre au métier ou des meilleures pratiques d'un secteur d'activités.

Le *text mining*, aussi appelé fouille de textes : ici la démarche est inverse, on part sans a priori et on va décrire des documents en les regroupant, par les mots ou les groupes de mots les plus discriminants. En fait, on applique les techniques de modélisation utilisées pour des valeurs quantitatives telles que la classification hiérarchique (Malrieu et Rastier, 2001). Une fois l'analyse réalisée, les termes identifiés peuvent enrichir et faire évoluer la taxinomie.[29]

Cette recherche concernera plus particulièrement :

- les méthodes de collectes de donnée automatisées, de sélection et d'intégration rationnelle dans des corpus d'analyse textométrique de documents fournis par les programmes créés par l'auteur,
- le développement de méthodes d'analyses mieux adaptées à l'exploration textométrique de corpus hétérogènes mais thématisés (comparaison de textes rédigés pour des supports informationnels différents, catégorisations multilingues, textométrie multilingue, procédures d'affichage et de mises en parallèle de textes multilingues, etc.).

L'objectif est de proposer des formes plus simples, plus transparentes et moins *dirigées* de la veille informationnelle. Au lieu de rechercher dans les textes les instances de *patrons* ou de notions définis a priori, nous privilégirons la collecte et la structuration des formes d'expression émergentes dont la fréquence tend à augmenter, au cours d'une période donnée, parmi les textes mis à disposition des médias.

Nous utiliserons de manière privilégiée des corpus de textes français, anglais et chinois, langues pour lesquelles nous possédons une compétence plus développée. De même, nous commencerons à appliquer nos méthodes à des corpus de veille rassemblés sur le réseau Internet à partir de réponses à de vastes questions qui touchent à la protection de l'environnement, secteur pour lequel la demande se manifeste plus fortement en ce début de siècle.

Les acteurs économiques, conscients des enjeux de la veille, se trouvent souvent démunis face à la complexité de l'approche, au volume des données et à leur hétérogénéité. Ce contexte impose la mise en œuvre de veille informatisée, ne pouvant qu'être acquise auprès des professionnels de la veille, experts en fouille informationnelle et en TAL.

Au-delà de ces applications aux domaines particuliers du nucléaire et de l'environnement, notre projet se propose de développer une méthodologie générique de la veille informationnelle, basée sur les documents disponibles sur la *toile* quels que soient la langue et le domaine dont ils relèvent.

[28] Science des lois et des principes de la classification des organismes vivants; par extension, science de la classification. http://www.cnrtl.fr/definition/taxonomie, (consulté le 15/03/2010).
[29] http://www.statistique-2013.fr/02-analyse-textuelle.shtml, (consulté le 15/01/2015).

Par ailleurs, et c'est un des points les plus importants pour notre travail, nous ferons en sorte que les prescriptions méthodologiques dont l'efficacité aura été prouvée par nos analyses puissent être rassemblées en un ensemble de procédures informatisées directement utilisables par une communauté élargie de chercheurs et d'analystes.

1.4.3 Textométrie multilingue

« *La textométrie ne se confond pas avec la linguistique de corpus. Toutes deux fondent leurs investigations sur un corpus numérique, dont la constitution est déterminante. Comme son nom l'indique, la linguistique de corpus poursuit un objectif de description et de modélisation de la langue. La textométrie, centrée sur le texte, a pu être mobilisée par diverses sciences humaines (histoire, littérature, sciences politiques etc.)* » (Pincemin, 2011)

« *Alors qu'elle se propose de dévoiler la structuration lexicale des textes, l'approche lexicométrique trahit d'emblée la réalité textuelle dans ce qu'elle a de fondamental : sa 'séquentialité' syntagmatique. En effet, au terme d'une segmentation – opération initiale qui vise à identifier les composantes lexicales du corpus étudié – la statistique textuelle produit un inventaire hiérarchisé qui classe les formes lexicales suivant leur fréquence d'apparition. Or, même si cette liste permet par des calculs contrastifs de mesurer la distribution du vocabulaire dans le discours et d'y repérer les ventilations irrégulières qui traduisent des préférences d'emploi, la décontextualisation des lexies crée une rupture que la statistique seule ne peut compenser.*» (Martinez, 2012)

Les travaux de l'analyse de discours s'effectuent généralement sur des séquences de textes, sur lesquels on tire des hypothèses puis on vérifie celles-ci sur des discours. La lexicométrie est appliquée strictement au lexique, la textométrie aux textes. Quant à la textométrie multilingue, elle consiste à mettre en pratique la textométrie sur des corpus textuels de deux et/ou plusieurs langues.

1.5 Méthode analytique pour la textométrie multilingue

Nous appelons cette méthode, évaluations et analyses *critériées* par domaines disciplinaires et par séries chronologiques.

La recherche des méthodes et technologies performantes du traitement du langage naturel et de la communication a été longtemps la préoccupation principale de la communauté scientifique, après des décennies de travaux sans véritable fil conducteur, ainsi sont intronisées la linguistique computationnelle et l'ingénierie linguistique. Leurs natures hautement interdisciplinaires s'inscrivant dans les sciences, voire dans le monde de la Connaissance au sens très large du terme, permettent de mieux cerner la communication digitale véhiculée en langage naturel et d'interpréter le comportement humain. La moindre effervescence pourrait, comme on dit, « apporter de l'eau au moulin » des chercheurs en sciences humaines et sociales.

La linguistique, une discipline traitant l'interférence et l'interdisciplinarité à travers la nature des lexiques, les propriétés de la phonologie, de la morphosyntaxe et de la sémantique, suggère les démarches et les outils conceptuels permettant d'analyser les phénomènes des mots et/ou des segments textuels en fonction de leurs contextes. Nous proposons une méthode analytique, telle que décrite dans les tableaux 1.2 et 1.3 ci-dessous. Ceux-ci présentent la mise en œuvre d'une méthode analytique *critériée* par deux axes, un axe linguistique et un axe statistique. Sur l'axe linguistique, quatre domaines seulement sont pris en compte. Nous sommes conscients que la phonétique et l'étymologie sont des domaines en relation étroite avec la linguistique, mais dans notre présent travail, nous les écartons pour des raisons pratiques. Cette méthode d'analyse par critères sera appliquée à nos deux corpus et présentée dans la section 4.4.

Tableau 1.2 Évaluations et analyses « *critériées* » par domaines disciplinaires et séries chronologiques sur l'axe linguistique

Analyses *critériées*		**Axe du temps et série chronologique (en semaine, mois ou année)**									
		1999-2005			**2006-2009**			**2010-2012**			**etc.**
Langues / sphères		FR	US	CN	FR	US	CN	FR	US	CN	etc.
Axe linguistique	Morphosyntaxique										
	Sémantique										
	Lexique et/ou Lexicographique										
	Pragmatique (deixis)										
Intérêts manifestés	**Peu intéressant, Intéressant, Très intéressant**										

Le tableau 1.3 ci-dessous, regroupe les principaux modules d'analyse de la textométrie sur l'axe statistique. Chaque module relève des outils et des connaissances statistiques spécifiques aux comptages des mots et des formes à travers la volumétrie de chaque corpus constitué.

Tableau 1.3 Évaluations et analyses « *critériées* » par domaines disciplinaires et séries chronologiques sur l'axe statistique

Analyses *critériées*		**Axe du temps et série chronologique (en semaine, mois ou année)**									
		1999-2005			**2006-2009**			**2010-2012**			**etc.**
Langues / sphères		FR	US	CN	FR	US	CN	FR	US	CN	etc.
Axe statistique	Ventilation										
	Spécificité										
	AFC										
	Carte des sections										
	Cooccurrence et/ou poly-Cooccurrence										
	Segments répétés										
	Inventaire distributionnel										
Intérêts manifestés	**Peu intéressant, Intéressant, Très intéressant**										

En effet, dans un processus de veille, le veilleur se retrouve face à des masses de données qui sont souvent désordonnées et de provenances diverses. Dans le but d'exploiter ces volumes de manière efficace, il est nécessaire de posséder et d'appliquer des méthodes prospectives afin d'ouvrir des brèches analytiques.

Dans la veille active, l'objet à rechercher existe, mais les relations entre l'objet et les événements sont à révéler. Ces deux tableaux fonctionnent comme deux grilles d'évaluation appliquées sur tous les mots-clés ou tous les événements calculés et analysés par ces deux axes. L'évaluation s'effectue, critère par critère, afin de déterminer les degrés d'importance de chacun des axes, sur une échelle de 0 à 4 pour l'axe linguistique, et sur une échelle de 0 à 7 pour l'axe statistique. Dans les deux tableaux 1.2 et 1.3, nous avons retenu 4 domaines appartenant à la linguistique et 7 domaines à la statistique. Si plus de la moitié des critères de chaque axe sont attestés, alors, l'information recherchée devient intéressante. Il s'agit d'une démarche empirique de retenir plus de la moitié des critères.

Dans le cas où la veille est passive, c'est-à-dire, une veille où l'objet recherché n'est pas vraiment déterminé, la fouille d'informations est à effectuer dans la masse des données avec les mêmes échelles d'évaluation, alors cette méthode permet d'ouvrir des brèches efficaces d'investigation.

Cette méthode dite, « évaluations et analyses *critériées* par domaines disciplinaires et par séries chronologiques », est ainsi élaborée à partir des flux d'information et des données classées langue par langue et triées chronologiquement dans le continuum espace-temps. Elle est basée sur la textométrie dont l'axe linguistique et/ou l'axe statistique servent de socle, puis modulée avec toutes les autres disciplines. Les résonances textuelles monolingues, bilingues et multilingues à travers les informations ou événements sont appelées « poly-résonances » textuelles.

A l'issue des analyses croisées, si la moitié des critères de chacun des axes est renseignée, alors la période est considérée comme intéressante et nous focalisons nos analyses sur cette dernière. L'absence d'une série chronologique (Salem, 1988, 1991 ; Delanoë, 2010) dans l'un des corpus à caractère de veille comparable, conduirait à une élimination directe de cette période.

Dans le cas de la veille au sens général du terme, l'axe statistique devient parfois plus déterminant que l'axe linguistique, ou autre axe. Par exemple, d'autres axes tels que l'axe économique, l'axe sociologique, l'axe politique, l'axe psychologique, pourraient être intégrés dans cette méthode à l'aide d'indicateurs ou indices dénombrables pour chacune de ces disciplines, par exemple, indice des prix à la consommation, indice de niveau d'instruction, indice de salaire de base par secteur d'activité et catégorie socioprofessionnelle.

1.6 Corpus, alignements, comparabilité

Du recueil de documents spécialisés au corpus de veille multilingue thématique

Dans cette section, nous décrivons les concepts de base relevant de la notion de corpus. Dans un premier temps, nous définissons les termes corpus, texte, corpus parallèle, texte parallèle, corpus bilingue, corpus multilingue, bi-texte, multi-texte, texte comparable, corpus comparable. Nous aborderons par la suite les caractéristiques et les problèmes de ces corpus. Dans un second temps, nous examinerons plus particulièrement la notion-clé de corpus de veille multilingue thématique.

1.6.1 Corpus *versus* textes

En sciences humaines et sociales, le corpus est un *« recueil réunissant ou se proposant de réunir, en vue de leur étude scientifique, la totalité des documents disponibles d'un genre donné, par exemple épigraphiques, littéraires, etc. Corpus des inscriptions grecques, latines; le corpus des métriciens et des musicologues, etc. »* (TLFi, 1960). En linguistique, c'est un *« ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique, etc.»* (TLFi, 1960). Alors qu'un texte est une *« suite de signes linguistiques constituant un écrit ou une œuvre »* (TLFi, 1960).

En effet, une précision se doit d'être apportée aux sens de ces deux notions dans les études qui suivent. Nous parlons souvent de corpus de textes, corpus de mots, corpus d'articles, etc., la notion de corpus étant le contenant et les textes, le contenu. Le corpus serait le quantificateur ou classificateur de la notion de texte(s) dans un domaine spécifique, nous parlons d'un corpus d'articles du Monde, de deux corpus de textes de Balzac, tout comme un récipient et son contenu, tels qu'une bouteille d'eau, un vaste océan d'eau salée sur Ganymède, la plus grosse lune de Jupiter, etc. Or, la notion de texte(s) peut, elle aussi, devenir le contenant incluant le contenu par sa désignation. Un corpus de textes pourrait être également considéré comme un texte dans le sens figuré. Par exemple, un texte, des écrits ou des textes de Molière serait à la fois le support, le quantificateur et son contenu textuel. Donc, ce serait délicat d'affirmer la supériorité du corpus par rapport aux textes. Indéniablement, la perception de ces deux notions varie selon leur contexte.

Le corpus, selon la langue dans laquelle les textes sont rédigés, peut être monolingue, bilingue ou multilingue (Bowker et Pearson, 2002 : 12), textes issus d'une ou plusieurs langues ou de langages spécialisés. Il peut se décliner soit en corpus parallèle, soit en corpus comparable.

Dans la littérature, le corpus de référence se définit comme suit :

> *« …nous distinguons les corpus dits de référence, c'est-à-dire conçus pour être représentatifs d'une langue en général, tels que le British National Corpus ou le Corpus of Contemporary American English et les corpus spécialisés qui sont représentatifs d'un genre ou d'un domaine spécialisé. Tous ces corpus se définissent comme un ensemble de textes sur support électronique, qui ont été assemblés selon des critères spécifiques en fonction d'un objectif précis. Nous ne parlons donc pas ici de corpus d'exemples, ni de textes assemblés au hasard des trouvailles, sans cadre défini. Cette définition d'un corpus représente la définition canonique adoptée par la linguistique de corpus»* (Kübler, 2011).

1.6.2 Corpus parallèles, textes parallèles

Un corpus parallèle est *« un corpus qui contient des textes source et leur traduction »* (McEnery et Xiao, 2007), ou encore, les corpus parallèles sont des *« ensembles de textes accompagnés de leurs traductions dans une ou plusieurs langues »* (Bowker et Pearson, 2002 : 92).

Nous parlons de corpus parallèle bilingue, s'il est constitué d'une langue source et d'une langue cible, par exemple le corpus Hansard[30].

Quant au corpus parallèle multilingue, celui-ci peut se composer d'une langue source avec plusieurs langues cibles, par exemple le corpus MeLLANGE[31] (Castagnoli et al. 2011) ou le corpus du site « Linguee [32]». Parfois, la langue de départ peut être multiple. Dans ce cas, leurs traductions ne sont pas toujours directes, par exemple, la Bible et les textes de l'Union Européenne (Goeuriot, 2009; Kübler, 2011). Il est à noter que l'Ancien Testament a été rédigé en hébreu et en grec. Un texte voté par le parlement de l'Union Européenne est écrit et/ou traduit dans toutes les langues des états membres.

L'une des premières pages de l'ouvrage *« Parallel text processing »* (Véronis, 2000) définit le terme anglais « *parallel text* » en faisant référence à la Pierre de Rosette[33], l'exemple emblématique du texte parallèle. Si nous admettons que la définition du corpus, mentionnée plus haut, est plus vaste que celle du texte, c'est-à-dire, le corpus peut comprendre plusieurs textes, alors, nous l'appelons corpus parallèle. Les deux volets, source *versus* cible, du corpus sont parallèles dans leur totalité, mais ne sont pas toujours symétriques au niveau des unités textuelles, mots, expressions phrases, paragraphes, séquences, chapitres, etc. (Lebart et Salem, 1994, Chapitre 2). Par exemple, le traducteur peut prendre deux paragraphes de la langue source puis les traduire en un seul paragraphe dans la langue cible. De ce fait, de sérieux problèmes de correspondance directe des unités, entre les deux volets, se manifestent dans le corpus parallèle. Nous l'appelons, dans ce cas, corpus parallèle non aligné.

[30] Corpus composé de textes parallèles en anglais et en français canadien, tirés de documents officiels du discours parlementaire canadien de 1970 à 1988, souvent considéré comme une référence de corpus parallèle.
[31] http://mellange.eila.jussieu.fr/, consulté le 25 avril 2015.
[32] http://www.linguee.com, consulté le 25 avril 2015. Il convient de noter que le site Linguee utilise vraisemblablement une collection de corpus triés. Il est très difficile de savoir quel est le texte source et le texte cible de ces corpus.
[33] Pierre trouvée par Champollion en Egypte en 1799, pierre gravée d'un même texte en 3 écritures différentes (égyptien en hiéroglyphes, égyptien démotique et alphabet grec) et 2 langues égyptien ancien et grec ancien.

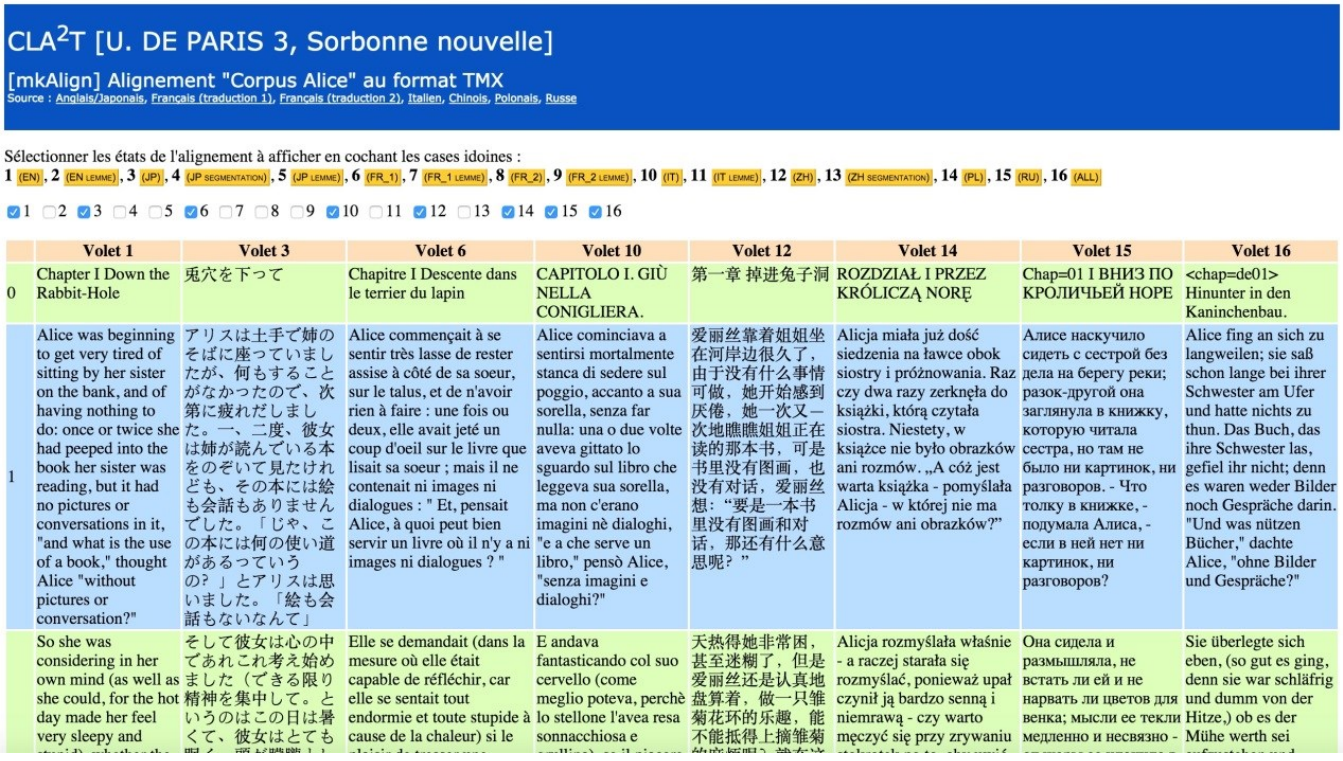
Pour résoudre ce problème et mieux exploiter le corpus, il faut procéder à une étape de traitement appelé alignement. Cette étape consiste à aligner ces unités textuelles entre les volets afin de récupérer leurs correspondances textuelles et sémantiques. Ainsi, nous obtenons le corpus ou texte parallèle aligné par paragraphe, par phrase, par expression ou par mot. Dans la pratique courante, nous nous limitons souvent à l'alignement au niveau du chapitre, du paragraphe ou de la phrase.

1.6.3 Les corpus alignés

Un corpus multilingue est dit aligné, si les unités ou éléments textuels ou para-textuels (paragraphes, phrases, termes, etc.) sont mis en correspondance, langue par langue, dans le corpus multilingue parallèle qui le compose.

D'une manière générale, un corpus parallèle est supposé déjà aligné. Mais cela nécessite un choix pour son unité d'alignement et un traitement technique long et coûteux. En effet, ce processus de l'alignement de corpus parallèles consiste plus précisément à effectuer *« la mise en correspondance des différents niveaux d'unités »* (Véronis, 2000) entre deux ou plusieurs volets.

Dans la littérature, le texte ou corpus parallèle aligné s'appelle parfois bi-texte (Harris, 1998) ou multi-texte (Véronis, 2000) quand il s'agit de plusieurs langues ou volets, par exemple, le corpus «*Corpus Alice, Alice au pays des mesures*» (figure 1.2).[34]


CLA²T [U. DE PARIS 3, Sorbonne nouvelle]

[mkAlign] Alignement "Corpus Alice" au format TMX

Source : Anglais/Japonais, Français (traduction 1), Français (traduction 2), Italien, Chinois, Polonais, Russe

Sélectionner les états de l'alignement à afficher en cochant les cases idoines :

1 (EN), 2 (EN LEMME), 3 (JP), 4 (JP SEGMENTATION), 5 (JP LEMME), 6 (FR_1), 7 (FR_1 LEMME), 8 (FR_2), 9 (FR_2 LEMME), 10 (IT), 11 (IT LEMME), 12 (ZH), 13 (ZH SEGMENTATION), 14 (PL), 15 (RU), 16 (ALL)

☑1 ☐2 ☑3 ☐4 ☐5 ☑6 ☐7 ☐8 ☐9 ☑10 ☐11 ☑12 ☐13 ☑14 ☑15 ☑16

	Volet 1	Volet 3	Volet 6	Volet 10	Volet 12	Volet 14	Volet 15	Volet 16
0	Chapter I Down the Rabbit-Hole	兎穴を下って	Chapitre I Descente dans le terrier du lapin	CAPITOLO I. GIÙ NELLA CONIGLIERA.	第一章 掉进兔子洞	ROZDZIAŁ I PRZEZ KRÓLICZĄ NORĘ	Chap=01 I ВНИЗ ПО КРОЛИЧЬЕЙ НОРЕ	<chap=de01> Hinunter in den Kaninchenbau.
1	Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, "and what is the use of a book," thought Alice "without pictures or conversation?"	アリスは土手で姉のそばに座っていましたが、何もすることがなかったので、次第に疲れだしました。一、二度、彼女は姉が読んでいる本をのぞいて見たけれども、その本には絵も会話もありませんでした。「じゃ、この本には何の使い道があるっていうの？」とアリスは思いました。「絵も会話もないなんて」	Alice commençait à se sentir très lasse de rester assise à côté de sa soeur, sur le talus, et de n'avoir rien à faire : une fois ou deux, elle avait jeté un coup d'oeil sur le livre que lisait sa soeur ; mais il ne contenait ni images ni dialogues : " Et, pensait Alice, à quoi peut bien servir un livre où il n'y a ni images ni dialogues ? "	Alice cominciava a sentirsi mortalmente stanca di sedere sul poggio, accanto a sua sorella, senza far nulla: una o due volte aveva gittato lo sguardo sul libro che leggeva sua sorella, ma non c'erano imagini nè dialoghi, "e a che serve un libro," pensò Alice, "senza imagini e dialoghi?"	爱丽丝靠着姐姐坐在河岸边很久了，由于没有什么事情可做，她开始感到厌倦，她一次又一次地瞧瞧姐姐正在读的那本书，可是书里没有图画，也没有对话，爱丽丝想："要是一本书里没有图画和对话，那还有什么意思呢？"	Alicja miała już dość siedzenia na ławce obok siostry i próżnowania. Raz czy dwa razy zerknęła do książki, którą czytała siostra. Niestety, w książce nie było obrazków ani rozmów. „A cóż jest warta książka - pomyślała Alicja - w której nie ma rozmów ani obrazków?"	Алисе наскучило сидеть с сестрой без дела на берегу реки; разок-другой она заглянула в книжку, которую читала сестра, но там не было ни картинок, ни разговоров. - Что толку в книжке, - подумала Алиса, - если в ней нет ни картинок, ни разговоров?	Alice fing an sich zu langweilen; sie saß schon lange bei ihrer Schwester am Ufer und hatte nichts zu thun. Das Buch, das ihre Schwester las, gefiel ihr nicht; denn es waren weder Bilder noch Gespräche darin. "Und was nützen Bücher," dachte Alice, "ohne Bilder und Gespräche?"
	So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and	そして彼女は心の中であれこれ考え始めました（できる限り精神を集中して。というのはこの日は暑くて、彼女はとても	Elle se demandait (dans la mesure où elle était capable de réfléchir, car elle se sentait tout endormie et toute stupide à cause de la chaleur) si le	E andava fantasticando col suo cervello (come meglio poteva, perchè lo stellone l'avea resa sonnacchiosa e	天热得她非常困，甚至迷糊了，但是爱丽丝还是认真地盘算着，做一只雏菊花环的乐趣，能不能抵得上摘雏菊	Alicja rozmyślała właśnie - a raczej starała się rozmyślać, ponieważ upał czynił ją bardzo senną i niemrawą - czy warto męczyć się przy zrywaniu	Она сидела и размышляла, не встать ли ей и не нарвать ли цветов для венка; мысли ее текли медленно и несвязно -	Sie überlegte sich eben, (so gut es ging, denn sie war schläfrig und dumm von der Hitze,) ob es der Mühe werth sei

Figure 1.2 Corpus parallèle multilingue, «*Corpus Alice, Alice au pays des mesures*»

[34] http://www.tal.univ-paris3.fr/mkAlign/corpus/Alice-ENG-JP-FR-IT-CH-PL-RU-ALL/ALIGNEMENT-ALICE.html, (consulté le 24 avril 2015).

1.6.4 Corpus comparables, textes comparables,

Le terme ***comparable***, est utilisé pour indiquer que les textes partagent certaines caractéristiques ou certains traits (Bowker et Pearson, 2002).

Les textes comparables sont « *des textes de genre, domaine, époque, etc. similaires* » (Véronis, 2000). Quant au corpus comparable, celui-ci : *«… se définit comme incluant des textes originaux dans différentes langues et répondant aux mêmes critères de genre, de temporalité, de registre, etc. Ici, aucun texte n'est la traduction de l'autre et on parle aussi de textes natifs. Un corpus comparable peut aussi être multilingue. Nous incluons aussi dans cette définition les corpus comprenant des textes rédigés dans une langue originale et des textes qui ont été traduits dans cette langue… »* (Bowker et Pearson, 2002; Kübler, 2011). Un corpus comparable réunit « *des documents textuels dans des langues différentes qui ne sont pas des traductions les uns des autres* » (Bowker et Pearson, 2002 : 93).

Les corpus comparables sont composés de documents écrits en deux ou plusieurs langues ayant une composition ou une structure similaire (ou quasi-similaire). Dans son article en anglais, Teubert (Teubert, 1996) parle de « *composition* » pour désigner ces structures. Les textes du corpus sont donc sélectionnés selon des critères de domaine, de période, etc. ou encore selon des critères purement linguistiques. Cette démarche assure un minimum de caractéristiques communes aux textes rassemblés (Firth, 1957). Cependant, Teubert ne nous précise pas ce qu'est une composition similaire, ni les critères qui permettent de la constituer. C'est pourquoi nous préférerons cette définition, plus générale : *« Les corpus comparables sont composés de documents en plusieurs langues, qui, sans être des traductions, ont toutefois des caractéristiques en commun »* (Bowker et Pearson, 2002 : 93).

Les critères généralement utilisés sont le thème abordé, le média, la période de rédaction, mais il y en existe d'autres. Ils sont parfois regroupés en deux catégories : les critères qualitatifs (critères utilisés en stylistique tels que le genre, l'auteur, la période, le média, etc.) et les critères quantitatifs **(**critères basés sur les mesures de fréquences de certains traits linguistiques comme par exemple la fréquence de certains termes) (Déjean et Gaussier, 2002).

Le degré de comparabilité d'un corpus varie selon l'ensemble des critères de comparabilité sélectionnés (Morin et Daille, 2006). Pour le moment, il semblerait qu'il n'y ait pas de corpus comparable de référence, puisque tout corpus comparable est constitué pour être comparé dans la mesure où seulement certaines comparabilités sont respectées. Cependant, nous pouvons utiliser le projet NERC (Network of European textual Reference Corpora) (Teubert, 1996), comme base pour constituer des corpus comparables (Sharoff, Rapp, Zweigenbaum et al., 2013).

Comme mentionné plus haut, nous pouvons ainsi appliquer, par assimilation, aux textes comparables, considérés comme des sous-ensembles de corpus comparables, le même principe que pour les textes parallèles et les corpus parallèles. A l'heure actuelle, aucun corpus comparable de référence n'a été élu par la communauté scientifique. Seules deux catégories principales de corpus comparables se distinguent :

> *« Les corpus comparables généralistes : composés généralement d'articles de journaux. Les documents sont souvent extraits de journaux nationaux et portent sur une même période, voire une même thématique. Les corpus comparables spécialisés : composés de documents émanant d'un domaine spécialisé, souvent scientifique, faisant appel à un langage spécialisé »* (Koehn, 2004 ; Goeuriot, 2009).

1.6.5 Caractéristiques et problèmes liés aux traitements des corpus parallèles et comparables

Dans le domaine public, en dépit d'un grand nombre de traductions disponibles, il existe relativement peu de textes parallèles et comparables compilés en corpus choisis parmi les langues européennes et le chinois. Toutefois, des quantités croissantes de textes bilingues et multilingues ont été mises à disposition notamment par les travaux recensés dans *Workshop on Statistical Machine Translation* (WMT16[35]). Pour utiliser les textes alignés, il faut tenir compte du fait qu'ils comportent chacun des particularités spécifiques. Les textes sont rédigés différemment selon le type de sujet traité. Par exemple, les articles juridiques utilisent un vocabulaire très précis à la différence des textes relevés dans les réseaux sociaux (par exemple forums, blogs, Twitter), qui sont beaucoup plus hétérogènes. Dans le cadre de nos travaux, toutes ces caractéristiques sont à prendre en compte pour l'utilisation des données, car elles peuvent avoir une incidence sur la qualité des analyses. Il est également important de pouvoir disposer d'un choix de corpus adéquats. Lors de l'évaluation des explorations, il faudra prendre en compte le registre du langage des corpus utilisés, afin de pouvoir déterminer leurs performances et leurs défauts. Les textes comparables présentent des avantages notables par rapport aux textes de traduction en parallèle. D'une part, ils n'ont pas les problèmes de sens induits par d'éventuelles erreurs de traduction et sont vierges de toute influence d'autres textes, d'autre part ils sont disponibles en très grand nombre dans le domaine public (Nakamura-Delloye, 2007).

Toutefois, il reste très difficile de construire des textes parallèles à partir de textes comparables alignés. Des études sur l'alignement ou l'extraction de mots correspondants à partir de textes comparables ont déjà été réalisées, jusqu'à des tentatives, aux résultats encore peu concluants, d'alignements de phrases entières (Munteanu et Marcu, 2002). Cependant, cela s'améliore dans les travaux plus récents (Rapp, Zweigenbaum et Sharoff, 2016 ; Rapp, Sharoff et Zweigenbaum, 2016).

1.6.6 Les corpus multilingues thématiques

Les corpus multilingues sont en général utilisés pour fournir des logiciels de traduction automatique, pour générer des exercices ou encore des supports d'étude dans le cadre de l'apprentissage des langues et pour alimenter des ressources de la fouille d'informations (Zanettin, 1998) multilingues. Dans leurs formes, les corpus de veille multilingues peuvent ressembler aux corpus comparables et/ou parallèles, toutefois, leurs objectifs sont différents. Dans cette présente recherche, nous aborderons des méthodes de fouille textuelle multilingue explorant précisément les corpus multilingues thématiques afin d'extraire des « renseignements » intéressants pour une intelligence ou une veille plus compétitive.

Disponibilité des corpus multilingues

Les NTIC permettent désormais d'avoir accès à un éventail très large de documents multilingues en ligne, à titre indicatif, « *en février 2015, Le Monde diplomatique comptait 35 éditions internationales en 19 langues : 30 imprimées et 5 électroniques*»[36].

Cependant, *« ...la constitution d'un corpus parallèle aligné est longue et demande de la part de l'utilisateur des compétences techniques que souvent le traducteur n'a pas ; d'autre part, les ressources disponibles permettant de créer des corpus parallèles sont rares ou protégées par des droits d'auteurs »* (Kübler, 2011).

[35] http://www.statmt.org/wmt16/index.html, (consulté le 01/07/2016).
[36] http://www.monde-diplomatique.fr/int/, (consulté le 27/04/2015).

1.6.7 Terminologie et spécificités en chinois

Nous soumettons une traduction en chinois des termes (Bourigault et Slodzian, 1999) utilisés dont la plupart sont déjà reconnues par la communauté scientifique :

- comparable(s) : 可比 kě bǐ;
- corpus : 语料库/yǔ liào kù; 文集/wén jí;
- corpus monolingue(s) : 单语语料库/dān yǔ yǔ liào kù ou 单语文集/dān yǔ wén jí;
- corpus multilingue(s) :
  多语语料库/duō yǔ yǔ liào kù ou 多语文集/duō yǔ wén jí;
- parallèle(s) : 平行/píng xíng ou 对等/duì děng;
- texte(s) : 文本/wén bén;
- texte(s) comparable(s) : 可比文本/kě bǐ wén bén;
- texte(s) parallèle(s) :
  平行文本/píng xíng wén bén ou 对等文本/duì děng wén bén.

Concernant le système linguistique du chinois et les traitements automatiques, il est à noter que par la structure des phrases, les concepts de caractères chinois, les morphèmes et lexèmes chinois sont très différents de ceux des langues européennes.

Les caractéristiques linguistiques et textuelles chinoises sont susceptibles de poser des défis particuliers lors des traitements automatiques :

- les variations régionales et/ou stylistiques qui peuvent exister parmi les caractères et leur encodage (Chine continentale, Hongkong, Taiwan, Singapour),
- les conventions textuelles spéciales de l'imprimerie et de la ponctuation,
- les nombreuses ambiguïtés de la langue,
- l'absence de marquages grammaticaux clairement définis pour de nombreux mots inconnus, comme les noms, les abréviations et les translittérations (Wong, Li et al, 2010). Par exemple : pour les toponymes, la ville de Lucerne en Suisse, se nomme en chinois à la fois, 琉森/liú sēng ou 卢塞恩/lú sài en. Ces deux formes différentes constituent une variante libre en linguistique.

Le système d'écriture de la langue chinoise fonctionne en milliers ou parfois en dizaines de milliers de caractères, appelés également sinogrammes ou Hanzi, dans lequel chaque caractère correspond pour la plupart des cas au moins à un morphème et à une syllabe phonologiquement. La notion d'espace entre des unités lexicales dans une phrase n'existe pas dans l'écriture chinoise ; toutefois, la ponctuation est similaire au système linguistique occidental. Le texte se rédige sans la séparation de mots par les espaces blancs et l'appréhension du texte se fait par le découpage personnel du locuteur. Le niveau de la connaissance du lexique de ce dernier déterminera la finesse du résultat de cette segmentation de la chaîne textuelle en unités sémantiques et/ou lexicales distinctes. Les fonctions grammaticales d'un mot sont multiples et varient selon le contexte. Ce phénomène de changement de fonctions grammaticales est plus marqué en chinois qu'en français et en anglais.

Voici un exemple reflétant la complexité de la segmentation de la langue chinoise :

- 小心肝/xiǎo xīn gān

Littéralement les trois caractères signifient : petit, cœur, foie. Deux segmentations sémantiques sont possibles :

- 小 | 心肝/xiǎo | xīngān : petit(e) chéri(e)
- 小心 | 肝/xiǎo xīn | gān : attention au foie

Comme nous pouvons le constater, selon le découpage du mot, le sens est complétement différent. Par conséquent, la segmentation en mots du texte en chinois joue un rôle déterminant dans tout traitement automatique du chinois et de l'information.

Les mots inconnus posent des problèmes importants lors de la segmentation en mots des textes chinois. Les entités nommées telles que les noms propres qui se réfèrent à des personnes, des lieux, des noms d'organisation constituent des sources importantes de mots inconnus. Les méthodes de reconnaissance des noms utilisent des indices tels que la structure commune du lexique dans ces diverses parties, le repérage de caractères utilisés régulièrement dans les noms, la prise en compte des données contextuelles, etc. Ces indices sont obtenus par extraction manuelle ou automatique à partir de corpus de textes établis par des institutions scientifiques. Les résultats de la segmentation et de la reconnaissance des mots inconnus sont en général évalués selon cinq critères : rappel, précision, F-mesure, rappel des mots inconnus et rappel des mots connus (Wong, Li et al, 2010). Il faut savoir qu'il existe également plusieurs « standards » de segmentation.

Cette complexité de la segmentation du texte en chinois souligne à l'évidence des difficultés pour l'alignement dans les corpus parallèles notamment où la langue chinoise est présente. Des traitements supplémentaires sont exigés pour la segmentation du texte chinois avant ou après l'alignement des volets du corpus multilingue.

Par ailleurs, ces difficultés ne sont pas les mêmes selon la direction (source-cible versus cible-source) de traduction réalisée, car la langue chinoise omet ou ajoute souvent des éléments de phrase, contrairement aux langues anglaise et française. D'une manière générale dans la langue chinoise, on a tendance à répéter des éléments afin d'éviter les confusions possibles (*cf.* section 7.2.1).

Tableau 1.4 Difficultés et spécificités de l'alignement de la langue chinoise

	En français ou en anglais	**En chinois et traduction**
Omission d'élément (phrase averbale)	Exemple 1 : Il fait beau.	天气很好。/tiānqì hěn hǎo. /Temps très beau.
	Exemple 2 : Il fait froid.	天很冷。/tiān hěn lěng. /Ciel très froid.
Explication	Sujet + verbe + COD/adjectif/attribut	Sujet + COD/Adjectif
Ajout d'éléments	Exemple 3 : Parce qu'il y a le soleil, les rayons du soleil éclairent la terre.	因为有了太阳，所以大地得到了阳光。Yīnwèi yǒu le tàiyàng, suǒyǐ dàdì dédào le yángguāng. / Parce qu'il y a le soleil, donc la terre peut bénéficier/profiter des rayons de soleil.
	Exemple 4: *"He wasn't far out. (The most recent calculation, based on enormous computer programmes at a number of world centres, including the Hadley Centre for Climate Prediction and Research yields global temperature increases of 1.5-6C for a doubling of carbon-dioxide levels.)"*	而他的计算结果和事实相差并不远.(最近在一些世界性的研究中心"包括哈德利气候预报研究中心"用庞大的计算机程序进行了计算.结果发现在二氧化碳水平增加了一倍之后.全球各地的温度升高了1.5－6摄氏度.) / Ér tā de jìsuàn jiéguǒ hé shìshí xiāngchà bìng bù yuǎn. (Zuìjìn zài yīxiē shìjiè xìng de yánjiū zhōngxīn" bāokuò hā dé lì qìhòu yùbào yánjiū zhōngxīn" yòng pángdà de jìsuànjī chéngxù jìnxíngle jìsuàn.) / Cependant, ses résultats de calcul ne sont pas très loin de la réalité.
Explication	La coprésence de connecteurs syntaxiques (cause et conséquence), 因为/yīnwèi/parce que … 所以/suǒyǐ/donc, « parce que…donc » est obligatoire, ce qui n'est pas le cas en français. Le connecteur « cependant » est absent dans l'exemple anglais. De plus, des éléments ont été ajoutés en chinois.	

Selon les constats du tableau 1.4 ci-dessus, il est plus facile d'analyser des textes chinois alignés à partir d'un original français ou anglais plutôt que l'inverse. Dans les exemples 1 et 2, « Il fait beau. » et « Il fait froid. », il est à noter qu'il s'agit d'une structure attributive dans une forme impersonnelle (le sujet il). En chinois, les verbes sont absents et le sujet impersonnel est remplacé par des noms communs comme « temps » et « ciel ». Quant aux exemples 3 et 4, ce sont des connecteurs « donc et cependant » et des éléments « résultats et réalité » qui sont ajoutés en chinois. L'ajout ou la répétition d'éléments dans les textes chinois par rapport aux versions françaises ou anglaises complexifient parfois la vérification de l'alignement.

Les expressions de la cause (Bertin, 1997, 2001, 2002, 2003) ou les connecteurs, mots ou expressions participant à l'établissement de la trame du discours, peuvent parfois modifier considérablement la structure des phrases et/ou des paragraphes lors de la traduction. La modélisation linguistique de ces connecteurs peut constituer une aide précieuse au traitement automatique des langues.

Les spécificités de chacune de ces trois langues seront abordées de manière détaillée dans le chapitre 3.

A ce point de la présentation du cadre de notre recherche, il est également intéressant d'aborder la notion de comparabilité de corpus.

1.6.8 Comparabilité

Dans le cas des études bilingues sur les corpus comparables, nous pouvons citer, en autres, les approches suivantes :

> *« La notion de comparabilité des corpus est une notion assez vague et peut prendre de multiples formes : comparabilité des vocabulaires, des genres, etc. A notre sens, la comparabilité des corpus bilingues doit être examinée sous l'angle de l'objectif visé et de la méthode qui les exploite et est liée à la notion de représentativité.» ... « La notion de comparabilité doit aussi être liée à l'objectif visé et à la méthode employée pour y parvenir. »* (Chiao, 2004).

La comparabilité des corpus bilingues peut se spécifier par les quatre critères suivants :

- *« Comparabilité de la couverture lexicale »*
- *« Comparabilité des fréquences relatives des mots »*
- *« Comparabilité des cooccurrences relatives »*
- *« Comparabilité des similarités distributionnelles entre mots »* (Chiao, 2004).

Des travaux (Sharoff, Babych et Hartley, 2006 ; Li, 2012) ont apporté un nouvel éclairage sur la comparabilité et les corpus comparables.

Par ailleurs, la comparabilité des corpus multilingues traite également les trois différents niveaux suivants :

> *« La profusion de documents accessibles dans des langues variées sur le Web incite à puiser dans ce réservoir pour constituer des corpus comparables. Néanmoins, cette tâche ne saurait se réduire à la simple collecte de documents partageant un vocabulaire commun. Il est nécessaire de respecter des caractéristiques communes telles que le thème et le domaine (Bowker et al., 2002) qui sont fixées avant la construction du corpus et qui sont fonction de sa finalité (McEnery et al., 2007). De nombreux travaux traitent de la construction de corpus à partir du Web (Baroni et al., 2006, Chakrabarti et al., 1999) mais aucun, à notre connaissance, n'est consacré à celle des corpus comparables, qui doit répondre à différentes contraintes. Nous fixons ainsi la comparabilité à trois niveaux : le domaine, le thème et le type de discours. »* (Goeuriot, Morin et Daille, 2009).

Dans nos recherches de veille multilingue textométrique, mis à part les critères importants que nous venons de citer ci-dessus, concernant la comparabilité des corpus, nous devons également examiner la convergence et la divergence textuelles de chacun des corpus retenus dans la mesure du possible.

Dans la section suivante, nous aborderons les outils proposés pour la veille multilingue.

1.7 Les logiciels pour la veille multilingue

Dans le contexte de nos études, nous utiliserons les logiciels Lexico 3 (Salem et Lamalle, 2009) et Le Trameur (Fleury, 2014), programmes de référence pour les analyses des données textuelles. Lexico 3 a l'avantage, par sa robustesse, de traiter les données de grands volumes et de restituer des représentations graphiques avec clarté. Des retours aux textes immédiats et des fonctions statistiques proposés par ces logiciels permettent d'orchestrer habilement les traitements statistiques, informatiques, les analyses quantitatives et qualitatives des textes.

Une série de thèses (Martinez, 2003 ; Zimina, 2004 ; Née, 2009; Sansonetti, 2010 ; MacMurray, 2012 ; Miao, 2012) ont été soutenues abordant différentes notions de la textométrie, ainsi dans les chapitres qui suivent, nous rappellerons simplement quelques fondements de cette école.

L'une des principales lacunes des systèmes de fouille de texte est leur incapacité à relier l'information extraite du contexte dans lequel un texte a été produit. Ces systèmes définissent avec difficultés l'événement correspondant en fait à un objet du monde réel. Un événement est constitué d'un réseau complexe de références, en laissant des empreintes lexicales dans le texte. Des techniques d'exploration de textes plus traditionnels utilisent des annotations qualitatives prédéterminées afin de formuler des interprétations sur les événements. La statistique textuelle se réfère aux informations textuelles quantitatives issues des analyses qualitatives. L'objectif de ce chapitre est donc de tester la statistique textuelle comme un moyen de l'exploitation appliquée à la fouille d'informations et plus particulièrement en utilisant des calculs de cooccurrences pour détecter des événements statistiquement significatifs. En analysant les trois sous-corpus de veille à caractère comparable par les calculs de cooccurrences, cette méthode espère révéler leurs empreintes lexicales, découvrant ainsi de nouvelles informations qui pourraient autrement passer inaperçues par des techniques de fouille d'information standard (MacMurray et Shen, 2010).

1.8 Notion d'événement

L'objectif général pour les systèmes d'exploration de texte est défini comme la détection d'informations ou d'événements pertinents en reliant ces derniers à d'autres événements produits dans le texte. Cependant, déterminer ce que sont exactement des renseignements pertinents ou événements ne se révèle pas une tâche facile, encore moins d'arriver à les projeter dans le monde réel.

Comme mentionné ci-dessus, l'une des principales lacunes des systèmes de « *Text Mining* » est leur incapacité à relier l'information extraite (Srivastava, Cooley et al, 2000) d'un contexte plus large dans lequel un texte a été produit. Il est difficile de définir un «événement» comme correspondant en fait à un objet du monde réel. Comme indiqué dans un certain nombre d'articles allant de la reconnaissance d'entités nommées à l'analyse du discours des noms propres, la désignation réelle des événements change avec le temps, non seulement sous forme graphique, mais aussi dans le sens (Slodzian, 2000 ; David, 2004 ; Poibeau, 2005 ; Moirand, 2007 ; Krieg-Planque, 2009 ; Née, 2009). De même que David stipule que *«Les productions journalistiques sont disséquées, comparées par des spécialistes qui s'appliquent à rendre visible et intelligible un travail médiatique assujetti à une actualité ontologiquement instable et volatile. »* (David, 2004).

Les événements ne sont donc pas seulement des entités ou des modèles, tels que définis par la plupart des systèmes d'extraction de l'information (Grishman et Sundheim, 1996 ; Grishman, 2003 ; Poibeau, 2003 ; Wright, 2006 ; Tufféry, 2012), mais sont plutôt directement liés aux corpus et ne feront que donner des informations sur le corpus dans lequel ils apparaissent. En outre, en essayant d'identifier un «événement», il faut noter qu'il est plus qu'une expression autonome (Veniard, 2007). Un « événement » est constitué d'un réseau d'autres références, soit dans le même article, soit d'une série d'articles (Adam, 1997 ; Veniard, 2007). Cette recherche est principalement basée sur les sept caractéristiques ci-dessous d'un événement dans les textes narratifs telles que définies par Adam (Adam, 1997) et Cicurel (Cicurel, 1994) :

1. Le cœur de l'événement : la description de l'événement par ses protagonistes, décrite par des journalistes ou expliquée par les scientifiques.
2. Les événements passés : d'autres événements de même nature, l'événement en cours est donc comparé à des événements passés.
3. Le contexte : l'atmosphère générale dans laquelle l'événement a eu lieu.
4. La périodicité du noyau de l'événement : la reproductibilité de l'événement.
5. L'arrière-plan ou commentaire : l'explication de l'événement.
6. Les réactions verbales : les réactions à l'événement par une variété de haut-parleurs - victimes, experts, représentants, etc.
7. Les histoires similaires : les histoires ne sont pas directement liées à l'événement, mais en relation avec l'atmosphère générale associée à l'événement.

Chacune de ces caractéristiques peut donner lieu à un certain nombre d'articles individuels ou peut être discutée au sein du même article. Ce modèle montre comment les événements sont discutés et reliés par la presse écrite comme un réseau de pièces complexes de l'information. Suite à ces arguments, deux hypothèses peuvent être formulées :

1.) L'entité nommée impliquée dans un événement aura une fréquence plus élevée et un plus grand nombre de cooccurrences car examinée par une série d'articles de journaux,

2.) Les événements laissent des «empreintes» lexicales dans le texte pouvant être révélées à l'aide des statistiques textuelles par la détermination de ce qui est statistiquement significatif dans un article donné.

1.9 Unités mesurables et intelligence

L'intelligence est un terme polysémique et abstrait.

Selon la définition du Trésor de la Langue Française informatisé disponible sur le site du CNRTL, il s'agit de :

> *« [Chez les êtres animés] Fonction mentale d'organisation du réel en pensées chez l'être humain, en actes chez l'être humain et l'animal ».*

En anglais, la définition du mot intelligence selon le dictionnaire d'Oxford[37] est :

1. *The ability to acquire and apply knowledge and skills.*
2. *The collection of information of military or political value.*

En chinois, deux traductions pour ce terme sont possibles:

1. 智力/zhì lì/intelligence des êtres humains
2. 情报/qíng bào/renseignement ou information

[37] www.oxforddictionaries.com

L'acception économique de l'intelligence varie en fonction de son contexte et évoque des champs sémantiques extrêmement multiples et complexes. Dans le cadre de notre recherche, nous nous préoccupons uniquement de son aspect idéologique, c'est-à-dire, dans le contexte économique et non philosophique, sans quoi on risquerait de passer des années à philosopher au travers les différents domaines associés. Par exemple, comme on dit, vaut-il mieux « se poser la question du bas de laine » ou « gagner des positions stratégiques sur les marchés ou le négoce » ?

La démarche intellectuelle qui constitue à expliciter la chaîne de relation entre les unités mesurables et l'intelligence au sens économique (figure 1.4, ci-dessous), n'en demeure pas moins digne d'intérêt.

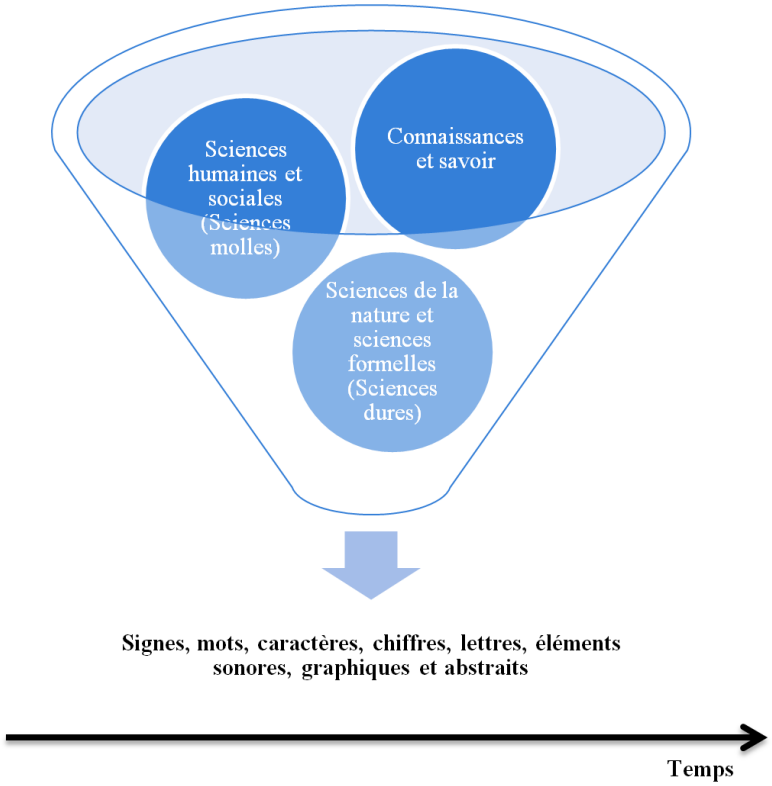


Figure 1.3 Perception des éléments et des sciences

Jusqu'alors ce qui était des données intangibles, ne relevant pas de la quantification, sont devenues désormais des objets de recherche tels que : éléments sonores, éléments graphiques, signes, chiffres, lettres, dessins, figures, graphies, images et mots. La figure 1.3 ci-dessus illustre que les sciences n'ont jamais trahi les objets d'études. L'homme se cherche, et cherche à comprendre, utiliser, manipuler et maitriser les différents types d'éléments et de données à travers leurs évolutions, autrefois peu quantifiables. Les sciences dures, sciences reposant sur le calcul, les sciences humaines, appelées parfois sciences molles, les connaissances et le savoir de la vie sont véhiculés par les données énoncées dans l'air du temps.

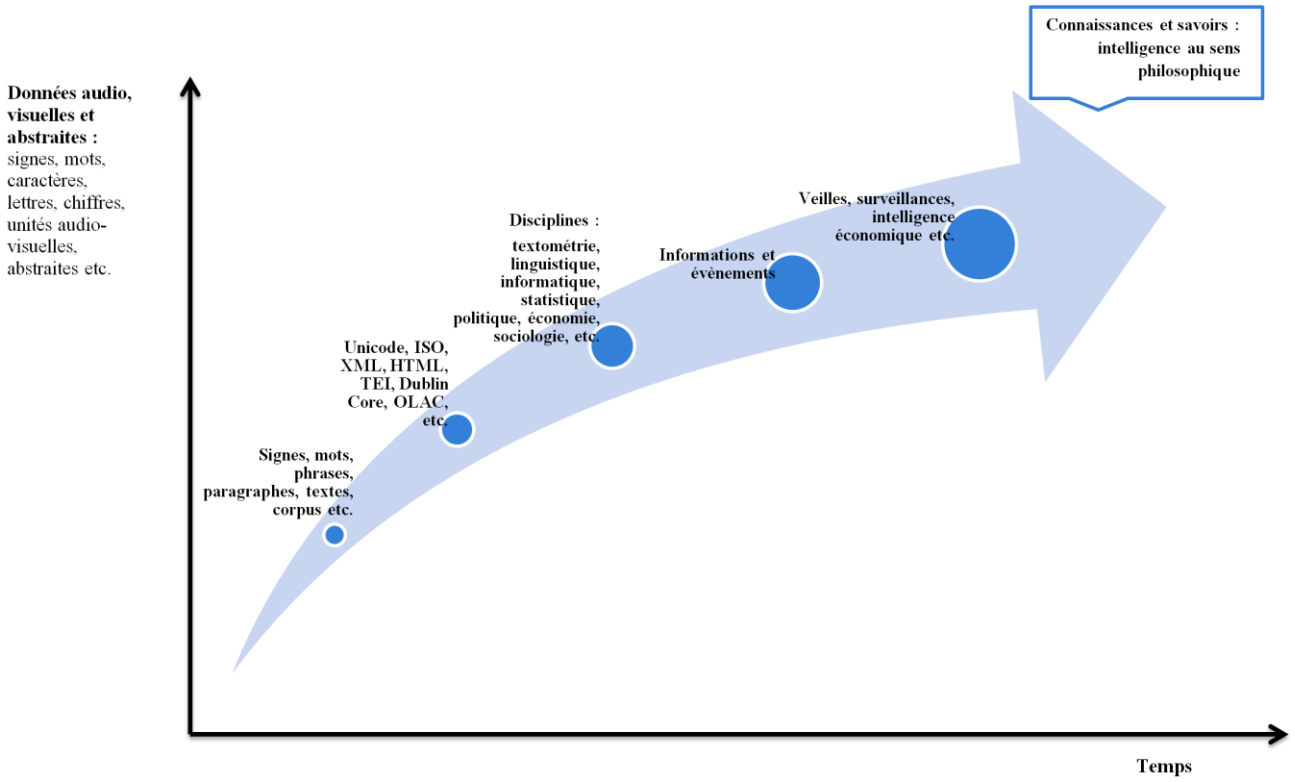


Figure 1.4 Epistémologie de la textométrie et l'intelligence

Afin de mieux comprendre et expliciter l'évolution de l'intelligence dans le contexte économique avec les unités mesurables notamment textuelles, une infographie épistémologique des relations entre la textométrie et l'intelligence a été tracée dans la figure 1.4 ci-dessus. L'évolution de l'informatique et des NTIC a bouleversé la faisabilité de la numérisation et la prise en charge des signes linguistiques (Unicode, ISO, XML, etc.), supports des connaissances (signes, chiffres, lettres, mots et langues), soumis à des contraintes physiques autrefois (carapaces, bambou, soie, papier, etc.). La diffusion et circulation d'informations (Liu, 2007) (technologie du web, informatique), la révélation des événements (information et communication) et la transmission du savoir (cultures et connaissances) ont dépassé complètement la notion du temps (vitesse de diffusion et circulation d'information). De nos jours, l'intelligence de l'homme (perception de l'univers) ne cesse d'être confrontée à l'emprise de répétition des médias (animations audio-visuelles), des défis des renseignements et de guerres de l'information (données, mégadonnées en signes, caractères, chiffres, lettres et audio-visuelles). Les méthodes pragmatiques telles que la statistique textuelle appliquée, veille informationnelle, intelligence compétitive, ouvrent de nouveaux horizons pour les recherches et développements des sciences et technologies. L'ère nouvelle se numérise tous azimuts, comme par exemple, les démarches administratives, les documents numériques, les courriers électroniques, les achats en ligne, la billetterie en ligne, l'archivage numérique.

La textométrie appliquée, située au carrefour des sciences molles et sciences dures, propose de mettre en relief efficacement les relations des différentes connaissances interdisciplinaires et transdisciplinaires, de manipuler au plus près les données qui véhiculent les informations entre autres stratégiques, de naviguer de manière concrète dans les masses de données textuelles.

Conclusion du chapitre

Nous avons donc vu le cadre conceptuel préalable à notre étude. Nous avons ainsi pu définir les notions de veille économique et d'intelligence économique. La veille économique correspond à un processus de fouille d'informations. La veille économique se rattache davantage à l'idée d'une prise de décisions au moyen d'informations disponibles. Dans le contexte de notre société d'information, la veille et l'intelligence économique deviennent indispensables aux acteurs économiques désirant rester compétitifs. Notre étude traitera plus particulièrement de la veille multilingue, c'est à dire une veille économique dans plusieurs langues : ici l'anglais, le chinois et le français. Ce terme peut avoir des acceptions différentes selon les langues. Ainsi, en chinois les termes ont une connotation militaire avec une idée de renseignement et de stratégie. En français et en anglais, les termes se rattachent davantage à la notion de renseignement. Le chapitre 2 aura pour objet de définir le thème d'étude.

Partie 2 Thématiques, sphères de communication et corpus

«Avoir une autre langue, c'est posséder une deuxième âme.»
- Charlemagne

« L'âme d'un peuple vit dans sa langue.»
- Johann Wolfgang von Goethe

« L'histoire du commerce est celle de la communication des peuples.»
- Montesquieu

« La musique est peut-être l'exemple unique de ce qu'aurait pu être - s'il n'y avait pas eu l'invention du langage, la formation des mots, l'analyse des idées - la communication des âmes.»
- Marcel Proust, *La Prisonnière* (1923)

Comme nous l'avons montré ci-dessus, le processus de veille obéit à un ordre défini. Ce processus nécessite de cerner en premier lieu les thèmes abordés : ici les énergies et l'environnement. A partir de ces thèmes, deux corpus de textes sont construits et seront traités au moyen d'outils statistiques et informatiques. Il est nécessaire d'utiliser un corpus de textes comparable, notion qui est définie plus haut comme :

> *« incluant des textes originaux dans différentes langues et répondant aux mêmes critères de genre, de temporalité, de registre, etc. Ici, aucun texte n'est la traduction de l'autre et on parle aussi de textes natifs. Un corpus comparable peut aussi être* ***multilingue****. Nous incluons aussi dans cette définition les corpus comprenant des textes rédigés dans une langue originale et des textes qui ont été traduits dans cette langue…* » (Bowker et Pearson, 2002; Kübler, 2011).

En suivant ce processus de veille multilingue évoqué plus haut, notre étude s'attachera à définir les notions d'énergies et d'environnement. Nous traiterons ensuite la problématique de l'énergie d'un point de vue mondial. La question des sources d'énergie alternative et de leur utilisation fera aussi l'objet de notre étude. Nous nous intéresserons plus précisément à la question du nucléaire. Enfin nous verrons les impacts de l'énergie sur l'environnement. Le présent chapitre permettra donc de clairement cerner les thèmes qui feront l'objet d'une veille multilingue.

2. Energies et environnement dans le monde

Présentation des deux thèmes

Afin d'illustrer la veille multilingue et ses méthodes, nous présenterons une série de résultats relatifs à des travaux liés à deux thèmes d'actualité, les énergies et l'environnement, au cœur des débats des sociétés dans lesquelles nous vivons (Mérenne-Schoumaker, 2007).

Les énergies apparaissent comme les locomotives de toutes les économies du monde et de la vie de l'humanité. Un type d'énergie soumis à de fortes polémiques, le nucléaire, polarise notre attention, et plus particulièrement, les centrales et l'évolution de leurs réacteurs, dont la *star* des nouveaux réacteurs de troisième génération, *EPR* (réacteur pressurisé européen), sujet suscitant encore beaucoup d'interrogations dans au moins trois sphères de communication de la planète.

Le choix des énergies est déterminant pour l'environnement, car leurs liens ne sont pas anodins sur les facteurs du changement climatique et notamment la production de gaz à effet de serre (Bertel et Naudet, 2004 : 22). Les impacts et les consommations des énergies s'identifient comme une obligation afin de préserver l'environnement dans lequel nous vivons et de lutter contre le changement climatique, nous imposant de maîtriser les deux thèmes (énergies et changement climatique) qui vont de pair. D'ailleurs, le projet de décision de la COP 21 confirme notre constat :

> *«Reconnaissant que les changements climatiques représentent une menace immédiate et potentiellement irréversible pour les sociétés humaines et la planète et qu'ils nécessitent donc la coopération la plus large possible de tous les pays ainsi que leur participation dans le cadre d'une riposte internationale efficace et appropriée, en vue d'accélérer la réduction des émissions mondiales de gaz à effet de serre, ...»* - Convention-cadre sur les changements climatiques, 12 décembre 2015

Aujourd'hui, ces deux domaines étroitement liés sont l'objet de nombreux débats à propos de l'utilisation de l'énergie nucléaire qui entraîne de multiples interrogations, comme par exemple ses impacts sur notre environnement ou encore la sécurité nucléaire et ses conséquences sur les populations, la faune et la flore.

Contexte socio-économique et enjeux

Des réflexions introduites par les concepts de l'I.E imposent de considérer l'environnement comme un point de départ de ce travail de veille textométrique et de mettre en évidence les discours tenus sur le nucléaire au sein des corpus relatifs à l'énergie et aux problèmes environnementaux. L'environnement relève intrinsèquement de deux thèmes fondamentaux de l'humanité : la santé et la sécurité. Toutefois, une question de « veillabilité[38] » se pose sérieusement au cours de ce travail. Afin d'opérer une série d'observations autour de la notion de nucléaire et ses conséquences environnementales, mais également de relever d'autres impacts provoqués par diverses causes, les informations blanches[39] nous suffisent-elles et qu'en dégagent-elles ? Comment le nucléaire, au fur et à mesure, impacte notre environnement, sachant que le nucléaire n'est pas la seule cause de changement ? Et quelles seront les évolutions socio-économiques ? Les textes de la presse et des médias qui diffusent des informations en France, aux Etats-Unis et en Chine vont-ils nous révéler une cartographie de l'information ?

Dans les paragraphes suivants, afin de mieux cerner le contexte de nos thèmes retenus, un panorama de la situation énergétique (Bertel et Naudet, 2004) et environnementale est dressé.

2.1 L'énergie aujourd'hui dans le monde

En 2015, la population mondiale s'élève à environ 7,3 milliards d'habitants[40] pour atteindre 9 milliards de personnes en 2050 d'après les prévisions. Par cette croissance démographique, la demande d'énergie est de plus en plus importante afin d'assurer le confort de tous, celle-ci étant plus grande dans les pays émergents tels que les pays de la BRICS (Brésil, Russie, Inde, Chine, Afrique du Sud), mais aussi la question de la croissance économique est le deuxième facteur de cette augmentation en demande d'énergie. Par ailleurs, la diminution des ressources énergétiques traditionnelles pose un sérieux problème aux Etats. Dans une soixantaine d'années, les énergies telles que le charbon, le pétrole, voire l'uranium seront quasiment épuisées. Il faut anticiper ce changement et trouver dès maintenant de nouvelles sources d'énergies renouvelables et non polluantes. Or, ces états qui se développent ont des besoins de plus en plus importants en énergie, notamment pour le développement de leur économie. C'est pourquoi produire de l'énergie dans une voie durable et verte est vital (Dessus, 2011).

De nos jours, nous disposons de multiples sources d'énergie, fossiles, renouvelables et nucléaire qui procurent une quantité d'énergie plus ou moins importante avec des conséquences plus ou moins dramatiques. Il faudra trouver d'autres moyens de produire de l'énergie, en abondance et à un moindre coût. Par exemple, un des avantages du nucléaire est son moindre coût par rapport aux autres énergies, notamment au niveau du coût de l'électricité, moitié moins chère en France qu'en Allemagne (Oriol, Meinzel et al, 2013).

[38] Mise en adéquation de la veille
[39] Se reporter à l'annexe A, section A.3 : Sources formelles et informelles
[40] www.ined.fr (consulté le 31/07/2015)

Les énergies actuelles demandent beaucoup de moyens et n'assurent aucunement l'avenir de notre planète. Ces énergies dites fossiles (charbon, pétrole et gaz naturel) couvrent la majorité des besoins mondiaux, avec une part d'environ 80% contre 7% pour l'énergie nucléaire et 13%[41] pour l'énergie renouvelable, et sont souvent sources de pollution et par conséquent néfastes pour la faune et la flore. Quant au nucléaire, il suscite des questions nombreuses et complexes, par exemple que ce soit sur la sécurité énergétique, la temporalité (un enchaînement de plusieurs événements) mais aussi le risque environnemental soulevant de grands débats politiques (conséquences des catastrophes du type de Fukushima survenue en mars 2011) et parfois de vives réactions de la part des populations concernées.

Face à ces énergies, la filière des énergies renouvelables (Vernier, 2014) est beaucoup plus saine pour tous. Elle ne demande pas de transports sur des distances importantes puisque les sources sont dispersées dans le monde. L'environnement naturel offre partout des énergies renouvelables que l'homme peut utiliser directement ou convertir là où il en a besoin (énergies du soleil, du vent, de l'eau, ou encore des plantes).

Les puissants pôles du globe dont la France, les Etats-Unis, la Chine (Kreft, 2006), ont un fort besoin énergétique (Gouysse, 2010). Les deux premiers, puissances économiques de longue date, ont une production d'énergie à l'image de celle du monde, contrairement à la Chine, un pays émergent. Sa production énergétique provient à 87% des combustibles fossiles (Lafargue, 2013 : 22) et 13% des énergies renouvelables, dont une part inférieure à 1% représente l'énergie nucléaire[42]. Pour l'année 2010, les émissions de CO2 du secteur énergétique s'établissaient à plus de 30 milliards de tonnes. La Chine devance les Etats-Unis en émettant 25 % des émissions de CO2. Ces deux pays représentent plus de 40 % des émissions mondiales[43] (Henriet et Maggiar, 2012 : 148).

Par la croissance démographique et le développement technologique, les besoins énergétiques du globe vont monter en flèche ces prochaines années, d'où la nécessité de trouver un apport énergétique intéressant tout en préservant l'environnement.

L'énergie pour quelle utilisation ?

Utiliser de l'énergie est capital dans toutes les activités, c'est une aide précieuse pour chacun. En effet, elle a été utilisée et banalisée à partir de la Révolution industrielle avec l'utilisation du charbon dans l'industrie et s'est développée jusqu'à maintenant avec l'usage majoritaire de l'électricité. L'énergie permet de produire en grande quantité à moindre coût par rapport à une main d'œuvre. C'est pourquoi la part de la consommation d'énergie dans l'industrie est importante en France et aux Etats-Unis avec environ 21%[44] de la consommation énergétique globale et en Chine avec 48%[45] en 2012. Cela s'explique par le développement rapide de ce pays et la multiplication d'entreprises.

Toutefois, le secteur du transport est un des secteurs les plus gourmands en énergie. En effet, en corrélation avec la croissance démographique, le nombre des différents moyens de transport s'accroît (bus, train, avion, voiture). Ainsi, les besoins énergétiques de ce secteur sont de plus en plus importants, tandis que 95%[46] proviennent de la combustion du pétrole, une ressource limitée qui se

[41] www.connaissancedesenergies.org/fiche-pedagogique (consulté le 31/07/2015)
[42] www.iea.org/statistics/statisticssearch/report/?year=2012&country=CHINA&product=Balances (consulté le 31/07/2015)
[43] www.developpement-durable.gouv.fr/IMG/pdf/chapitre_2.pdf (consulté le 01/08/2015)
[44] https://flowcharts.llnl.gov/ (consulté le 31/07/2015)
[45] www.iea.org/statistics/statisticssearch/report/?year=2012&country=CHINA&product=Balances (consulté le 31/07/2015)
[46] www.cea.fr/jeunes/themes/l-energie/la-production-d-energie/l-energie-pour-quoi-faire (consulté le 31/07/2015)

raréfie. La France est un pays qui consacre une part plus importante de sa consommation énergétique pour les transports, 31.6%[47], suivie des Etats-Unis 27,1%[48] et de la Chine avec 14%[49].

Enfin, l'énergie est aussi utilisée dans le secteur tertiaire et résidentiel, surtout pour le chauffage. La population croissante a besoin notamment de se chauffer, s'éclairer, se rafraîchir. De plus, tous les produits ménagers, informatiques, etc. nécessitent une source d'énergie pour assurer leur fonctionnement. Avec la banalisation de ces outils, s'explique la hausse de la consommation d'énergie, principalement celle de l'énergie électrique. 80% de l'électricité chinoise sont fournis par le charbon (Martinot et Li, 2007) contre 8,6% par les énergies propres (Voïta, 2010). La France, étant le pays le moins peuplé parmi les deux autres puissances, a une consommation énergétique de 44,8%[50] dans ces secteurs, contre 25,3%[51] en Chine et 20%[52] aux Etats-Unis.

Quelles énergies ?

Des énergies fossiles

Les énergies fossiles sont les plus utilisées. Toutefois, elles comptent de nombreux inconvénients : leur quantité est limitée, leurs prix varient souvent, elles dégagent énormément de gaz à effet de serre. Parmi celles les plus utilisées nous trouvons le charbon, bête noire des défenseurs de l'environnement, source d'énergie de la révolution industrielle du XIX^{e} siècle, mais encore massivement utilisé aujourd'hui, aussi bien dans les pays développés comme l'Allemagne que chez les émergents, et ce, malgré son caractère polluant avéré, le pétrole (la première source d'énergie mondiale aujourd'hui) et le gaz naturel provenant des couches géologiques du sous-sol.

Le charbon produit 41% de l'électricité dans le monde, plébiscité pour son prix bas[53]. En 2014, la production mondiale a atteint 39,33 milliards de tonnes équivalent pétrole. Les trois principaux producteurs mondiaux sont la Chine, les États-Unis et l'Inde[54]. La Chine représente à elle seule plus de 40% de la demande mondiale. Le charbon représentait 73% de son mix-énergétique en 2014, toujours selon Enerdata. Mais l'an dernier, pour la première fois depuis 1999, la consommation de charbon du pays a baissé, en grande partie en raison du ralentissement de la croissance économique. Il est à noter que l'usage accru du charbon dans les centrales européennes coïncide avec l'avènement du pétrole et du gaz de schiste aux Etats-Unis.

Le charbon est pointé du doigt en raison de la pollution qu'il engendre. Tout d'abord lors de son extraction, car les mines génèrent de grandes quantités de CO2 et polluent parfois les nappes phréatiques. Vient ensuite la pollution liée à l'exploitation même du charbon.

[47] www.developpement-durable.gouv.fr/IMG/pdf/Ref_-_Bilan_energetique_de_la_France.pdf (consulté le 31/07/2015)
[48] https://flowcharts.llnl.gov/ (consulté le 31/07/2015)
[49] www.iea.org/statistics/statisticssearch/report/?year=2012&country=CHINA&product=Balances (consulté le 31/07/2015)
[50] www.developpement-durable.gouv.fr/IMG/pdf/Ref_-_Bilan_energetique_de_la_France.pdf (consulté le 31/07/2015)
[51] www.iea.org/statistics/statisticssearch/report/?year=2012&country=CHINA&product=Balances (consulté le 31/07/2015)
[52] https://flowcharts.llnl.gov/ (consulté le 31/07/2015)
[53] Selon Enerdata, cabinet intelligence et consulting, https://www.ademe.fr/sites/default/files/assets/documents/89845_brochure-perspectives-energetiques-mondiales.pdf (consulté le 02/08/2015)
[54] www.bp.com/content/dam/bp/pdf/Energy-economics/statistical-review-2015/bp-statistical-review-of-world-energy-2015-full-report.pdf (consulté le 31/07/2015)

En Chine, il y a une vraie prise de conscience de cette pollution et celle-ci a fait d'importants gains d'efficacité dans l'industrie électrique en fermant des vieilles centrales et en en ouvrant des nouvelles, plus efficaces et moins consommatrices en charbon (Martinot et Li, 2007).

Les Etats-Unis sont de gros producteurs et aussi consommateurs d'énergies polluantes, mais l'utilisation des énergies renouvelables (Le Leuch, 2010) progresse régulièrement face à la pression mondiale et à la question du changement climatique qui anime les débats de la politique mondiale et par conséquent celle de l'Amérique. Depuis l'arrivée au pouvoir de Barack Obama, l'énergie et l'environnement sont devenus des priorités (Méritet, 2009). La transition énergétique est toutefois lente, quant au nucléaire, les Américains restent marqués par l'image d'un nucléaire non économique, « *nuclear is not competitive* » (Chavardès, 2009).

Des énergies renouvelables

Il faut se rendre à l'évidence le monde a besoin d'énergie, de plus en plus d'énergie due d'une part à la croissance démographique et/ou économique, d'autre part à la diminution des ressources énergétiques traditionnelles. Il faut anticiper ce changement et trouver dès maintenant de nouvelles sources d'énergies renouvelables et non polluantes (Multon, 1998, 2003).

Les énergies renouvelables sont celles qui sont les plus développées actuellement, afin de limiter les émissions de CO_2 produites en grande quantité par les énergies fossiles. Ces énergies sont inépuisables et disponibles dans la nature comme par exemple, l'énergie solaire, l'énergie éolienne, l'énergie hydraulique ou encore géothermale, mais également le traitement des déchets et la biomasse. Cependant la part de ces énergies renouvelables ne progresse que très lentement.

Dans le monde en 2010, la part de ces énergies renouvelables atteint 17% des besoins de consommation d'énergie. Par exemple, pour l'Union Européenne, ces énergies représentent 15% de la consommation brute d'énergie contre 7,9% en 2004 et 12 ,1% en 2010[55], mais n'atteignent pas encore les 20%, objectif pour 2020. La Chine diversifie également sa production d'énergies en développant les énergies renouvelables (Guermond, 2007 ; Guermond et Ma, 2011). Le nouveau président chinois XI Jinping s'engage à « stopper » la croissance des émissions de CO2 à l'horizon 2030[56]. Les responsables publics sont de plus en plus conscients des avantages offerts par les énergies renouvelables, notamment la sécurité énergétique, la dépendance réduite aux importations, la réduction des émissions de gaz à effet de serre, la prévention contre les pertes dans la biodiversité, l'amélioration de la santé, la création d'emplois, le développement rural, etc. (Swaminathan, Rahmanian, Bertini et al, 2013).

Le soleil est la plus grande source d'énergie sous forme de lumière et de chaleur utilisable, elle est à la portée de tous, et illimitée. Les énergies ainsi produites par l'astre sont donc propres et n'émettent pas de gaz à effet de serre. L'énergie solaire est la source de toutes les énergies (sauf l'énergie nucléaire, géothermique et marémotrice). Les rayons émis sont source de chaleur, et puisque la température n'est pas la même partout, des transferts d'énergie thermique, par des mouvements de convection se perpétuent dans l'atmosphère. Ceci explique l'existence de vent ainsi que des courants océaniques de surface ; la convection liée à la rotation de la Terre provoque également des courants océaniques profonds. Ces différents phénomènes sont alors exploités et génèrent l'énergie éolienne (Multon, Roboam et al, 2004), hydraulique et marine. Toutes ces énergies sont dites renouvelables, et sont

[55] http://ec.europa.eu/eurostat/web/energy/data/main-tables (consulté le 01/08/2015)
[56] https://www.lenergieenquestions.fr/tag/Chine (consulté le 01/03/2015)

aujourd'hui des énergies à privilégier car elles n'émettent pas de gaz à effet de serre. La COP21 encourage ces initiatives :

> *«Considérant la nécessité de promouvoir l'accès universel à l'énergie durable dans les pays en développement, en particulier en Afrique, en renforçant le déploiement d'énergies renouvelables, ...»* - Convention-cadre sur les changements climatiques, 12 décembre 2015

Le solaire

Les rayons du soleil sont exploités avec les énergies suivantes : l'énergie solaire photovoltaïque (utilisation de modules photovoltaïques pour produire de l'électricité), l'énergie solaire thermique (utilisé pour le chauffage domestique ou des eaux sanitaires), l'énergie solaire thermodynamique (production de l'électricité grâce à une production de chaleur).

En France, l'énergie solaire est surtout utilisée pour la consommation des particuliers. De plus, le pays n'a pas un fort ensoleillement, sauf dans le Sud. Cette énergie représente 4,9% de la production d'électricité et 4,6% de cette même production mais d'origine renouvelable en France en 2013[57].

Aux Etats-Unis, le potentiel de l'énergie solaire est immense. En effet, on y trouve la plus grande centrale solaire du monde, en Californie. Cette énergie représente 14,3% de la production d'électricité d'origine renouvelable dans le pays et 5% en Chine[58].

L'éolien, l'hydraulique

Les pays comme la France, les Etats-Unis, voire la Chine sont en train de miser sur cette autre source d'énergie renouvelable qu'est le vent.

En Chine, la production d'électricité par « Eole », fait de rapides progrès et augmente (Ravignan de, 2008); des sites favorables à l'implantation d'éoliennes sont nombreux comme dans le désert de Gobi et sur les hauts plateaux tibétains et même l'éolien « offshore » continue de croître (Sawin, Bhattacharya et al, 2012). Quant à l'hydraulique, c'est une solution privilégiée des Chinois, énergie intéressante grâce aux puissants fleuves traversant la Chine en construisant des barrages[59]. Cependant, la Chine s'intéresse depuis de longues années à la petite hydraulique (Zhu et Pan, 2007).

Des énergies innovantes

Les biocarburants, ou agro-carburants, représentent l'ensemble des carburants de toutes formes (liquide, gaz, solide) provenant d'une matière vivante dans un écosystème, la biomasse. Leur utilisation est toujours associée à celle d'un carburant fossile, c'est une source d'énergie utilisée dans les transports. Il existe aujourd'hui deux filières industrielles: l'éthanol, fabriqué à partir d'amidon et de sucre (betteraves, céréales, etc.) mélangé à l'essence et le biodiesel à partir d'huile végétale, de graisse animale ou huile usagée, utilisé dans les moteurs diesel.

Cette innovation au grand potentiel a pour objectif de réduire les émissions de gaz à effet de serre, de préserver les ressources d'énergie fossile mais aussi de diversifier les sources d'énergie : une alternative intéressante à la production d'énergie, mais sans dégrader l'environnement. Cet objectif de

[57] www.connaissancedesenergies.org/fiche-pedagogique/chiffres-cles-production-d-energie (consulté le 31/07/2015)
[58] http://jeunes.edf.com/article/le-solaire-photovoltaique-en-france,176 (consulté le 31/07/2015)
[59] www.duclair-environnement.org/2014/12/12/en-chine-un-projet-hydraulique-pharaonique-entre-en-activite/ (consulté le 01/08/2015)

la réduction des gaz à effet de serre est le leitmotiv de toutes les COP, leitmotiv que l'on retrouve à divers niveaux du projet de décision de la COP21 :

> *« Insistant avec une vive préoccupation sur l'urgence de combler l'écart significatif entre l'effet global des engagements d'atténuation pris par les Parties en termes d'émissions annuelles mondiales de gaz à effet de serre jusqu'à 2020 et les profils d'évolution des émissions globales compatibles avec la perspective de contenir l'élévation de la température moyenne de la planète nettement en dessous de 2 °C par rapport aux niveaux préindustriels et de poursuivre l'action menée pour limiter l'élévation des températures à 1,5 °C,… »*

Toutefois, il existe trois générations de biocarburant, caractérisées par différentes origines de la biomasse et des procédés utilisés mais seule la première génération est industrialisée et produite en faible quantité. De plus, cette alternative est toujours source de débat en raison de sa concurrence avec la ressource alimentaire.

L'énergie des marées (houle, courants marins) est également largement exploitée par la Chine ainsi que la biomasse marine pour produire des biocarburants à partir des algues. L'utilisation des technologies dans ce secteur devrait connaître une forte croissance (Wang, Yuan et al, 2011).

Une autre possibilité d'énergie : l'hydrogène (H), élément chimique le plus léger et l'un des premiers éléments à avoir existé dans l'univers, élément le plus important composant notre monde et présent partout, notamment dans l'eau, les hydrocarbures, mais n'existant pas à l'état pur. Aujourd'hui se porte l'étude sur la production de dihydrogène (H_2), une molécule particulièrement énergétique. La combustion de 1kg dihydrogène procurerait 3 fois plus d'énergie que celle de 1kg d'essence[60]. De plus, cette combustion n'est pas du tout polluante car elle ne dégage que de l'eau selon l'équation suivante : $2H_2 + O_2 \rightarrow 2H_2O$. C'est pourquoi la molécule est très intéressante dans l'objectif de diminuer les émissions de CO_2. Il est possible de produire du dihydrogène à partir d'eau (électrolyse), d'hydrocarbure ou de la biomasse (gazéification).

Des idées innovantes

En règle générale, on admet que l'homme cherche à aménager son territoire en restant compétitif, par le biais d'infrastructures (axes de transports, bâtiments, etc.). Toutefois, ces aménagements ont également un impact sur l'environnement en raison de l'utilisation de différents matériaux. C'est pourquoi il est important de construire « intelligent » tout en pensant à la Terre en restreignant les pertes d'énergie et en utilisant les énergies renouvelables. L'objectif est de réduire les dépenses énergétiques, aujourd'hui possible grâce aux systèmes d'isolation, limitant les transferts d'énergie thermique, la mise en place de différents appareils diminuant les émissions de gaz à effet de serre comme les systèmes de ventilation. Les énergies renouvelables peuvent être mises au service de ces nouvelles infrastructures, telles que les panneaux solaires ou les éoliennes par exemple.

En France, le secteur de l'habitat représente 20%[61] des émissions de gaz à effet de serre, donc il parait important de développer des bâtiments « intelligents »[62]. Cependant, peu de grands projets concrets

[60] www.cea.fr/jeunes/themes/les-energies-renouvelables/l-hydrogene/caracteristiques-de-l-hydrogene (consulté le 31/07/2015)
[61] www.cea.fr/jeunes/themes/les-energies-renouvelables/l-essentiel-sur-les-batiments-intelligents%20 (consulté le 31/07/2015)
[62] Un bâtiment intelligent est un bâtiment équipé de systèmes de contrôle-commande et de monitoring pour optimiser les consommations d'énergie

n'ont encore abouti comme le remarque Olivier Cottet[63] « *la France a une position de leader d'opinion en Europe dans le domaine de la performance énergétique. Nous sommes très volontaristes même si nous sommes encore trop souvent conceptuels, avec un foisonnement de labels et de normes un peu compliqués* ».

En Asie, le marché des bâtiments intelligents est en pleine expansion, liée à deux phénomènes, la croissance démographique et l'urbanisation. Ainsi il y est aussi important de construire tout en préservant l'environnement. En Chine 36 villes intelligentes sont en cours[64] de réalisation, pays où il est nécessaire de réduire la consommation d'énergie au vu de ses besoins énergétiques actuels.

Quant à l'énergie nucléaire, elle repose sur le phénomène de la fission nucléaire : un noyau père lourd au contact d'un neutron se désintègre en deux noyaux fils, plus légers. Cette collision forme de l'énergie qui est utilisée et très importante comme le confirme « *un seul gramme d'uranium fournit autant d'énergie que 3 tonnes de charbon.* »[65]. Toutefois, les centrales nucléaires utilisent les noyaux d'uranium, mais malheureusement, le stock de ceux-ci diminue de plus en plus et bien que cette énergie n'émette pas de gaz à effet de serre, les déchets produits sont importants et néfastes. C'est pourquoi il est nécessaire de mettre en place une gestion de ces déchets radioactifs.

Energies et énergie nucléaire

Le nucléaire est un mot vilipendé, pourtant ce terme relatif au noyau de l'atome[66] fait peur (la bombe, les sous-marins, la force de dissuasion, les menaces de la part de certains pays, etc.), un mot synonyme de guerre, un mot qui revient surtout quand surviennent des catastrophes telles Tchernobyl en avril 1986 et plus récemment Fukushima en mars 2011. Après ces moments de panique qui saisissent la planète entière, suivent de grands débats philosophiques, de grandes déclarations de nos politiques et puis la réalité reprend le dessus.

Comme le relate Guillaume Roquette[67], *« …Trois ans presque jour pour jour après Fukushima, on n'a jamais construit autant de nouveaux réacteurs nucléaires, … En Grande-Bretagne, David Cameron annonce la mise en chantier de deux réacteurs EPR, …. Quant à l'Allemagne, seul pays au monde à avoir annoncé un arrêt total du nucléaire après Fukushima, elle s'interroge sérieusement sur le bien-fondé de sa transition énergétique … »* - Convention-cadre sur les changements climatiques, 12 décembre 2015

Si nous prenons la définition du nucléaire au sens physique du terme : « *l'énergie nucléaire c'est l'énergie dégagée par la fission du noyau d'atomes lourds ou la fusion de noyaux d'éléments légers …* »[68]. En résumé, lors de la fission, il y a production d'énergie mais aussi de chaleur, cette chaleur est utilisée dans la production de vapeur qui elle-même sert à créer l'électricité. Cependant, l'atome fait toujours peur, pourtant le nucléaire a d'autres fins que celles de la guerre et aujourd'hui, il est surtout utilisé dans la production d'électricité (Bertel et Naudet, 2004 : 27).

La situation française du « quasi tout nucléaire » est historique, puisque c'est à la fin du XIX^{e} siècle en 1896 qu'Henri Becquerel découvre la radioactivité et deux ans plus tard Pierre et Marie Curie, le

[63] Olivier Cottet, directeur marketing du Programme HOMES, programme européen dans le bâtiment vert dont Schneider Electric est à l'origine
[64] www.infohightech.com/les-batiments-intelligents-en-asie-en-2012/ (consulté le 31/07/2015)
[65] www.cea.fr/jeunes/themes/l-energie/la-production-d-energie/les-differentes-energies (consulté le 31/07/2015)
[66] http://www.cnrtl.fr/definition/nucl%C3%A9aire
[67] Guillaume Roquette Le Figaro magazine 14 mars 2014, édito
[68] www.cnrtl.fr/lexicographie/nucléaire (consulté le 01/08/2015)

polonium et le radium[69] (se reporter à l'annexe D). Dès la fin de la deuxième guerre mondiale, les politiques français, le Général de Gaulle[70], puis le président Giscard d'Estaing, décident très tôt de développer l'énergie nucléaire, la France devenant ainsi un des premiers pays producteurs de cette énergie[71].

Actuellement, la France doit faire face notamment aux vieillissements de ces centrales nucléaires, aux coûts élevés de leurs démantèlements[72], aux dépassements en coût et en durée de la construction de l'EPR à Flamanville.

Quant aux Etats-Unis, l'opinion publique a toujours été hostile au nucléaire, hostilité qui s'explique par trois raisons principales (Chavardès, 2009). La première est le traumatisme de l'accident de *Three Mile Island*, événement fortement médiatisé, le « *nuclear is not safe* », la deuxième raison est une conséquence de la première et concerne la mise en sécurité des centrales existantes, le « *nuclear is not competitive* », et la troisième concerne le problème de stockage des milliers de combustibles irradiés. Cependant l'opinion générale évolue et reconnaît que mener une politique énergétique diversifiée permettrait de réduire les émissions de CO2.

Et en Chine, la part du nucléaire dans le mix énergétique est encore faible, mais elle augmente sa capacité en réacteurs nucléaires (Machenaud, 2005). Même si celle-ci reste prudente, elle devrait mettre en chantier ou en service plus de 150 réacteurs d'ici 2025[73] (se reporter à l'annexe G).

Les questions suscitées par le nucléaire sont denses, nombreuses et complexes, par exemple que ce soit sur la sécurité énergétique, la temporalité mais aussi le risque environnemental. Toutes ces questions soulèvent de grands débats politiques (Schneider, 2001).

2.2 Energie et environnement

En physique, l'énergie est définie comme la « *capacité d'un corps ou d'un système à produire du travail mécanique ou son équivalent* » ou encore comme l'« *ensemble des forces susceptibles de mouvoir les machines nécessaires à la production industrielle ou à la vie domestique* »[74].

Comme on l'a vu, le besoin d'énergie non renouvelable pour n'importe quel type d'activité dégrade l'environnement. En effet, les énergies fossiles et nucléaires nécessitent une matière première qui s'épuise au fur et à mesure, alors les réserves diminuent. Aujourd'hui, la réserve de pétrole est estimée pour 40 années, de charbon à 230 années, de gaz naturel à 70 années et d'uranium à 50 années[75]. C'est pourquoi il faut valoriser les énergies renouvelables, leurs ressources étant illimitées. Mais il faut tout de même savoir que même ces « énergies vertes » ont un certain impact sur l'environnement cependant négligeable face aux autres énergies.

L'environnement est très touché par la production d'énergie avec les émissions de gaz à effet de serre, les déchets radioactifs, aussi les gouvernements tentent d'y remédier comme en France avec la stratégie de traitement-recyclage.

---

[69] www.cnrtl.fr/definition/polonium et www.cnrtl.fr/definition/radium (consulté le 01/08/2015) et se reporter à l'annexe D, Le nucléaire et la politique française.
[70] www.ina.fr/video/I11074500/le-general-de-gaulle-a-propos-de-l-energie-atomique-video.html (consulté le 01/08/2015), extrait du discours du Général de Gaulle du 25/09/1963 à Orange
[71] www.vie-publique.fr/politiques-publiques/politique-nucleaire/histoire-politique-nucleaire-civil/ (consulté le 01/08/2015)
[72] « Politiques, stratégies et coûts de démantèlement : un tour d'horizon international », AEN, 2003, n°21.2
[73] www.lemonde.fr/idees/article/2012/11/05/l-avenir-chinois-du-nucleaire-mondial_1785878_3232.html (consulté le 01/08/2015), et se reporter à l'annexe G, La politique énergétique chinoise.
[74] http://cnrtl.fr/definition/bhvf/énergie (consulté le 31/07/2015)
[75] www.cea.fr/jeunes/themes/l-energie/la-production-d-energie/energie-et-environnement (consulté le 1/07/2015)

Environnement

Quant au terme environnement, quelle en est la définition ? L'une d'elles est la suivante : « *Ensemble des choses qui se trouvent aux environs, autour de quelque chose …* » et par extension « *Ensemble des éléments et des phénomènes physiques qui environnent un organisme vivant, se trouvent autour de lui …* »[76].

Deux des plus grands émetteurs de CO2[77] sont entre autres la Chine (9,9 milliards) et les Etats-Unis (5,2 milliards), mais ces deux pays commencent à prendre de sérieuses mesures et les bénéfices se font sentir sur l'environnement. Quant à la France, elle fait également des progrès dans ce domaine. Le ministère de l'Écologie, du Développement durable et de l'Energie a publié un rapport en 2006 qui annonce une *« diminution considérable de 12 % des émissions de gaz à effet de serre … C'est surtout grâce au progrès tangible effectué dans le domaine de l'agriculture* ». D'autres bons chiffres concernent l'utilisation des sources d'énergies renouvelables par les Français, et la baisse de leur consommation de chauffage. Par contre, plusieurs constats négatifs : l'augmentation de la température et de la pollution marine (par exemple en Méditerranée) ou encore la progression du nombre de personnes souffrant d'allergies[78].

Les impacts du nucléaire sur la santé de l'homme et sur les écosystèmes sont l'objet de débats nationaux et internationaux. Il est nécessaire d'améliorer la connaissance des conséquences environnementales liées à la présence ou aux rejets de substances radioactives et chimiques en lien avec les activités nucléaires. Des questions comme « quels sont les effets d'une exposition à de faibles doses de radiations ? » ou « quels seront les effets sur les prochaines générations ? » restent remplies d'incertitudes.

Indéniablement, nous savons que le nucléaire comporte des dangers pour les organismes, risques de radioactivité lors du dégagement de l'énergie des noyaux. La radioactivité est un phénomène physique naturel. Dans notre environnement, nous sommes constamment exposés à une radioactivité naturelle, donc à faible dose. Par contre, lors de catastrophes nucléaires, la radioactivité est colossale et dans ce cas, les risques pour la santé en sont grandement accentués. Les conséquences immédiates sont des décès, personnes irradiées ou personnes atteintes de cancer de la thyroïde, des milliers de personnes contaminées et déplacées, des mutations génétiques et des organismes contaminés.

Le nucléaire laisse son empreinte sur la faune et la flore, exposées aux mêmes radiations que les humains. La contamination des sols, mesurée en becquerels, fait suite à des accidents nucléaires civils ou militaires. Mais un des grands débats du nucléaire porte sur les déchets. Les centrales nucléaires produisent beaucoup de déchets radioactifs (Landais, 2012), déchets qu'il va falloir traiter pour en recycler une partie, les autres devront être stockés. Mais comment faire ? Les populations ne font plus confiance au nucléaire, et pourtant cette énergie devra faire l'objet de surveillance et de contrôle afin de rassurer celles-ci, cette énergie devient indispensable au XXI^{e} siècle. C'est à ce niveau que les politiques doivent intervenir et prendre des décisions fondamentales en réglementant cette énergie (contrôles, règles, etc.).

[76] http://www.cnrtl.fr/lexicographie/environnement (consulté le 31/07/2015)
[77] http://www.energie-environnement.fr/emissions-de-co2-etat-des-lieux/, chiffres en milliards de tonnes de CO2 émis en 2013 (consulté le 01/08/2015)
[78] *2014, l'environnement en France, les grandes tendances*, Ministère de l'écologie, du développement durable et de l'énergie : http://www.developpement-durable.gouv.fr/L-environnement-en-France-Edition,41611.html (consulté le 02/08/2015)

Face au changement climatique et à la dégradation de l'environnement, les politiques sont dans l'obligation de réagir et d'appliquer des mesures adéquates. Pour ce faire, la Chine s'intègre peu à peu dans les organisations mondiales, elle accède notamment à l'OMC (2001) et au G20 (2008). Les plans quinquennaux (Béduneau-Wang, Shan et al, 2010 ; Price, Levine et al, 2011 ; Alexeeva et Roche, 2014) fixent aussi des objectifs de réduction des émissions polluantes et des organismes d'état sont créés et chargés de mettre en application la politique environnementale[79].

Conclusion du chapitre

L'énergie et l'environnement sont des thèmes faisant aujourd'hui l'objet de nombreux débats passionnés. L'énergie est en effet l'un des moteurs de nos économies. Le choix d'un type d'énergie est stratégique pour tout décideur. Il apparaît donc que ce thème est particulièrement pertinent pour faire l'objet d'une veille textométrique et d'une étude d'intelligence économique. Les informations relevées grâce à un processus de veille textométrique et d'intelligence économique doivent aussi être replacées du point de vue des impacts environnementaux et sociaux. C'est la raison pour laquelle nous avons analysé la situation énergétique mondiale. Nous avons ainsi pu montrer que les besoins énergétiques vont s'accroître du fait de l'augmentation de la population mondiale. Cela nous a amené ensuite à nous intéresser aux différentes utilisations possibles de l'énergie. Actuellement les secteurs économiques utilisant le plus d'énergie sont le secteur tertiaire et le secteur des transports. Nous avons ensuite proposé le panorama des différents types d'énergie existants et plus particulièrement sur le nucléaire et ses impacts environnementaux. Le chapitre 3 sera consacré aux langues des trois pays.

[79] http://www.ccpit-france.org/Dossiers_speciaux_2.htm (consulté le 01/08/2015)

3. Trois sphères distinctes mais connectées

Une fois les thèmes cernés, il est nécessaire de mieux comprendre les langues des trois pays : France, Etats-Unis et Chine qui feront l'objet de notre étude. Dans cette perspective, nous examinerons d'abord les différences et les points communs de ces trois pays d'un point de vue géopolitique. Nous serons amenés à évoquer la situation de leur presse, et ainsi à justifier les choix de nos corpus. Nous verrons ensuite que ces langues ont des systèmes linguistiques bien distincts. Ces différences linguistiques se retrouvent aussi bien dans la phonologie que dans la syntaxe. En particulier, la langue chinoise se caractérise par un système graphique fruit d'une longue histoire et très différent du français et de l'anglais. Ce système complique ainsi le processus de segmentation de cette langue. Différents logiciels de segmentation du chinois existent cependant. Il s'agit de :

- *ICTCLAS*
- *Hylanda*
- *Jieba*
- *Stanford Word Segmenter*

Nous rappellerons les fondements algorithmiques et statistiques de ces outils. Nous les mettrons ensuite en œuvre sur des textes de nos corpus et nous en analyserons les résultats.

La France, les États-Unis et la Chine

La veille textométrique multilingue, objet de notre recherche, nous a amené bien évidemment à nous poser la question des langues à retenir, langues que nous avons limitées à trois.

Le choix devant être représentatif des systèmes politiques, des cultures et des économies de la planète, trois mondes ont ainsi été cernés, retenus comme essentiels, sans, pour autant, les considérer comme exclusifs.

Ils représentent en effet trois pôles de l'économie mondiale, trois politiques différentes avec chacune leurs zones d'influence et, trois positions face aux problèmes de production, de consommation et de gestion des pollutions qui en résultent.

Ces trois mondes sont :

- la France, citée en premier, le pays du français, la langue de ce travail,
- les Etats-Unis d'Amérique, première puissance mondiale, avec l'anglais, version américaine,
- la Chine, pays le plus peuplé, multilingue, dont le mandarin est la langue parlée par le plus grand nombre de locuteurs dans le monde.

Ces trois mondes s'imposent pour des raisons bien différentes, même si d'autres zones existent, comme celles du monde arabe, du monde russe, dont l'espace russophone couvre majoritairement le pays le plus étendu de la planète, et du monde hispanophone, langue latine comme le français, qui aurait pu être une autre option, nonobstant la langue du pays de rédaction de cette étude et le faible emploi de l'espagnol au sein des grandes organisations mondiales. L'Inde, de même, est un acteur essentiel par sa position démographique mais son développement est problématique au niveau de l'environnement. N'ayant pas de langue fédératrice elle s'appuie sur l'anglais, la langue de sa colonisation.

3.1 Les sphères de communication et les langues

Ce sont trois pays incontournables de la vie politique et économique mondiale, mais aussi des pays au rayonnement international par leurs histoires et leurs pensées philosophiques.

La France, co-fondatrice de l'Union Européenne, en est un des deux moteurs avec l'Allemagne. S'appuyant sur des conditions naturelles favorables, sur le choix du productivisme et sur un secteur agro-alimentaire puissant, la France est productrice et exportatrice des produits agricoles. Le développement industriel du pays prend son essor avec la révolution industrielle qui court tout au long du XIX^{e} siècle. La France dispense un message universaliste en matière de Droits de l'Homme depuis le Siècle des Lumières et la Révolution de 1789. Elle défend une politique volontariste en matière environnementale et organise en fin d'année 2015 la conférence COP21[80] sur le climat. En ce début de siècle, elle fait face à des problèmes de croissance et de chômage.

Les Etats-Unis ont été la puissance du XX^{e} siècle et sont les grands rivaux de la Chine pour le nouveau siècle. Ils ont déplacé le centre de gravité économique du monde, de l'Europe vers l'Amérique, puis de l'Atlantique vers le Pacifique. Leur puissance les a jusqu'à récemment écartés de la prise de conscience dans le domaine de l'environnement, considérant «*le niveau de vie des Américains comme non négociable*[81] ».

Quant à la Chine, une très ancienne puissance, dépositaire d'une culture millénaire, longtemps restée en autarcie, elle s'ouvre, de manière prodigieuse en ce début de XXI^{e} siècle, au progrès, à la participation et à la coopération internationale en adhérant à des organismes internationaux, comme par exemple, l'Organisation mondiale du commerce, fer de lance de ses exportations. Devenue « usine du monde », elle relève des défis sociaux et économiques considérables en régulant par exemple sa démographie et la répartition des profits de sa croissance et elle accélère son développement et son rayonnement national et international. Première importatrice de matières premières qu'elle transforme et exporte en grande partie, elle est à un tournant de son histoire en matière de croissance et de positionnement énergétique et environnemental.

3.2 Comparaison des sphères de communication

Elles portent, en ce qui concerne l'étude, sur deux grands aspects, d'une part socio-culturel avec les modes de vie qui en découlent, et d'autre part sur leur positionnement vis-à-vis de leurs besoins énergétiques et des conséquences sur l'environnement.

En France, le développement de la puissance économique des XIX^{e} et XX^{e} siècles, conforté par les Trente Glorieuses, s'essouffle ; les crises pétrolières successives, ainsi qu'une désindustrialisation progressive, la rendent moins concurrentielle et mettent son modèle social en danger dans un monde ouvert. Elle perd progressivement sa position de grande puissance face aux pays émergents et lutte pour maintenir son rang dans le but de maintenir son rayonnement et son influence dans le monde. Elle s'appuie pour ce faire, mais de plus en plus en concurrence sur :

- sa zone d'influence en Afrique,
- son rayonnement et son rôle au sein du Conseil de Sécurité aux Nations-Unies,
- la place de la langue française dans les organisations internationales et la francophonie,
- son domaine maritime : il est à la deuxième place mondiale derrière celui des États-Unis,
- l'adhésion de nombreux peuples à sa culture et à ses valeurs, comme les Droits de l'Homme.

[80] Conférence des Parties (COP21), conférence des Nations-Unies sur les changements climatiques

[81] « Le mode de vie des Américains n'est pas négociable », affirmait, en 1992 lors de la conférence de Rio (Dollfus, 1999 : 32), Georges Bush (père), président des Etats-Unis.

La France, souvent désireuse d'apporter un modèle aux autres pays, trouve sa place dans la politique environnementale et dans la lutte pour le développement durable. Elle prend conscience des enjeux énergétiques et revoit son « tout nucléaire », tout en essayant de préserver sa place industrielle dans la filière. Elle se montre volontariste quant aux énergies nouvelles et aux économies d'énergie, face à une Allemagne parfois contradictoire qui renonce au nucléaire au profit d'une filière charbon/gaz, très polluante.

Aux Etats-Unis, la superpuissance s'impose sans complexe, tant au niveau politique avec une place de quasi gendarme du monde, que militaire, avec les récents conflits, parfois mal gérés, comme la guerre du Golfe en 1991. Sur le plan économique, son modèle capitaliste, commercial et financier, s'impose progressivement sur toute la planète, en particulier dans les nouvelles technologies. La recherche et le développement placent les grandes sociétés de la « *Silicon Valley* » en position de maîtres à penser de la planète de demain. Sa monnaie étalon lui permet d'agir à peu près librement, malgré un fort déficit. Ses problèmes économiques et sociaux internes ne sont cependant pas résolus et sa place de pourvoyeur de « rêve américain » peut être parfois remise en question, en dépit de la fascination qu'exerce sa réussite dans les autres pays.

Plus gros consommateurs d'énergie par tête d'habitant, les Etats-Unis sont le mauvais élève qui a du mal à imposer une conduite aux pays émergents comme la Chine et l'Inde, demandeurs de droits à consommer et donc à polluer à leur tour. D'abord hostiles à toutes formes de contraintes, les États-Unis ont pris récemment conscience de la nécessité de contenir la menace que fait peser sur la Planète le changement climatique.

En Chine, l'ouverture et le développement très rapides constituent un exemple sans précédent d'accession au rôle de grande puissance mondiale. Ce cheminement bouscule la Planète et ses modèles économiques. En effet, dirigé par un parti unique, communiste, le pays encourage sa population à s'enrichir, sans renoncer à sa position de parti fort peu soucieux de démocratie, de transparence et de droits de l'Homme. Ce laboratoire de recherche et de développement remet en question les critères habituels et interpelle quant aux issues possibles au niveau du pays. Le positionnement de la Chine, sa capacité de gestion de sa démographie, de ses conflits régionaux, sociaux, culturels et de son évolution face aux enjeux, ne permettent donc pas d'imaginer la suite, ni sur le plan de l'ouverture à la démocratie et donc au multipartisme, ni aux risques de conflits internes en cas de difficultés à répartir la croissance.

En tout état de cause, la Chine est confrontée à des problèmes de pollution considérables, liés à son développement et à son rôle de producteur à bas coût, essentiel à la poursuite de son essor. Le pays ne pourra faire l'économie de réformes, mener une politique énergétique moins axée sur le charbon, tout en évitant le tout nucléaire.

Ses choix d'évolutions en matière de communication et de gestion des médias apparaissent donc comme essentiels.

3.3 Rôle de la Presse

L'étude des textes telle que décrite dans les méthodes traditionnelles antérieures à l'ère numérique ne peut s'affranchir de la connaissance des niveaux différents de transparence des informations publiées dans les textes analysés.

En France, on admet que l'existence de la presse date du début de l'imprimerie. La liberté de la presse a varié au gré des régimes politiques : niée sous l'Ancien Régime, affirmée dans l'article 11 de la Déclaration des Droits de l'Homme de 1789, adoptée enfin avec la loi du 29 juillet 1881. Après la période noire de l'occupation allemande, les ordonnances de 1944 réaffirment la liberté et le pluralisme de la presse. Actuellement, les médias et la presse relatent l'information selon l'Etat de droit, de la variation des opinions et des organes de presse.

Aux Etats-Unis, la liberté de la presse figure dans les revendications des pères de la révolution du XVIII^{e} siècle : les journalistes sont nombreux à jouer un rôle politique, tel Benjamin Franklin. Sa liberté est garantie par le premier amendement de la Constitution, avec certaines restrictions, telles l'incitation à l'émeute, l'accès aux documents classifiés, la position durant les conflits, l'attaque du 11 septembre (écoutes), *USA Patriot Act*[82] (une loi antiterroriste), etc.

En Chine, l'ensemble des actions de communication est géré par la propagande de la RPC, le sens du mot n'ayant pas le même caractère péjoratif qu'en français. Cela permet la mise en application de la ligne politique du Parti communiste au gré des circonstances. La censure peut alors être appliquée, afin d'influencer la population et de faire évoluer son opinion.

Peu habituée à la libre expression, la population commence à ressentir l'impérieuse nécessité d'une information libre et transparente : il existe en particulier une demande d'utilisation non contrôlée d'internet et des réseaux sociaux.

Les principaux journaux

En France, les quotidiens les plus connus sont Le Monde, Le Figaro ou encore Libération. Pour notre corpus comparable, nous choisissons le quotidien Le Monde. Ce quotidien a été créé en 1944 par Hubert Beuve-Méry. Le général De Gaulle souhaitait doter la France d'un quotidien de référence ouvert sur le monde, comme son titre l'indique. Le Monde est un journal généraliste. Même si le journal ne le reconnaît pas lui-même, il est habituellement considéré que le journal adopte une ligne éditoriale de centre-gauche. Il traite ainsi d'actualités politiques, économiques et internationales. En dépit des difficultés et des critiques comme celles de Péan et Cohen, figurant entre autres dans le livre « La face cachée du Monde » (Péan et Cohen, 2003), Le Monde demeure un quotidien de référence en France, d'où notre choix. Il classe ses articles à la fois par thème : politique, économie, sciences, sport…, mais aussi géographiquement avec la catégorie «International». Il comporte aussi, comme tout journal, sa Une, qui contient les articles jugés les plus importants par la rédaction de ce journal.

Aux Etats-Unis, l'un des quotidiens les plus connus est le *New York Times*. Il a été créé en 1851 par Henry Jarvis Raymond et George Jones[83]. Tout comme Le Monde, le *New York Times* est un quotidien généraliste. Sa préférence politique va très souvent au Parti Démocrate[84]. Il soutient le candidat représentant le Parti Démocrate aux élections présidentielles de 1988, 1992, 1996, 2000, 2004 et 2008.

[82] « *Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001* » (Loi pour unir et renforcer l'Amérique en fournissant les outils appropriés pour déceler et contrer le terrorisme) http://www.selectagents.gov/resources/USApatriotAct.pdf (consulté le 15/06/2016)

[83] Henry Jarvis Raymond, journaliste et homme politique et George Jones, banquier, ont fondé en 1851 le New York Daily Times, devenu en 1857 le *New York Times*.

[84] http://www.nytimes.com/inte:ractive/2008/10/23/opinion/20081024-endorse.html (consulté le 15/06/2015).

Sa large diffusion et son positionnement politique, voisin de celui du journal Le Monde, en font un choix pertinent pour notre corpus comparable de veille textométrique. Le *New York Times* est organisé en trois sections principales :

- la section «nouvelles» consacrée aux actualités nationales et internationales, nouvelles classées selon leur thème ou le lieu de leur provenance. Ainsi, la section « nouvelles » comprend les sous-sections : international, politique, santé, etc.,
- une section consacrée aux « opinions », comprenant l'éditorial et le courrier des lecteurs,
- une section intitulée «suppléments» correspond aux loisirs, aux arts, à la littérature et au cinéma.

En Chine, le principal journal est le Quotidien du Peuple dont la première édition a été publiée en 1948. Il s'agit du journal officiel du Parti Communiste Chinois, permettant ainsi de connaître la ligne idéologique officielle de ce parti sur les différentes problématiques d'actualité. A partir des sites *qq.com* et *sina.com*, nous avons extrait des articles de ce journal. Le site *qq.com* comprend aussi une catégorie contenant les articles placés en «Une» du site. Les articles issus de la presse chinoise sont ensuite classés géographiquement (une rubrique pour les informations nationales et une rubrique pour les informations internationales) et thématiquement (une rubrique est par exemple consacrée aux articles concernant les faits militaires). Le site *sina.com* qui contient une version anglaise adopte un classement similaire. Un dernier site d'actualité cette fois-ci centré sur l'environnement : *chinadialogue.net* a aussi été utilisé pour notre corpus parallèle. Les articles de ce site sont classés dans les catégories suivantes : commerce, catastrophes naturelles, villes, changement climatique, protection de l'environnement, législation écologiste, nourriture et santé, pollution et gestion de l'eau.

L'approche du support linguistique de ces trois entités majeures dans le monde conditionne les bases de notre étude comparative.

3.4 Les trois langues du corpus

Le français, l'anglais et le chinois sont des langues très présentes dans le monde. Néanmoins, elles sont fort différentes, l'anglais et le français sont deux langues indo-européennes classées parmi les langues germaniques et romanes, quant au chinois, elle est classée dans la famille des langues sino-tibétaines, ce qui cause des difficultés dans l'analyse des textes appréhendés par la veille textométrique.

Le français[85] est une langue très parlée dans le monde (quelques 274[86] Millions de locuteurs dispersés dans de nombreux pays), et en particulier, en Europe, en Afrique et au Québec. Ces locuteurs se regroupent dans la Francophonie. Certains pays, comme l'Algérie, avec près de 40 millions d'habitants, quoique non adhérent à la Francophonie, l'emploient de manière courante, autant que leur langue nationale. Langue de Molière, de Balzac, de Victor Hugo, des « Belles Lettres » en somme et du Siècle des Lumières, elle fut la langue des cours européennes durant des siècles, y compris l'Angleterre. Elle dispose d'un statut de langue internationale et sert jusqu'à présent de langue de travail dans les instances internationales telles qu'ONU, OCDE, OMC, OIT, Comité olympique international, etc.

L'anglais est l'une des langues les plus parlées au monde aux côtés du chinois et de l'espagnol. Elle est en outre la plus enseignée universellement dans les écoles et dans les cursus économiques. Elle est devenue une sorte d' « esperanto des affaires », langue d'usage international la plus répandue dans les domaines du commerce, de la diplomatie et de la politique, mais aussi dans les publications scientifiques et dans les activités du transport international.

Le chinois (ou mandarin) est la langue la plus parlée au monde dans le pays le plus peuplé du monde avec la diaspora chinoise. Sa diffusion à des locuteurs non chinois, malgré les efforts du gouvernement chinois (création des centres culturels Confucius en Afrique, télévisions chinoises par satellite CCTV...), est rendue plus difficile du fait de sa complexité. En effet, le chinois comprend 50 000 caractères. L'existence de quatre tons dans les différents mots est également source de difficultés pour un apprenant du chinois. De plus, la séparation de l'écrit et de l'oral impliquent la mémorisation de l'image (idéogramme) et d'un son ou vice-versa. Pour toutes ces raisons, l'analyse textométrique du chinois se révèle difficile.

L'étude comparative de ces trois langues fait ressortir la complexité de leur approche simultanée. Celle-ci provient tout autant des différences culturelles, historiques que des principes morphologiques et syntaxiques qu'elles mettent en œuvre.

Ces trois langues se distinguent fondamentalement par leur système d'écriture, alphabétique et idéographique, par leur prononciation monosyllabique et/ou multi-syllabique et par leurs tons. Elles appartiennent au groupe des langues sujet-verbe-objet (SVO), cet ordre représente environ 42%[87] des langues parlées dans le monde. Cependant, le sujet est généralement en tête de phrase en français et en anglais, alors qu'en chinois c'est le prédicat[88] qui prédomine. Par exemple, la phrase *il*[89] *pleut*, peut se traduire en anglais par *it's raining*, tandis qu'en chinois, on dit, 下雨了/xià yǔ le (littéralement, tomber pluie). Dans cet exemple, la locution chinoise du langage courant n'a pas besoin de sujet. Un autre exemple, dans le langage courant et familier, pour demander à quelqu'un s'il a mangé ou s'il est sorti, il suffit de dire, 吃了吗/chī le ma?, 去了吗/qù le ma?.

[85] http://www.diplomatie.gouv.fr/fr/politique-etrangere-de-la-france/francophonie-et-langue-francaise/
[86] http://www.francophonie.org/-Qu-est-ce-que-la-Francophonie-.html
[87] Russell Tomlin, *Basic Word Order: Functional Principles*, Croom Helm, London, 1986, p. 22.
[88] Définition : Dans un énoncé où l'on peut distinguer ce dont on parle et ce qu'on en affirme ou nie. Prédicat (logique). Terme qui dit quelque chose de l'autre. http://www.cnrtl.fr/definition/Pr%C3%A9dicat
[89] Dans cet exemple, « il » n'est pas un sujet réel, mais un pronom impersonnel.

Un dernier exemple qui relève du langage écrit, 妈妈一回到家，(她)就去厨房了。/ māmā yī huí dàojiā, (tā) jiù qù chúfáng le. / Maman est de retour à la maison, elle est allée directement à la cuisine, le sujet (她/tā/elle) n'est pas répété dans la deuxième partie de la phrase.

Dans ces trois cas, la phrase ou une partie de phrase ne possède pas de sujet, mais celui-ci est sous-entendu. C'est l'interlocuteur de la conversation qui est alors le sujet. C'est la situation de communication qui détermine les référents et la deixis.

Dans les sections qui suivent, nous allons commencer par une présentation de la langue chinoise, puis viendront les spécificités du français et de l'anglais par rapport au chinois. Nous terminerons par une section sur la segmentation de la langue chinoise.

3.5 Langue chinoise, une *scriptio continua*

3.5.1 Les spécificités de la langue chinoise

Avant d'étudier la textométrie du chinois, il est indispensable de se pencher sur les spécificités de cette langue, fort éloignée des langues indo-européennes. Ces particularités devront être prises en compte par tout logiciel de segmentation.

L'une de ses grandes caractéristiques est l'écriture sans espace, également appelée *scriptio continua*, *scriptura continua* ou *scripta continua*. Cette pratique ancienne existait aussi dans l'antiquité grecque et romaine ainsi que dans la France médiévale. Cependant, le locuteur européen repère d'abord les syllabes, puis les mots. Tandis qu'en chinois, de manière générale, chaque idéogramme est déjà associé à une seule syllabe, il ne reste qu'à identifier les mots et séquences de mots par groupement des caractères.

Dans la présente section, nous partirons de l'étude des caractères chinois, composants de base de tout texte chinois, pour ensuite traiter les notions de mots et de phrases.

Histoire du système d'écriture chinois

La langue chinoise est issue de la famille des langues sino-tibétaines. Elle se distingue ainsi des langues indo-européennes. Elle comprend actuellement de nombreux dialectes dont l'un des plus connus est le cantonais, la langue officielle étant le mandarin. Son système d'écriture est celui des caractères chinois[90].

Les inscriptions de caractères chinois les plus anciens ont, en fait, été trouvées sur des os oraculaires, c'est à dire des morceaux d'os ou de carapaces de tortue, durant la dynastie des Shang entre le XVII^{e} et le XI^{e} siècle av. J-C.

Les premiers caractères chinois sont des pictogrammes. Ils ont une forme très proche de l'entité qu'ils sont censés représenter[91]. Les caractères de l'époque ont alors une forme très différente selon le lieu ou encore le support d'écriture (carapace de tortue ou omoplate de bovin). Plusieurs facteurs seraient à l'origine de cette situation : la distance, et la moindre importance des échanges écrits. Une unification de l'écriture est apparue sous le règne de Qin Shi Huang Di (221 av. J-C.). Li Si publie alors en l'an 213 un index des caractères : le *Sancang* qui contient environ 3300 caractères. Par la suite, l'évolution

[90] La création des caractères chinois est attribuée au ministre CANG Jie du mythique empereur jaune. Celui-ci se serait inspiré des empreintes d'oiseau. Un autre livre le « *Daodejing* » affirme que les caractères chinois ont été créés par Laozi (ou Lao Tseu ou Lao Zi, un sage chinois, considéré a posteriori comme le père fondateur du taoïsme).
[91] Par exemple, le caractère « se reposer » représente un homme et un arbre, afin de suggérer l'idée d'un homme s'allongeant à côté d'un arbre.

des supports d'écriture va entraîner un enrichissement des styles d'écriture. On distingue ainsi sous les Qin huit styles d'écriture : 大篆/dà zhuàn/le grand sceau, 小篆/xiǎo zhuàn/le petit sceau, 刻符/kè fú/l'écriture lapidaire, 虫书/chóng shū/l'écriture insecte, 摹印/mó yìn/l'écriture pour imprimer les sceaux, 署书/shǔ shū/l'écriture pour faire des titres, 殳书/shū shū/l'écriture en forme de lance, 隶书/lì shū/le style clérical.

Entre la dynastie Han et la dynastie Jin, les Han[92] voient aussi apparaître l'écriture cursive, une des formes de la calligraphie chinoise. Elle se compose généralement de trois types utilisés dans l'art de l'écriture à savoir, 章草/zhāngcǎo/la cursive des scribes, 今草/jīncǎo/la cursive dite moderne et 狂草/kuángcǎo/la cursive dite sauvage. Il est à noter que 行书/xíngshū/le style courant s'ajoute à cet inventaire et s'emploie toujours dans l'écriture quotidienne.

Structure graphique des caractères

Les caractères chinois sont composés d'une trentaine de types de trait. Les plus courants sont le trait vertical, le trait horizontal, le point, le trait jeté descendant de droite à gauche, le trait appuyé descendant de gauche à droite, le trait relevé, et le crochet. Le tracé des caractères se fait, trait par trait, selon un ordre défini. Chaque caractère doit être inscrit dans un carré. Il doit y avoir par ailleurs le même espace entre chaque caractère. Les caractères sont composés de différents éléments phonétiques et sémantiques. L'un des éléments sémantiques du caractère est appelé la clé, une notion apparue dans le *Dictionnaire analytique des caractères chinois* conçu par Xu Shen (58 – 147). La clé permet de classer les différents caractères dans un dictionnaire, un second critère de classement étant le nombre de traits. Le dictionnaire Kangxi, compilé en 1716 sur l'ordre de l'empereur du même nom, recense 214 clés.

Pinyin

Au début du XX^{e} siècle, 胡适/hú shì et 鲁迅/lǔ xùn, participant à la revue «新青年/xīn qīng nián/Nouvelle Jeunesse» parue de 1915 à 1926, ont entamé des travaux de simplification des caractères chinois. Dans les années 1950-1960, la Commission de Réforme de l'écriture s'est montrée favorable à une réforme de l'écriture chinoise. Cette réforme s'est en particulier traduite par une phonétisation de l'écriture chinoise au moyen du pinyin dont l'adoption s'est réalisée en 1958 avec le 汉语拼音/Hànyǔ pīnyīn. Cinq tons ont aussi été utilisés. Le but était alors d'adopter une langue unique pour les différentes régions de la Chine. Ces travaux ont été rendus difficiles pour plusieurs facteurs. Un premier obstacle est l'existence de caractères homophones. Un second obstacle réside dans la diversité des dialectes en Chine. Il est à noter que d'autres systèmes avaient été auparavant adoptés. Ainsi, le missionnaire jésuite Matteo Ricci a, au XVI^{e} siècle, tenté d'effectuer une première romanisation de la langue chinoise. Par la suite, d'autres systèmes tels que celui de Wade-Giles[93], ont été mis en place au XIX^{e} siècle. Son système se différencie du pinyin par plusieurs aspects. L'un d'entre eux réside dans l'utilisation systématique d'un trait d'union pour séparer les syllabes. En 1918, le 注音拼音/Zhùyīn pīnyīn, ou appelé encore le *bopomofo*, a été adopté, système encore en vigueur à Taiwan. Il consiste en un alphabet de quarante lettres et comprend cinq tons.

[92] Les Han (chinois simplifié : 汉 ; chinois traditionnel : 漢 ; pinyin : hàn) constituent l'ethnie majoritaire en Chine. Ils sont issus de l'ancienne ethnie Huaxia et prennent le nom Han à l'époque de la dynastie Han (206 av. J-C à 220 ap. J-C).
[93] Parfois abrégé en Wade, une romanisation du chinois mandarin, système créé par Thomas Wade au milieu du XIX^{e} siècle et modifié par Herbert Giles, http://global.britannica.com/topic/Wade-Giles-romanization (consulté, le 01/02/2015).

Simplification des caractères

Une simplification des caractères chinois a aussi été mise en œuvre. Le but du pouvoir communiste chinois était alors de lutter contre l'analphabétisme. Les caractères chinois de l'époque étaient perçus comme inutilement complexes et restreignaient l'accès à la lecture d'une grande partie de la population. La rapidité du tracé est aussi recherchée. Le but est alors de réduire le nombre de traits utilisés pour écrire un caractère[94]. Afin de simplifier les idéogrammes sans rompre totalement avec l'histoire du système graphique chinois, plusieurs méthodes et règles sont employées. Par exemple, on a conservé des formes anciennes telles que la clé du nuage 云/yún/nuage, comme écrite dans le caractère 魂 /hún, âme, qui est formé de la clé du nuage (à gauche) et du mot fantôme (à droite). Autre exemple, la forme cursive de la clé de la parole dans les calligraphies est devenue officiellement son écriture simplifiée. La forme 言/yán/parole ou clé de la parole, est devenue 讠/clé de la parole. Nous pouvons, entre autres, la constater dans le caractère 讀/dú/lire et sa forme simplifiée 读/dú/lire.

Différents types de caractères

Il existe aussi différents types de caractères chinois :

- les pictogrammes correspondant à une imitation de l'objet qu'ils sont censés représenter. Un exemple peut être donné avec le caractère signifiant arbre. Celui-ci a la forme d'un arbre, 木 /mù/arbre (),
- les idéogrammes simples représentant une idée abstraite,
- les idéogrammes composés résultant de l'association de deux idéogrammes simples. Un nouveau sens est créé à partir de cette association. Par exemple, le caractère 好/hǎo/bon résulte de l'association du caractère signifiant femme et du caractère signifiant enfant,
- les idéo-phonogrammes représentent environ 90% des caractères en Chine. Ils sont constitués de deux parties : une partie phonétique et une partie sémantique. Par exemple, le caractère 认 /rèn/connaître ou reconnaître, comprend une partie sémantique qui est la clé de la parole, 讠 /clé de la parole, et une partie phonétique : 人/rén/l'homme.

[94] Les 514 caractères simplifiés présentés par la Commission de Réforme de l'Ecriture comportent alors en moyenne 8 traits au lieu de 16 pour les anciens caractères.

3.5.2 La formation des mots ou des idiotismes

Les caractères chinois peuvent être associés pour produire de nouveaux sens[95]. Les mots constitués de deux syllabes forment dans la langue chinoise moderne la majeure partie du lexique, à la différence du chinois classique essentiellement composé de mots monosyllabiques[96]. La segmentation et l'identification du texte reposent fortement sur la compétence du lecteur chinois.

Des combinaisons de caractères chinois peuvent aussi être employées pour leur aspect phonétique afin de retranscrire les mots étrangers, par exemple dans le cas des noms de pays tels :

印度/yìn dù/Inde

法国/fǎ guó/France

Les caractères employés ici ne le sont pas en raison de leur sens, mais en raison de leur prononciation.

Le chinois comprend aussi des combinaisons de quatre caractères appelées les 成语/chéng yǔ. Ces combinaisons correspondent souvent à des idiotismes[97]. Ils font partie de la langue écrite 文言/wényán utilisée de l'Antiquité chinoise jusqu'à 1919. Par exemple, nous comprenons que le 成语/chéngyǔ : 好声好气/hǎoshēng hǎoqì signifiant «être aimable et courtois lors d'un échange de paroles» ne produit un sens que dans sa globalité. Chacun des caractères : 好/hǎo/bon, 声/shēng/la voix et 气/qì/l'air pris isolément ne permettent pas de déterminer le sens de l'expression. Les quatre caractères des 成语/chéngyǔ/idiotisme ne respectent pas forcément la structure syntaxique du chinois moderne. Certains sont en effet tirés du chinois classique. Par ailleurs, le nombre restreint de caractères des 成语/chéngyǔ/idiotisme nécessite de s'affranchir des règles grammaticales pour exprimer l'idée principale derrière ces quatre caractères.

Les caractères chinois sont des morphèmes

Un morphème se définit comme la plus petite unité significative dans une langue. La notion de morphème se distingue de celle du mot qui peut contenir plusieurs morphèmes[98].

Dans le cas de la langue chinoise, le caractère chinois correspond à l'unité significative minimale. Chaque caractère pris isolément a en effet un sens, à l'exception de cas tels que 葡萄/pú táo/raisin, 蝴蝶/hú dié/papillon, etc. Comme nous l'avons vu plus haut, le sens du caractère peut cependant varier lorsqu'il est combiné à d'autres caractères. Les linguistes chinois considèrent en général que les morphèmes chinois sont monosyllabiques et sont des caractères. On distingue deux types de morphèmes chinois :

[95] Par exemple, le mot 中国/zhōng guó/Chine est formé de deux caractères de sens différents : 中/zhōng/milieu et 国/guó/pays, qui associés donnent le mot Chine.

[96] Certains mots composés de plusieurs caractères changent de sens si on leur retire un caractère. Par exemple, le mot 法国/fǎ guó/France devient le mot *la loi*, si on lui retire le caractère 国/guó/pays. D'autres mots tels que le mot 花儿/huā er/fleur, a le même sens que 花/hua/fleur. Les mots chinois peuvent aussi comprendre trois syllabes, comme par exemple 飞机票/fēi jī piào/billet d'avion.

[97] Définition : Construction qui apparaît propre à une langue donnée et qui ne possède aucun correspondant syntaxique dans une autre langue. http://www.cnrtl.fr/definition/idiotisme (consulté le 08/01/2016)

[98] Ainsi, dans le mot lentement, nous relevons deux morphèmes : lent- : un lexème c'est à dire une unité de sens dans un mot d'une phrase. Dans les langues flexionnelles, les mots chant et chants dérivent du même lexème. -ment : la désinence de l'adverbe en français.

- les morphèmes libres, qui ne perdent pas leur sens initial lorsqu'ils sont associés à un autre mot. Tel est le cas du morphème 完/wán/finir qui ne perd pas son sens lorsqu'il est associé au morphème 全/quán/tout pour former le mot 完全/wán quán/complètement.
- les morphèmes liés, qui une fois qu'ils sont rattachés à d'autres morphèmes perdent leur sens initial. Il existe beaucoup de morphèmes liés dans les mots composés. Par exemple, le mot 老乡/lǎo xiāng/une personne du même village ou parfois concitoyen, contient deux morphèmes :
    - 老/lǎo : vieux
    - 乡/xiāng : village ou parfois ville

Ces deux morphèmes perdent leur sens une fois combinés et constituent à ce titre des morphèmes associés.

3.5.3 La notion de mot

Le mot est une notion peu aisée à définir quelles que soient les langues. Le TLFi donne la définition suivante du mot :

> *« Son ou groupe de sons articulés ou figurés graphiquement, constituant une unité porteuse de signification à laquelle est liée, dans une langue donnée, une représentation d'un être, d'un objet, d'un concept »*[99]

Cette définition ne permet pas de voir en quoi un mot se distingue d'un morphème. Di Sciullo et Williams [100] définissent le mot comme l'unité syntaxique minimale ayant un sens. En ce qui concerne le chinois, plusieurs problèmes se posent dans la mesure où un caractère isolé tel que 中/zhōng/milieu peut former un mot. Néanmoins, comme nous l'avons vu plus haut plusieurs caractères peuvent former un mot. L'article *The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank 3.0* de Fei Xia retient plusieurs critères pour définir si deux morphèmes chinois (ou caractères chinois) forment un mot :

- Les deux morphèmes forment-ils un mot une fois réunis ?
- Fréquence de la combinaison de ces deux morphèmes.
- Possibilité d'insérer un morphème entre les deux morphèmes en question.

De manière générale, les mots chinois ne connaissent ni conjugaison, ni déclinaison, ni temps, ni désinence, ni genre, ni nombre, mis à part certaines exceptions utilisées dans des circonstances précises pour les nombres[101] seulement, car le chinois est une langue isolante.

Rappelons qu'une langue isolante se définit comme une langue où les mots sont invariables. Le chinois et le vietnamien sont des langues isolantes. Les langues isolantes s'opposent aux langues flexionnelles. Dans ce type de langue, les mots sont composés d'une racine, à laquelle il est ajouté un ou plusieurs morphèmes dont le but peut être d'expliciter le temps, le genre ou le nombre. Dans ce type de langue, les mots changent de forme pour des motifs grammaticaux. Par exemple, en français, le mot cheval devient au pluriel *chevaux*.

[99] http://www.cnrtl.fr/lexicographie/mot
[100] Anna-Maria Di Sciullo and Edwin Williams, *On the definition of word* (Linguistic Inquiry Monographs 14). Cambridge, Massachusetts: MIT Press, 1987. Pp. x + 118.
[101] Le caractère 们/men est souvent considéré comme un marqueur de pluriel en chinois. Par exemple, on a 我们/wǒ men /nous, 你们/nǐ men/vous (deuxième personne du pluriel), 学生们/xué shēng men/les étudiants. Or, lorsqu'un nombre s'ajoute devant les noms en pluriel, le marqueur 们/men disparait, comme 三个学生/sān gè xuéshēng/trois étudiants.

Les mots chinois ne connaissent ni déclinaison, ni genre, ni nombre. Par exemple le mot 学生/xué shēng/étudiant, peut aussi bien désigner un étudiant, une étudiante, des étudiant(e)s. L'ajout du caractère 们/men après le mot 学生/xué shēng, constitue une marque de pluriel. Les verbes chinois n'ont pas de conjugaison. Par exemple, le verbe 有/yǒu/avoir, peut aussi bien signifier j'ai, tu as, il a, nous avons, vous avez, ils ont, que j'aurai, tu auras, il aura, nous aurons, vous aurez, ils auront.

D'une manière générale, les adjectifs chinois peuvent être repérés par l'ajout du caractère 的/de, lorsqu'ils sont en position d'épithète comme par exemple dans le groupe de mots : 漂亮的衣服/piào liàng de yī fú/un beau vêtement. Le caractère 的/de, est ici employé pour donner une fonction d'épithète à l'adjectif 漂亮/piào liàng/beau. Cependant, dans d'autres constructions grammaticales, les adjectifs n'ont pas de désinence spécifique.

Par exemple, dans la phrase

他有三只小狗/tā yǒu sān zhi xiǎo gǒu/il a trois petits chiens,

le caractère 小/xiǎo/petit, n'a pas de désinence spécifique. Les adverbes chinois ne possèdent pas non plus de désinence.

Par exemple, dans la phrase :

中国人都喝茶/zhōng guó rén dōu hē chá/les chinois boivent tous du thé,

l'adverbe 都/dōu/tous, ne contient aucune désinence spécifique. Il convient de noter que l'ajout du caractère 地/de, permet de construire un adverbe à partir d'un adjectif.

Ainsi, dans la phrase

他慢慢地跑步/tā màn màn de pǎo bù/il court doucement/lentement,

le caractère 地/de, permet de construire un adverbe à partir de l'adjectif 慢/màn/lent. Il est également à noter que, dans la pratique actuelle non formelle du chinois, certaines personnes abusent de l'emploi de la préposition 的/de, à la place de 地/de, une particule grammaticale strictement réservée aux adverbes.

3.5.4 Les mots-outils

Mots-outils en chinois pour marquer le genre, le temps ou le nombre

Un mot-outil ou mot grammatical est un type de mot dont l'utilisation est davantage syntaxique que sémantique. Les mots chinois étant, comme nous l'avons dit précédemment, invariables, leur utilisation requiert une importance particulière pour marquer le genre, le nombre ou la personne. Appartiennent à la catégorie des mots outils selon Fabienne Marc[102]:

- les adverbes (pour la plupart des cas),
- les prépositions,
- les conjonctions,
- les particules.

Quelques exemples d'emploi de particules :

- 了/le pour indiquer un état accompli, le contexte de la phrase pouvant aider à trouver le temps d'emploi du verbe. Par exemple, la particule 了/le dans la phrase /我吃了饭/wǒ chī le fàn/J'ai mangé/, permet de comprendre que le verbe 吃/chī/manger est ici à traduire au passé.
- 过/guò pour marquer l'expérience vécue, comme par exemple dans la phrase /我去过中国/wǒ qù guò zhōngguó/Je suis allé en Chine/.

Ces particules seront très utiles pour un logiciel de segmentation, dans la mesure où elles permettent de repérer le groupe verbal. Il faut savoir également que la particule 了/le n'est pas toujours collée au verbe en fin de phrase.

3.5.5 La notion de phrase

Le TLFi définit la notion de phrase comme suit : «*Tout assemblage de mots : expression, locution, tour figé ou non* [formant un sens complet]*»*.

Cette définition retient une définition large de la phrase. Cependant, elle ne permet pas de segmenter un texte en phrases. En effet, si tout assemblage de mots constitue une phrase comme le texte ci-après : « Je suis allé à la boulangerie acheter du pain. Je suis ensuite rentré chez moi ». Les segmentations possibles de la phrase ci-dessus sont les suivantes :

- Je suis
- suis allé à
- Je suis allé à la boulangerie acheter du pain.
- Je suis ensuite rentré chez moi
- Je suis allé à la boulangerie acheter du pain. Je suis ensuite rentré chez moi
- etc ....

Il convient aussi de noter que cette définition évince l'aspect sémantique de la phrase.

Un moyen usuel de reconnaître une phrase en français est la ponctuation et l'usage de majuscules. Ainsi, il est connu qu'une phrase commence par une majuscule et se termine par un point. L'usage de

[102] Fabienne Marc est maître de conférences en langue et linguistique chinoises à l'INaLCO Paris.

la ponctuation est relativement récent en chinois et remonte en fait à 1919[103]. Les signes actuellement utilisés sont :

- les points finaux 。?!
- les virgules , 、 ； ： °§
- les points de suspension ……
- les guillemets “ ” ‘ ’ ﹏﹃ ﹂﹁ 《 》 〈 〉
- les parenthèses （ ） [] 〔 〕 ｛ ｝ 【 】
- les tirets — —— ～
- les noms propres, les entités nommées, etc. introduits par les signes tels que -- ~~
- les points de soulignement.

La ponctuation sera un bon outil pour délimiter les différents segments d'une phrase.

3.5.6 L’ordre des mots dans la phrase

Dans la phrase chinoise, l'ordre des mots est : sujet verbe complément d'objet (SVO). Les adverbes et compléments circonstanciels sont en principe situés juste avant le verbe. Par exemple, dans la phrase /我们明天上课/wǒ mén míng tiān shàng kè/on a cours demain/, l'ordre des mots usuel est bien respecté avec comme sujet /我们/wǒ mén/nous/, comme complément circonstanciel de temps /明天/míng tiān/demain/ et comme verbe /上课/shàng kè/aller en cours/.

Cet ordre des mots n'a pas toujours existé selon certains auteurs (Li et Thompson, 1974). Ainsi le chinois classique aurait connu une structure du type Sujet Objet Verbe pour devenir ensuite une langue avec une structure sujet verbe objet entre le X^{e} et le III^{e} siècle av. J-C. Depuis le III^{e} siècle av. J-C, un retour à une structure sujet verbe objet serait en cours. Cette position est critiquée par plusieurs auteurs (Zhang, 1989 ; La Polla, 1990), qui reconnaissent, certes, que le chinois archaïque a connu une structure sujet-objet-verbe, mais qui réfutent l'idée d'un passage d'une structure du chinois moderne vers une structure sujet-verbe-objet[104].

Certaines phrases chinoises se terminent par un verbe. C'est le cas des structures en 把 Ba très utilisées en chinois: Sujet + Ba + Objet + Verbe.

Un exemple d'utilisation d'une telle structure est la phrase : 他把人打了/tā bǎ rén dǎ le/Il a frappé quelqu'un ou Il l’a frappé(e).

Ce type de structure permet de marquer une insistance sur l'objet. Il est à noter que dans le but d’enrichir les analyses d’informations, les relations couplées entre thème *versus* propos, mais aussi structure surface *versus* structure profonde demeurent également intéressantes.

[103] Monsieur HU Shi, philosophe et écrivain chinois a publié en 1919, le premier ouvrage 中国哲学史大纲/zhōngguó zhéxué shǐ dàgāng/*An outline history of Chinese philosophy*, ouvrage utilisant la ponctuation moderne.
[104] http://www.persee.fr/web/revues/home/prescript/article/crai_0065-0536_1997_num_141_2_15757

3.5.7 Le codage des caractères

Il s'agit d'un domaine en plein essor (UTF-8, UTF-16, etc.), mais qui s'est heurté à plusieurs obstacles. L'un des premiers obstacles réside dans la nature des caractères chinois, notamment dans la manière de les saisir et de les enregistrer. Etant donné le nombre important de caractères (environ 50 000), la réalisation d'un clavier géant contenant l'ensemble des caractères chinois paraît clairement inenvisageable.

Chaque texte chinois est représenté informatiquement comme une chaîne de caractères. Avant de commencer une opération de segmentation d'un texte chinois, il convient donc en premier lieu de s'assurer que l'ordinateur sera capable de reconnaître les caractères chinois. Pour les caractères anglo-saxons, le système ASCII, utilisé à partir de la fin des années 1960, comprenant 128 signes se révèle insuffisant pour la langue chinoise qui contient environ 50 000 caractères. Face à ce problème, le système Unicode (UTF 8) a été adopté. Il est actuellement capable de stocker un nombre important de caractères appartenant à diverses langues. En chinois, il consiste à attribuer un code à chaque caractère. Il convient de noter que le code sera identique pour ce même caractère, même s'il est utilisé dans une autre langue. A chaque caractère est aussi associé un ensemble de dessins, tracés, images possibles (en italique, en gras par exemple). Ces représentations sont appelées glyphes.

Des normes de codage des caractères chinois propres à la Chine ont aussi été créées. En 1980, la norme GB2312 a ainsi été adoptée. Elle contient entre autres 6 763 caractères simplifiés, ainsi que les Pinyin avec les quatre tons. En 1995, une extension nommée GBK et contenant 20 914 caractères chinois a été développée. En 2000, une nouvelle mise à jour de la norme GBK : la norme GB18030 a été adoptée par le comité de standardisation technique chinois. Elle reprend le codage des normes précédentes et s'assimile en fait à une version chinoise d'UTF-8. Les langues des minorités ethniques, Mongole, Ouïgoure, et Tibétaine sont aussi pris en charge. Cette norme[105] est actuellement incluse dans les systèmes d'exploitation vendus en Chine.

En ce qui concerne les caractères non simplifiés en vigueur à Taiwan et Hong Kong, la norme Big 5 a été mise en œuvre en 1984. Elle inclut 13 051 caractères chinois. Elle a connu deux extensions :

- Big five Plus adoptée en 1997 et contenant 20 914 caractères chinois.
- Big Five extension adoptée en 1998 et utilisée par Mac Os X et ajoutant 3 954 caractères chinois à Big Five Plus.

Hong Kong a également développé une extension à Big Five en 1995. Celle-ci s'intitule GCCS (*Government Common Character Set*). Différentes mises à jour, menées en 1999 (un nouveau nom de cette extension, HKSCS est apparu cette année-là), 2001, 2004 et 2008 ont permis d'ajouter 4 568 caractères traditionnels chinois à la liste de caractères chinois actuellement gérés par Big Five. Ce système est aussi géré par le système d'exploitation Mac Os X.

3.5.8 La saisie des caractères

Une fois les caractères encodés, il est nécessaire que l'utilisateur puisse saisir le texte à segmenter. Il existe actuellement plusieurs méthodes de saisie des caractères chinois à partir d'un clavier AZERTY ou QWERTY. Différents systèmes de saisie des caractères chinois ont auparavant existé, mais ils n'ont pas donné satisfaction. Nous distinguons essentiellement deux types de méthode de saisie des caractères : les méthodes fondées sur l'étymologie des caractères (telles que Canjie, Dayi et Wubi) et les méthodes fondées sur les Pinyin.

[105] http://www.iana.org/assignments/charset-reg/GB18030

En 1976, Chung Bong-Foo a mis en œuvre la méthode de saisie Canjie[106]. Par la suite en 1988, il a été mis en œuvre la méthode Dayi[107]. Nous citons enfin la méthode Wubizixing se fondant sur les traits de base suivants : le trait horizontal, le trait vertical, le trait descendant à gauche, le trait descendant à droite, et le crochet situés sur le clavier[108].

Il existe aussi des méthodes fondées sur l'utilisation du pinyin. L'utilisateur saisit alors la transcription en pinyin du caractère qu'il souhaite saisir. Une fois le pinyin saisi, une liste de caractères est proposée et l'utilisateur choisit le caractère qu'il souhaite voir afficher. Des exemples de méthodes de saisie au moyen des Pinyin sont : la méthode iBus pour les systèmes Unix, et la méthode SCIM (*smart common input method*) créée par SU Zhe (James Su) de l'université Tsinghua.

Il convient aussi de noter l'existence d'autres méthodes de saisie se fondant sur la transcription des caractères comme le Bopomofo (système de transcription phonétique des caractères chinois en usage à Taiwan). Certains claviers à Taiwan comprennent ainsi à la fois les lettres de l'alphabet, les symboles du Bopomofo, les caractères de la méthode Canjie et ceux de la méthode Dayi. Il est alors possible à l'utilisateur d'utiliser une méthode de saisie fondée sur le Bopomofo, la méthode Dayi ou la méthode Canjie.

Dans la section suivante, nous tentons d'aborder les implications textométriques dans les trois langues.

3.6 Implications textométriques des particularités linguistiques

Les unités de comptage jouent un rôle fondamental dans la textométrie et la lexicométrie. Les implications textométriques sont étroitement liées à la morphosyntaxe des trois langues (française, anglaise et chinoise) qui touche presque tous les niveaux d'organisation langagière, entre autres, le lexique et la syntaxe.

Unités de comptage

Les unités de comptage ou de mesure déterminent la valeur fondamentale des calculs statistiques dans le cadre de nos recherches. Cerner une unité de comptage dans un discours revient à identifier la forme unique d'un mot et/ou d'une notion. Ces formes recensées sont des révélateurs de la pensée du discours, nous appliquerons ces unités de comptage au regard de la problématique de nos recherches. En effet, dans le traitement automatique des langues, la normalisation des découpages de mots permet d'obtenir un bon inventaire de ces unités de comptage, et à chacune de ces unités, nous pouvons également attribuer une catégorie de type morphologique (morphème), syntaxique (syntagme), grammaticale (nom, adverbe, conjonction, etc.), etc. Lors des explorations textométriques, nous nous intéresserons davantage au découpage des mots qu'à ces catégories.

Les unités de comptage sont des éléments clés de la textométrie. Dans les lignes qui suivent, nous tenterons d'expliquer quels sont les impacts et les conséquences des particularités morphosyntaxiques des trois langues dans les travaux de la textométrie.

[106] Ce nom fait référence au mythique ministre qui a inventé les caractères chinois. Sur un clavier contenant 26 clés, par des combinaisons de touches, l'utilisateur forme des caractères, par exemple pour saisir le caractère 明 /míng/brillant, il tape sur la touche contenant le caractère 日/rì/jour, puis sur la touche contenant le caractère 月/yuè/mois ou lune.

[107] Celle-ci fonctionne sur le même principe que la méthode Canjie, 46 caractères sont répartis sur le clavier. L'utilisateur saisit des combinaisons de touches pour former un nouveau caractère.

[108] Le clavier contient les 25 composants les plus fréquents. L'utilisateur entre ensuite les caractères dans l'ordre des traits ou tape sur l'une des touches contenant l'un des composants les plus fréquents.

Homographes[109] et homonymes[110]

La phonétique de l'anglais et du français varie selon les aires géographiques où ces langues sont parlées. Les exemples les plus typiques sont l'anglais américain par rapport à l'anglais britannique et le français métropolitain par rapport au français québécois. La prononciation des mots relève des phénomènes plus complexes tels que les homographes et les homonymes.

- Homographes en anglais : *bank* (banque) / *bank* (rive), *light* (lumière) / *light* (légère), etc.
- Homographes en français : fils (fils de coton)/fils (de 15 ans), lis (une fleur de)/lis (tu lis), etc.

Il est à noter que le phénomène des homographes apporte directement des implications aux calculs de la textométrie car les écritures des formes sont identiques. Ces homographes peuvent également constituer des brèches intéressantes lors de l'interprétation des résultats. Quant aux homonymes, ceux-ci peuvent altérer les calculs statistiques de formes, nous pouvons ainsi en citer quelques exemples.

En anglais, un même graphème peut correspondre à plusieurs prononciations. Le graphème *oo* se prononce [ɔ:] dans le mot *door* et [u:] dans le mot *cool*[111]. Autres exemples, certaines lettres ne sont pas prononcées dans certains mots, comme le t dans *often*, d'autres sont aspirées comme le *h* dans *humor*. De plus, les homonymes restent assez récurrents en français et en anglais. Nous avons en anglais par exemple,

- *sail* (naviguer, faire de la voile) / *sale* (vente), (une prononciation pour 2 formes)
- *tale* (conte) / *tail* (queue), (une prononciation pour 2 formes)
- *will* (volonté) / *we'll* (we will), (une prononciation pour 2 ou 3 formes)
- *why* (pourquoi) / *Y*, etc. (une prononciation pour 2 formes)

En français, nous pouvons en citer également quelques-uns :

- maire / mer / mère (une prononciation pour 3 formes)
- ma / m'as / mas / mât (une prononciation pour 4 ou 5 formes)
- cahot / chaos / K.-O. (une prononciation pour 3 ou 4 formes)
- ta / t'as / tas (une prononciation pour 3 ou 4 formes)
- si / s'y / scie (une prononciation pour 3 ou 4 formes)

Si les phénomènes homographiques ne montrent que la richesse phonétique de l'anglais et du français, alors, les homonymes complexifient nos unités de comptage, en dépit de leurs écritures en alphabet latin.

En chinois, un phonème peut correspondre à plusieurs graphèmes ou caractères. Prenons un exemple, le phonème `mā` correspond à la fois au caractère 吗/`mā`, qui est particule interrogative, et au caractère 妈/`mā`, qui signifie maman. Réciproquement, un graphème peut correspondre à plusieurs phonèmes. L'idéogramme 只 se prononce à la fois `zhī` et `zhǐ`. Quand il s'agit d'un classificateur (souvent pour les quatre membres des êtres humains, les oiseaux, les gallinacés, les objets de forme cylindrique longue et fine, etc.), il se prononce `zhī`. Quand ce caractère signifie *seulement*, on prononce `zhǐ`, comme dans le mot 只有/`zhǐ yǒu`/seulement (littéralement seulement avoir ou seulement il y a).

[109] Définition : (Mot) dont la graphie est identique à celle d'un autre mot. http://www.cnrtl.fr/definition/homographe (consulté le 15/06/2015)
[110] Définition : (Mot, signifiant) qui a une prononciation et/ou une graphie identique à celle d'un autre mais un signifié différent. http://www.cnrtl.fr/definition/homonyme (consulté le 15/06/2015)
[111] Dans certains états des USA, le mot cool est prononcé [ɔ:].

A l'instar de la langue française et de la langue anglaise, ces phénomènes d'homonymie, d'homographie et de caractères polyphones sont également récurrents chez les Chinois, ce qui permet, entre autres, au secteur de l'information et de la communication de jouer avec les calembours. Ces jeux de mots modifient parfois la coupure des mots.

Lexique français et anglais

La plupart des mots de la langue française viennent du latin et du grec[112]. Le vocabulaire joue un rôle déterminant dans les capacités de créativité lexicale des langues. Une langue s'appauvrit lorsque son lexique ne génère plus de nouveaux mots. Dans la mesure où l'anglais et le français se trouvent dans une situation concurrentielle, le dynamisme de l'anglais se caractérise par la richesse de son lexique, tandis que le vocabulaire français s'anglicise, ainsi que le notaient voici plusieurs décennies déjà, Étiemble et aujourd'hui Claude Hagège par exemple.

> *« Dans Linguistique et colonialisme, p. 87-103, Louis-Jean Calvet utilise le taux d'emprunts comme critère par excellence pour évaluer la nature et l'étendue de la domination d'une langue donnée sur une autre. »* (Picone, 1992 : 12)

> *« Nous faisons écho à Jean Tournier* (1988 : 13) *en affirmant que le lexique de toute langue est « en évolution permanente » et que cette évolution « implique un constant enrichissement ». Il s'ensuit, d'après nous, que l'appareil qui permet à telle langue de générer son lexique joue un rôle primordial dans le maintien et la survie de la langue. »* (Picone, 1992 : 12)

Au plan étymologique, le lexique de l'anglais est très varié, ayant subi une influence celtique, normande, germanique, et franque. La lexicographie se partage ainsi entre les vocabulaires venant du vieil anglais, du scandinave, du français et du latin. Cela a permis à la langue anglaise de façonner les mots librement et a induit une sémantique à la fois riche et complexe. De plus, l'anglais utilise beaucoup de contractions et d'abréviations, ce qui peut être source de difficultés pour comprendre le sens d'une forme ainsi que la prise en compte de ces formes dans nos calculs statistiques, par exemple *A.M* pour *Ante Meridiem*, *NYT* pour *New York Times*, *NYC* pour *New York City*, etc. Il convient de noter que, d'une part, le français a exercé une influence importante sur l'anglais, notamment sur le lexique et sur l'orthographe, d'autre part, le français subit une influence de l'anglais. Plus récemment, dans les secteurs des nouvelles technologies, des médias, de la publicité, nous constatons un fort taux d'emprunt de mots anglais. Toutefois, pour certains, c'est une question de dosage, pour d'autres, comme l'Académie française, c'est une question à ne pas négliger. Ces influences mutuelles créent des faux-amis entre les deux langues, complexifiant ainsi le calcul des formes et leurs interprétations analytiques.

L'orthographe française[113] se caractérise par une dimension historique. Au Moyen-Age, l'écriture française est essentiellement phonologique, c'est à dire proche de la prononciation. A partir du XVI^{e} siècle, jusqu'à aujourd'hui plusieurs réformes ont fixé successivement l'orthographe du français (la Pléiade, naissance de l'Académie, etc.). Par exemple, le *h* est supprimé dans des mots tels que *aut(h)eur, méc(h)anique, t(h)résor*[114]. La dernière réforme de l'orthographe date de 1990. Cette

[112] Un exemple, le mot philosophie vient du verbe grec *philein* signifiant aimer et de *sophia* signifiant sagesse en grec ancien.
[113] Source : http://bbouillon.free.fr/univ/hl/Fichiers/Cours/orthog.htm (consulté le 15/06/2015).
[114] *Autheur*, Dictionnaire de l'Académie française, 1ère édition 1694 ; *Autheur,* « *Quelques-uns écrivent Auteur* », Dictionnaire de l'Académie française, 2ème édition 1718 ; Auteur, Dictionnaire de l'Académie française, 3^{e} édition 1740. Autre exemple, le mot *Mechanique* avec un *h* dans la première édition du Dictionnaire de l'Académie (1694), et une nouvelle orthographe, Mécanique, apparaîtra dans la quatrième édition du Dictionnaire, en 1762. Extrait de l'étude de Pierre Bouillon Québec, Québec, Canada.

réforme a modifié la graphie d'environ 1 500 mots. L'objectif de cette réforme était d'apporter des simplifications au niveau de l'orthographe telles que par exemple l'emploi des trémas placés sur la voyelle dans le mot aigüe, sans pour autant rompre la cohérence linguistique du français. L'orthographe française permet aussi de marquer les aspects grammaticaux des différents mots. Nous désignons par morphogrammes grammaticaux les caractères utilisés pour marquer les aspects grammaticaux d'un mot. Cependant, les mots français connaissent parfois des différences subtiles. Ainsi un trait d'union change le sens des mots et de la phrase, comme par exemple la différence d'interprétation du mot composé entre *langue-de-bœuf* qui signifie, entre autres, une dague d'origine italienne ou un champignon ou une plante et *langue de bœuf* qui désigne un mets. Le même phénomène se rencontre pour la locution cœur de bœuf. L'orthographe de ces formes différentes modifie le résultat de nos comptages de mots.

Aspects syntaxiques

La syntaxe de l'anglais est moins complexe que celle du français (Cornish, 2005) et se caractérise par une grande flexibilité, comme l'explique Parisse :

> « *La transformation d'un nom en verbe ou inversement n'implique qu'une simple modification de la position du mot dans la phrase, sans modification du mot lui-même. Les nouvelles formes lexicales peuvent se créer à partir de formes préexistantes et certaines oppositions entre nom et verbe, verbe et adjectif, adverbe et nom s'expriment de manière lexicale alors que ces positions pourraient être seulement flexionnelles ou positionnelles* (Parisse 2009 : 7-20).

Cette flexibilité explique l'efficacité reconnue de la langue anglaise en particulier dans les domaines technologiques et scientifiques, *« progressif, construit du déterminé au déterminant, et considéré comme apportant plus de clarté, cet ordre régressif de l'anglais est reconnu comme plus synthétique »* (Picone, 1992 : 10).

En français et en anglais, le temps est annoncé principalement par le verbe lors de sa conjugaison. Or, ce n'est pas le cas pour le chinois. D'une part, la morphologie du verbe chinois ne relève pas la notion du temps, d'autre part, le temps du procès[115] (au sens linguistique) est indiqué exclusivement par des adverbes. Dans le procès (toujours au sens linguistique), le français et le chinois se démarquent par la télicité[116] du verbe par rapport à l'anglais où les formes progressives, c'est-à-dire, l'aspect inchoatif[117], progressif et parfois envoyant à une modalité future, sont assez courantes.

La flexion sur les verbes en français et en anglais repose sur des marques morphologiques. Tandis qu'en chinois, seuls les segments d'indications temporelles et aspectuelles déterminent la navigation dans le temps et la désignation de l'aspect du procès (au sens linguistique).

La différence de l'aspect syntaxique des trois langues engendre des formes grammaticales qui complexifient parfois la coupure des mots.

[115] Définition : Notion générale en laquelle se résolvent les différentes notions exprimées par le verbe. http://www.cnrtl.fr/definition/proces (consulté le 15/06/2015)
[116] Accomplissement d'une action ou d'un événement.
[117] Définition : Qui indique le déclenchement ou la progression graduelle d'une action. http://www.cnrtl.fr/definition/inchoatif

3.7 Segmenter le texte chinois

Pour pouvoir effectuer des analyses textométriques, il faut impérativement segmenter les textes chinois à l'aide de segmenteurs automatiques seuls capables de fournir un travail rapide et homogène que l'être humain n'est pas à même de réaliser de manière satisfaisante. En effet, les termes obtenus par la segmentation sont découpés à partir de connaissances sur la morphologie chinoise, elles-mêmes sujettes à discussions. Or, les unités textométriques sont tout justement formées à partir de ces termes segmentés.

Nous examinons maintenant les problématiques liées à la segmentation du chinois. Du point de vue du processus de segmentation proprement dit, plusieurs difficultés sont liées à la nature de la langue chinoise. Ainsi le chinois est une langue isolante. Il en résulte que les différents mots chinois, notion précisée plus haut (*cf.* section 3.5), ne contiennent ni genre, ni nombre, ni désinence. Le segmenteur chinois ne pourra donc pas s'appuyer sur les désinences pour déterminer la fonction d'un mot dans la phrase. Plusieurs segmenteurs dont les plus connus sont Hylanda, ICTCLAS, Stanford, et Jieba ont été développés afin de contourner ces difficultés. Nous rappellerons en premier lieu leurs spécificités techniques, puis nous les mettrons en œuvre sur une phrase simple.

Une fois le texte saisi par l'utilisateur, le segmenteur doit déterminer quels sont les mots chinois figurant dans la chaîne de caractères. Il se réfère pour cela à une base de données lexicale comprenant l'ensemble des mots de la langue chinoise. Le problème réside dans le fait qu'une même phrase chinoise peut dans certains cas conduire à plusieurs segmentations possibles.

Par exemple, la phrase,

> 物理学起来很难/wù lǐ xué qǐ lái hěn nán/La physique est très difficile,

est analysée différemment par les quatre segmenteurs distincts Hylanda, ICTCLAS, Stanford, et Jieba.

Ainsi, le segmenteur ICTCLAS découpera la phrase comme suit[118]:

> 物 理/wùlǐ/la physique | 学/xué/étudier | 起 来/qǐ lái/se lever | 很/hěn/très | 难/nán/difficile.

La fonction grammaticale du mot 起来/qǐ lái/se lever dans cette phrase est résultative.

Les segmenteurs Hylanda, Stanford, et Jieba donnent un résultat différent. Ils découpent la phrase comme suit :

> 物 理 学/wùlǐxué/la discipline physique | 起 来/qǐ lái/se lever | 很/hěn/très | 难/nán/difficile.

Nous voyons que les deux découpages proposés sont corrects grammaticalement. En fait, seule une analyse sémantique permet de choisir le premier découpage, dans la mesure où, en tant qu'entité abstraite, *la physique*, n'a pas la capacité de *se lever*.

La multiplicité des choix de segmentation réside ici dans le fait que les mots 物理学/wù lǐ xué/la discipline physique et 物理/wù lǐ/la physique sont tous deux des mots valides, c'est-à-dire, les deux mots sont connus et reconnus dans le langage courant et validés par les dictionnaires officiels chinois.

[118] Pour des facilités de lecture dans cette section, chaque segment en chinois est séparé par une barre verticale « | ».

Ce type de problème, étant une erreur de segmentation, est appelé combinaison de sens multiples. Il s'agit d'un cas où à la fois une chaîne ABC et une chaîne AB sont des mots valides en chinois.

Difficultés de segmentation

L'ambiguïté de la notion de mot chinois et les variations stylistiques et régionales sont sources d'un bon nombre de difficultés de segmentation (Sun, Shen, Tsou et al, 1998).

En 1987, Liu (Liu, 1987) a souligné le fait que l'une des causes des problèmes de segmentation était due à la définition du mot chinois trop peu précise pour un logiciel de traitement automatique des langues. La difficulté de segmentation s'avère encore plus complexe pour les humains. La norme GB 13715 a défini l'unité de segmentation en chinois. Elle se réfère à la fois à des critères sémantiques et syntaxiques. Elle peut aussi inclure des noms de lieux. Cette norme peut servir de référence pour le traitement automatique du chinois, mais elle comporte beaucoup d'exceptions. D'autres normes fondées sur la norme GB 13715 ont par la suite été mises en place avec difficultés, difficultés mises en évidence par de nombreux chercheurs tels que Sun et Zhou (Sun et Zhou, 2001).

Il convient aussi de noter qu'un mot chinois peut connaître certaines variations de prononciation selon la région. Ainsi les chinois du nord disent souvent 一点儿/yī diǎn er pour dire un peu. Les chinois du sud prononcent plus souvent 一点/yī diǎn. Des variations existent aussi dans les particules finales. Ainsi, au nord-est, il est utilisé la particule 嘛/ma pour évoquer une idée de controverse. Les habitants du sud préfèrent davantage utiliser la particule 哟/yo. Les habitants du sud-ouest ont aussi tendance à redoubler les noms qu'ils utilisent, comme par exemple, pour désigner un sac, ils emploient 包包/bāo bāo au lieu de 包/bāo. Les Pékinois préfèrent souvent dire 包儿/bāo er. Ces éléments stylistiques et régionaux, considérés comme des variantes libres ou contextuelles, qui sont à prendre en compte par un segmenteur, montrent la difficulté d'établir des normes d'unité de segmentation.

3.7.1 Les segmenteurs automatiques

Il existe plusieurs segmenteurs dont l'objectif est de résoudre les problèmes évoqués plus haut. Pour évaluer ces segmenteurs, plusieurs critères entrent en ligne de compte. Le premier critère est bien sûr la justesse du découpage sur un plan sémantique. Un second critère est le temps d'exécution et la consommation des ressources mémoire allouées. La complexité algorithmique permet d'évaluer le temps d'exécution de l'algorithme.

3.7.1.1 Le segmenteur Hylanda

Le segmenteur Hylanda est un logiciel propriétaire développé par l'entreprise du même nom située à Tianjin en Chine. Il utilise vraisemblablement deux types d'algorithmes de segmentation :

- la recherche maximale en avant
- la recherche maximale en arrière

Ces algorithmes de segmentation prennent en entrée le texte à segmenter sous la forme d'une chaîne de caractères, et le dictionnaire des termes chinois. Ils sont essentiellement utilisés par le segmenteur Hylanda. Dans le cas de la méthode de recherche maximale en avant, le texte est parcouru de gauche à droite. Le but est alors de trouver le groupe de caractères le plus long dans le dictionnaire. La méthode de recherche maximale en arrière parcourt le texte de droite à gauche et donne la chaîne la plus longue. La méthode de recherche de segmentation bidirectionnelle applique les deux méthodes précédentes, et choisit la segmentation la plus courte. Par exemple, la phrase :

> 物理学起来很难/wù lǐ xué qǐ lái hěn nán/La physique est très difficile.

aurait été segmentée comme suit si l'on avait appliqué l'une des méthodes décrites précédemment :

> 物理学/wù lǐ xué/la discipline physique | 起来/qǐ lái/se lever | 很/hěn/très | 难/nán/difficile.

Cette méthode de recherche maximale (Chen et Liu, 1992) permettait d'obtenir de bons résultats de segmentation dans des temps acceptables. L'une des faiblesses de ces méthodes réside cependant dans le fait que si l'un des mots de la chaîne de caractères à segmenter n'est pas dans le dictionnaire, il risque d'y avoir des erreurs de segmentation. Or, l'apparition régulière de nouveaux mots chinois rend impossible la construction d'un dictionnaire exhaustif. De plus, l'aspect sémantique de la phrase n'est pas non plus pris en compte.

Les méthodes de recherche maximale, en avant, en arrière et bidirectionnelle nécessitent de mettre en œuvre un algorithme de recherche : nous devons effectuer une recherche[119] de chacun des k (k est un entier naturel) caractères et groupes de caractères de la chaîne dans le dictionnaire contenant N mots (N est un entier naturel). Cette recherche peut être effectuée au moyen de l'algorithme de recherche dichotomique, l'un des algorithmes de recherche le plus efficace, dont nous rappelons ici les principes et la complexité décrits dans la plupart des cours d'algorithmique. Cet algorithme part du principe que la liste, dans laquelle l'élément est cherché, est triée par ordre croissant. Afin de trier cette liste, l'idée serait d'attribuer un code à chaque mot chinois (caractères).

[119] Le logiciel Hylanda étant propriétaire, la méthode précise de recherche utilisée par ce produit n'est pas connue. Nous présentons donc ci-dessous une méthode de recherche bien connue : la recherche dichotomique. Nous évoquerons aussi les arbres lexicographiques et les tables de hachage : structures de données qui permettent d'améliorer l'efficacité d'une recherche.

Voici ensuite en pseudo code un exemple d'implémentation de cet algorithme.

```
RechercheDichotomique(elementCherche,premierElementListe,dernierElementListe,Liste)

    Si la liste est vide
     Renvoyer « élément non trouvé »
     milieuListe = partie entière de((premierElementListe + dernierElementListe)/2)
          Si elementCherche = Liste[milieuListe]
          Renvoyer « élement trouvé »
          Si elementCherche > Liste[milieuListe]
          RechercheDichotomique(milieuListe + 1,dernierElementListe, elementCherche, Liste)
          Si elementCherche < Liste[milieuListe]
          RechercheDichotomique(premierElementListe,milieuListe - 1, elementCherche, Liste)
 Fin
```

Donnons aussi un exemple d'utilisation de cet algorithme. On considère qu'on a codé 4 mots par les entiers 1 2 3 et 4. On cherche le mot 3. La liste utilisée et triée par ordre croissant est [1, 2, 3, 4]. A la première étape, on constate que le terme du milieu de cette liste est 2 et est plus petit que 3. On extrait donc de la liste [1, 2, 3, 4] la liste [3, 4], et on regarde l'élément du milieu de cette liste (c'est le premier élément car la liste a deux éléments). On trouve 3 et on a terminé l'exécution de l'algorithme.

Examinons maintenant le nombre d'opérations que cet algorithme effectue.

Nous constatons que la taille de la liste est divisée par 2 à chaque itération de cet algorithme. Ainsi au bout de m opérations, la taille N de la liste devient :

$$\frac{N}{2^m}$$

Le nombre d'instructions m doit être tel que

$$\frac{N}{2^m} \geq 1$$

car l'algorithme s'arrête lorsque la liste est vide. On doit donc avoir :

$$2^m \leq N$$

ce qui donne :

$$\exp(m\ ln(2)) \leq N \text{ ou } m \ln(2) \leq \ln(N)$$

Donc :

$$m \leq \frac{\ln(N)}{\ln(2)} \text{ ou } m \leq \log_2(N)$$

L'algorithme de recherche dichotomique suit donc une complexité logarithmique, car le nombre d'opérations est majoré par un logarithme de la taille N de la liste. Le nombre d'opérations à effectuer au total pour chacun des k éléments de la chaîne de caractères à segmenter est donc majoré par :

$$\frac{k \ln(N)}{\ln(2)} = \text{k } \log_2 N$$

On a donc un nombre d'opérations majoré par

$$k \frac{\ln(N)}{\ln(2)} = \text{k } \log_2 N$$

A titre d'exemple si $N = 10^6$, l'ordinateur réalisera au plus k log2(N) opérations. Un algorithme de recherche « naïf » parcourant pour chaque caractère l'intégralité des N entrées du dictionnaire aurait conduit à un nombre d'opération majoré par

$$k\ N = k \text{ milliers}$$

opérations. La mise en place d'une méthode de recherche dichotomique permet donc de gagner un temps de calcul important. En effet l'ordinateur doit effectuer un nombre d'opérations multiplié, pour $N = 10^6$, par environ $\frac{N}{\log_2 N} \sim \frac{10^6}{20} = 50\ 000$ par rapport à l'algorithme de recherche dichotomique. Cela est dû au fait que ln(N) est plus petit que N. Donc k ln(N) est plus petit que k N.

Plusieurs algorithmes de recherche s'appuient sur les arbres lexicographiques, qui permettent d'enregistrer l'ensemble des mots d'un lexique. Un arbre lexicographique est construit comme suit :

- la racine est un nœud conventionnel et n'a pas de signification particulière. Les nœuds fils de la racine contiennent les premières lettres des mots du dictionnaire rangés par ordre alphabétique.

- appelons *f* un nœud fils de la racine, et notons *a* la lettre qu'il porte. Les nœuds fils de f contiennent les deuxièmes lettres des mots du dictionnaire commençant par a. Chaque nœud a au plus 26 fils car l'alphabet comprend 26 lettres.

En poursuivant ainsi, il est possible d'obtenir un arbre lexicographique.

Une autre structure de données utilisée pour effectuer des recherches correspond aux tables de hachage. Celles-ci consistent à assigner à chaque élément d'une liste une clé. On accède ensuite aux éléments de cette liste au moyen de leurs clés. Cette structure de données permet d'effectuer des recherches plus rapidement.

Le segmenteur Hylanda a aussi la capacité de déterminer la catégorie grammaticale des unités segmentées. Il a aussi la possibilité de détecter des noms propres, des noms de lieux, ou encore des noms d'organismes.

3.7.1.2 Le segmenteur ICTCLAS

Le segmenteur ICTCLAS[120] (*Institute of Computing Technology Chinese Lexical Analysis System*) est fondé sur l'algorithme de Dijkstra et le modèle de Markov caché. Il a été créé par Kevin Zhang et Qun Liu de *l'Institute of Computing Technology, Chinese Academy of Sciences*. Ce segmenteur[121] donne une précision de segmentation de 97,58%. Il est capable de détecter plus de 98% (taux de rappel) des noms et prénoms chinois. Sa vitesse de segmentation atteint environ 31,5 kilobits par seconde en C/C++. La première étape consiste à découper le texte à segmenter en unités lexicales minimales. Par exemple le texte : 我是中国人/wǒ shì zhōng guó rén/Je suis chinois, sera découpé ainsi : 我/wǒ/je | 是/shì/être | 中/zhōng/milieu | 国/guó/pays | 人/rén/gens. On écrit ensuite sur la même ligne les mots commençant par le même caractère ou la même lettre. Dans notre exemple, chacun des mots 我/wǒ, 是/shì, 中/zhōng, 国/guó, 人/rén occupera une ligne distincte. Un graphe orienté est ensuite construit sur le principe suivant : le successeur d'un segment (ou mot) a comme numéro de ligne le numéro de colonne de son successeur. Les segmentations possibles de la chaîne de caractères sont ensuite données par l'algorithme de Dijkstra, dont nous rappelons les principes ci-dessous.

L'algorithme de Dijkstra a pour but de trouver le chemin minimal entre deux endroits sachant que l'algorithme du plus court chemin est une part mineure de cette méthode, le point important est le Markov caché (MMC) ou *hidden Markov model* (HMM). Il prend en entrée un graphe, c'est-à-dire, un ensemble de sommets (qui modélisent les endroits) et d'arêtes (qui modélisent les portions de chemin parcourues). Une arête est une paire de sommets. A chaque arête est associé un nombre qu'on appelle le poids. Le poids d'une paire de sommets, qu'on note A et B, est aussi, lorsqu'ils sont non reliés par une seule arête, la somme des poids des arêtes qui mènent de A à B. Il y en a donc potentiellement plusieurs. Le but de l'algorithme de Dijkstra sera alors de trouver les chemins (constitué des arêtes à parcourir) de poids minimal afin d'aller d'un sommet A vers un sommet B. L'algorithme s'initialise donc en affectant au sommet initial le poids 0. Les autres sommets du graphe sont affectés d'un poids infini. Il est ensuite relevé les sommets adjacents à A (c'est-à-dire reliés par une arête à A). Le sommet choisi, noté S, est celui relié à A par une arête de poids minimal. Le même procédé est appliqué à nouveau au sommet S. L'algorithme s'arrête lorsqu'on est parvenu au sommet final. Du point de vue de la complexité, il est possible de montrer que le nombre maximal est au plus n^2(avec n le nombre de sommets).

Le segmenteur ICTCLAS utilise aussi le modèle de Markov caché. Rappelons au préalable les grands principes du modèle de Markov à travers l'exemple d'une fourmi qui peut aller dans deux positions possibles une position A ou une position B. Le modèle de Markov est composé d'états et de probabilités de transition d'un état à l'autre. Par exemple considérons deux endroits A et B. Considérons une fourmi initialement située en A. Supposons qu'une fois en A, la fourmi va en B avec la probabilité 0,5. Notons $X = (X_k)_{0 \le k \le n}$, un processus de Markov, X_t correspondant à la position de la fourmi au temps t (nous supposons que t est dans l'ensemble des entiers naturels), t peut prendre deux valeurs A ou B. Comme X est une chaîne de Markov d'ordre 1, les probabilités de transitions suivantes au temps t + 1 ne dépendent que de la position au temps t et pas de la position aux temps précédents $t-1$, $t-2$, etc. :

$$P(X_{t+1} = B \mid X_t = A\,, X_{t-1}\,, X_{t-2}, \dots) = P(X_{t+1} = B \mid X_t = A\,) = 0{,}5$$

et

$$P(X_{t+1} = A \mid X_t = A\,, X_{t-1}\,, X_{t-2}, \dots) = P(X_{t+1} = A \mid X_t = A\,) = 0{,}5$$

[120] http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/English.html
[121] http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/codes.html

(La première équation nous donne la probabilité que la fourmi soit en B au temps t + 1 sachant qu'elle était en A au temps t. La seconde équation nous donne la probabilité que la fourmi soit en A au temps t + 1 sachant qu'elle était en A au temps t).

$$P(X_{t+1} = B \mid X_t = \text{B}) = 1 \text{ et } P(X_{t+1} = A \mid X_t = \text{B}) = 0$$

car la fourmi ne quitte jamais l'état B lorsqu'elle y est.

Les chaînes de Markov ne nécessitent donc pas de connaître les états passés et les états futurs du modèle étudié, ce qui allège les calculs à effectuer.

Dans le cas d'un modèle de Markov caché, il n'y a pas une connaissance totale de la séquence d'états. Chacun des états émet des observations qui sont connues. A titre d'exemple, si on considère des sauts observables de la fourmi aux temps 1, 2, 3, 4 et 5, toutes les positions de la fourmi ne seront cependant pas connues (par exemple seules les positions de la fourmi aux temps 1 et 2 seront connues). Une séquence d'observations sera alors émise au cours du temps.

Un modèle de Markov cachée sera donc caractérisé par cinq paramètres :

1. l'ensemble des N états,
2. l'ensemble des M observations,
3. la matrice de transition de taille NxN représentant les probabilités de passer d'un état i à un état j (i et j compris entre 1 et N, la ligne i et la colonne j représente la probabilité de passer de l'état i à l'état j),
4. la matrice d'émission de taille NxM qui correspond à la probabilité pour chacun des N états d'émettre l'une des M observations possibles.
5. la matrice de taille 1xN qui représente la probabilité de commencer dans l'un des N états.

Les principes du modèle de Markov caché sont utilisés pour calculer le poids (ou la longueur) des arêtes du graphe orienté constitué à partir de la chaîne de caractères à segmenter. Chacune des unités lexicales minimales trouvées dans la chaîne de caractères à segmenter correspond à l'ensemble des états. De manière générale, une unité lexicale correspond à un état. On note S_t l'unité choisie au temps t. Le concepteur du logiciel ICTCLAS (Zhang et al, 2003) nous donne l'expression[122] suivante pour le poids de l'arc reliant une unité lexicale minimale i à une unité lexicale minimale j :

$$-\log(P(S_t = \textit{unité lexicale minimale } i \mid S_{t-1} = \textit{unité lexicale minimale } j))$$

La fonction qui à x associe $-\log(x)$ étant décroissante, nous en déduisons que plus la probabilité $P(S_t = \textit{unité lexicale minimale } i \mid S_{t-1} = \textit{unité lexicale minimale } j)$ sera grande, plus

$-\log\big(P(S_t = \textit{unité lexicale minimale } i \mid S_{t-1} = \textit{unité lexicale minimale } j)\big)$ sera petit.

Donc, en appliquant le critère de l'algorithme de Dijkstra (choix du chemin le plus court), le choix de l'unité lexicale minimale i au temps t se fera donc en choisissant l'unité lexicale minimale i telle que la probabilité de passer de l'unité lexicale minimale i au temps t, sachant que nous sommes à l'unité

[122] Remarquons que cette expression est bien de signe positif. En effet, si $0 < P(S_t = \textit{unité lexicale minimale } i \mid S_{t-1} = \textit{unité lexicale minimale } j) \leq 1$ alors par les propriétés du logarithme
$\log\big(P(S_t = \textit{unité lexicale minimale } i \mid S_{t-1} P(S_t = \textit{unité lexicale minimale } i \mid S_{t-1} = \textit{unité lexicale minimale } j)) = \textit{unité lexicale minimale } j)\big) \leq 0$.
On remarque de plus que quand $P(S_t = \textit{unité lexicale minimale } i \mid S_{t-1} = \textit{unité lexicale minimale } j))$ tend vers 0, alors également du fait des propriétés du logarithme
$-\log\big(P(S_t = \textit{unité lexicale minimale } i \mid S_{t-1} = \textit{unité lexicale minimale } j)\big)$ tend vers $+\infty$

lexicale j au temps t - 1, soit maximale. Il s'agit dans ce cas d'une décision locale, sinon l'algorithme du plus court chemin choisit un optimum global.

Prenons par exemple les 26 lettres de l'alphabet et la langue française. Certaines suites de lettres telles que *ks* ne se rencontrent que très rarement voire jamais. A l'inverse des suites de lettres telles que *li* ou *la* sont plus fréquentes. La probabilité de rencontrer la suite de lettres *ks* sera donc faible, tandis que les probabilités de rencontrer les lettres *la* ou *li* seront plus élevées. Pour segmenter un texte, le segmenteur ICTCLAS utilisant l'algorithme cité ci-dessus, préférera les suites de lettres *la* et *li* aux suites de lettres *ks*. En effet la distance définie plus haut séparant les lettres l et a et les lettres l et i sera plus petite que celle séparant les lettres k et s dans le cas du français.

Un exemple de modèle de Markov caché pour le traitement automatique des langues peut consister dans l'identification des séquences de lettres qui ont conduit à l'apparition d'un mot.

3.7.1.3 Le segmenteur Stanford

Le segmenteur Stanford a été développé à l'Université de Stanford et est fondé sur le modèle des champs aléatoires conditionnels qui constituent une amélioration du modèle de Markov caché. Nous rappelons ci-dessous le modèle des champs aléatoires conditionnels[123] expliqué sur le site des auteurs du logiciel.

- une variable aléatoire X représentant la séquence des observations à étiqueter ou à segmenter.
- une variable aléatoire Y représentant la liste des étiquettes. Nous pouvons considérer Y comme une variable aléatoire, représentant la séquence des étiquettes dans le texte à segmenter. L'objectif sera alors d'estimer Y à partir de la séquence des observations X.

On définit alors un modèle conditionnel caractérisé par la loi de Y sachant X. La séquence d'étiquettes sera donc définie conditionnellement par la séquence d'observations. Contrairement au modèle de Markov caché, l'étiquette choisie au temps t ne dépendra pas seulement de l'observation au temps t - 1 mais de toute la séquence d'observations. Avec les *Conditional random fields* (CRF ou champs markoviens conditionnels) linéaires d'ordre 1 (habituel), l'étiquette à la position t dépend aussi de l'étiquette à la position t - 1. Nous comprenons alors que ce type de modèle est plus adapté pour segmenter une séquence de caractères chinois. On construit ensuite un graphe G de sommets S, et d'arêtes A. L'ensemble des sommets S indexe la séquence d'étiquettes. On note cela comme suit : Y_s avec s dans l'ensemble des sommets S. Il convient de noter que l'état du sommet parcouru au temps t ne dépend que des sommets voisins (c'est à dire pour lesquels un chemin partant du sommet actuellement parcouru existe), et des probabilités de transition entre chacun des sommets. On montre que la séquence des étiquettes ou des délimiteurs (Y) suit une distribution de la forme [124]

$$P(Y \mid \mathrm{X}) \;=\; 1/\mathrm{A}\exp\Big(\sum_{i=1}^{N}\sum_{l=1}^{K} l_k\, f_k\ (Y_i, Y_{i-1}, \mathrm{X}, \mathrm{i})$$

Avec N le nombre d'observations,

A une constante de normalisation de sorte que la somme des probabilités P(Y|X) soit égale à 1. On a alors,

[123] http://nlp.stanford.edu/software/jenny-ner-2007.ppt (consulté le 01/03/2016)
[124] https://people.cs.umass.edu/~wallach/technical_reports/wallach04conditional.pdf (consulté le 01/03/2016)

$$A = \sum_{i=1}^{N} P(Y_i \mid X)$$

f_k des fonctions caractéristiques dont nous décrivons le principe plus bas.

K : le nombre de fonctions caractéristiques utilisées dans le modèle.

l_k: les poids affectés à chaque fonction caractéristique.

Les fonctions caractéristiques f_k sont des fonctions à valeurs dans R+. Elles seront cependant très souvent des fonctions à valeurs dans {0,1}. Nous voyons qu'elles dépendent de la séquence observée X, et des sommets de la séquence Y reliés seulement au sommet courant et au sommet précédent. Par exemple nous pouvons considérer la fonction caractéristique suivante f ;

f = 1 si l'étiquette du sommet courant est un marqueur de verbe et si le sommet courant est un point d'interrogation, et f = 0 sinon.

Un poids $l_k > 0$ indiquera une contribution positive de la fonction caractéristique f_k à la probabilité des P(Y|X). A l'inverse un poids $l_k < 0$ indiquera une contribution négative à la probabilité P(Y|X). Différentes méthodes statistiques, que nous ne détaillons pas ici, existent pour estimer ces paramètres l_k.

Le segmenteur de Stanford exploite aussi des aspects liés au lexique et aux noms pour traiter la décision de segmenter une séquence de caractères en affectant la valeur 1 si le caractère est séparable et 0 sinon. Il s'agit ici d'un codage d'une segmentation par des étiquettes sur des unités individuelles, nécessitant de modéliser le problème pour le modèle des champs aléatoires conditionnels (CRF). Par exemple, dans la séquence de caractères 我学汉语/wǒ xué hàn yǔ/J'apprends le chinois, on aura pour chacun des caractères les valeurs suivantes :

- 我/wǒ/je : 1, pour exprimer que le caractère est à séparer de 学/xué.
- 学/xué/apprendre : 1, pour exprimer que le caractère est à séparer de 汉/hàn.
- 汉/hàn/chinois : 0 , pour exprimer que le caractère est indissociable de 语/yǔ.
- 语/yǔ langue : 0, pour exprimer le caractère est indissociable de 汉/hàn.

Il convient aussi de noter que le segmenteur de Stanford est capable d'exploiter des relations grammaticales dans une phrase pour améliorer la segmentation. Par exemple, le segmenteur exploitera la fonction grammaticale des mots 我/wǒ/je : le sujet, 学/xué/apprendre : le verbe et 汉语/hàn yǔ/la langue chinoise : le complément d'objet direct. Au moyen de l'analyse des formes et des fonctions grammaticales des différentes unités de la chaîne de caractères à segmenter, le segmenteur de Stanford peut repérer des formes ne figurant pas dans le dictionnaire utilisé. Selon le site de présentation du segmenteur de Stanford[125], le taux de reconnaissance de mots inconnus est de 84,84%. Au moyen du modèle CRF (Champs markoviens conditionnels ou *Conditional Random Fields*), le segmenteur de Stanford offre aussi la possibilité de reconnaître des noms propres.

[125] http://nlp.stanford.edu/projects/chinese-nlp.shtml (consulté le 15/03/2016)

3.7.1.4 Le segmenteur Jieba

Jieba est un segmenteur, écrit sous forme de package Python, capable de gérer les caractères chinois simplifiés et traditionnels. Il s'utilise avec l'instruction : import jieba.

Il offre de nombreuses fonctionnalités telles que :

- la possibilité d'utiliser un dictionnaire personnalisé.
- l'extraction d'un mot dans une chaîne de caractères.
- l'activation ou la désactivation du modèle de Markov caché, pour la reconnaissance des mots inconnus, avec l'algorithme de Viterbi que nous décrivons plus bas.

Les fonctions de ce segmenteur sont :

- *jieba.cut* qui prend en paramètre la chaîne à segmenter, et le choix d'utiliser le modèle de Markov caché.
- *jieba.cut_for_search* qui permet d'extraire les mots de la chaîne de caractères. Par exemple, pour la chaîne :

  我在国立东方语言文化学院学汉语/
  wǒ zài guólì dōngfāng yǔyàn wénhuà xuéyuàn xué hànyù/
  J'apprends le chinois à l'Institut national des langues et civilisations orientales (INALCO).

  L'instruction *jieba.cut_for_search* renverra cette chaine segmentée en mots :

  我 /wǒ/je | 在 /zài/être | 国 立 /guólì/national | 东 方 /dōngfāng/oriental | 语 言 /yǔyàn/langues | 文化/wénhuà/civilisations | 学院/xuéyuàn/institut | 国立东方语言文化学院/INALCO[126] | 学/xué/apprendre | 汉语/hànyù/langue chinoise.

Ce type de fonction est particulièrement bien adapté pour rechercher un mot dans une chaîne de caractères.

Afin d'obtenir des gains en rapidité, il utilise une structure de données informatique appelée arbre lexicographique ou « trie[127] » ou encore arbre préfixe. Dans ce type de structure de données, la racine de l'arbre est la chaîne vide. Dans le cas de l'alphabet latin ses descendants seront les 26 lettres de l'alphabet. Dans le cas du chinois, les descendants correspondront aux différents caractères chinois disponibles dans le dictionnaire. Les descendants du nœud A seront les mots commençant par A par exemple air, avion, etc. Ceux du nœud B seront des mots commençant par B balle, bien, etc. Ensuite, les expressions commençant par balle seront les descendants du nœud balle. A la différence d'un arbre binaire de recherche, aucun nœud ne stocke la chaîne de caractères à laquelle il est associé. La chaîne cherchée se déduit en fait de la position du nœud dans l'arbre. Les structures de données "trie" sont aussi utilisées par exemple pour faire des suggestions à l'utilisateur d'un moteur de recherche. Par exemple, lorsque l'utilisateur tape *goo*, le moteur de recherche suggère les nœuds descendants du nœud "*goo*" comme par exemple *google*.

[126] L'entité nommée 国立东方语言文化学院 / guólì dōngfāng yǔyán wénhuà xuéyuàn / L'Institut national des langues et civilisations orientales peut être segmentée en au moins deux façons : 国立 | 东方 | 语言 | 文化 | 学院 ou 国立 | 东方 | 语言文化 | 学院

[127] Le terme vient de *retrievable memory*.

Jieba utilise aussi la technique de la programmation dynamique pour chercher les combinaisons de caractères les plus probables. Les probabilités de combinaison de caractères sont calculées, en fonction de leur fréquence d'emploi dans la langue utilisée. Par exemple, en français, une succession d'un t et d'un x est très peu probable, et sera en conséquence affectée d'une probabilité nulle. Il s'agit d'une technique de programmation visant à résoudre certains problèmes d'optimisation pouvant avoir plusieurs solutions. Le terme a été employé pour la première fois dans les années 1950 par Richard Bellman[128]. Elle se base sur le principe de « Diviser pour régner », dans la mesure où elle consiste à utiliser des solutions de sous-problèmes inclus dans le problème d'optimisation pour aboutir à une solution globale du problème à résoudre. Cette méthode se base sur le principe d'optimalité de Bellman qui s'énonce comme suit : « Si (C) est un chemin optimal pour aller d'un point A à un point B, et si le point C appartient à (C), alors les sous-chemins de (C) allant de A à C et de C à B sont optimaux ». Dans le domaine du traitement automatique des langues, l'algorithme d'Earley utilise les principes de la programmation dynamique. Cet algorithme consiste à parcourir la chaîne à analyser de gauche à droite. A chaque étape du parcours de la chaîne de caractères, on construit des arbres d'analyse syntaxique partielle. L'analyse syntaxique de la phrase est obtenue en combinant ces arbres d'analyses syntaxiques partielles.

Pour la reconnaissance de mots (connus et inconnus), comme nous l'avons dit précédemment, Jieba utilise le modèle de Markov caché décrit plus haut, et l'algorithme de Viterbi se basant aussi sur le principe de la programmation dynamique. Cet algorithme part d'une séquence de mots et d'une séquence de délimiteurs. Il vise à calculer la séquence de délimiteurs la plus probable pour la séquence de mots donnée. A l'instar des méthodes de programmation dynamique, il utilise pour le i-ème mot les résultats des calculs effectués pour le i – 1 ième mot.

Une autre caractéristique intéressante de Jieba est d'utiliser les modules de programmation parallèle de Python activables au moyen de l'option ; jieba.enable_parallel(n) avec n le nombre de processus parallèles. Chaque processus parallèle traite une partie de la chaîne à segmenter. Le résultat est alors obtenu beaucoup plus rapidement.

3.7.2 Comparaison des différents segmenteurs existants

Nous venons de rappeler les spécificités des quatre segmenteurs dans la section précédente. Nous allons les mettre en œuvre sur deux textes extraits de nos corpus de veille. Mais au préalable, nous proposons une segmentation subjective et personnelle. Cependant, les comparaisons inter-segmenteurs ne tiendront pas compte de cette segmentation subjective, en revanche, les analyses sémantiques s'appuieront sur les cinq résultats de segmentation (4 automatiques + 1 subjective).

Les différents résultats de ces deux phrases seront analysés et comparés, deux par deux, à l'aide d'un logiciel spécialement conçu pour cette tâche. Dans la section suivante, nous comparons les segmentations des deux textes extraits dans leur totalité, afin de montrer les avantages et inconvénients de ces quatre segmenteurs.

[128] Richard Ernest Bellman, mathématicien américain, inventeur de la programmation dynamique.

3.7.2.1 Critères de comparaison des quatre segmenteurs

Les segmentations peuvent être différentes selon le segmenteur. L'un des premiers critères d'évaluation du segmenteur sera un critère sémantique. Il s'agit de s'assurer que la segmentation obtenue est la plus proche du sens original du texte, autrement dit, le sens du texte ne doit pas être modifié à cause de la segmentation. De plus, le segmenteur devra si possible nous montrer les catégories grammaticales entre les différents mots des phrases. Afin de mieux illustrer les difficultés, nous allons prendre d'abord trois exemples en français montrant l'importance de la coupure des mots :

- « pomme de terre », n'est ni une pomme, ni la terre.
- « Hôtel de ville », n'est ni un hôtel, ni une ville, ni un hôtel dans la ville.

Si cette séquence est coupée en trois mots, le sens change complètement.

- « porte-parole » n'est ni une porte, ni une parole.

Si on ignore le tiret de ce mot, alors, le sens n'est plus le même.

Or, si nous raisonnons par analogie, dans le quotidien de la langue chinoise, l'écriture sans blanc est omniprésente, ce phénomène est appelé *scriptio continua* (*cf.* section 3.5.1). Il peut s'illustrer par le mécanisme suivant : si le français moderne s'écrivait comme le latin, alors, la phrase *fermer le théâtre* deviendrait *fermerlethéâtre*.

La compréhension de ce segment contracté, *fermerlethéâtre*, se base fortement sur l'identification des mots à l'intérieur de cette suite de caractères. La séparation des mots n'est pas figée, seules des chaînes de caractères se regroupent à travers la lecture en fonction de la connaissance et de l'interprétation du lecteur. Dans notre cas, plus de trois segmentations sont possibles, donnant plus de trois sens différents :

1. fer | mer | le | thé | âtre : découpage donnant des morphèmes opérants, la validation de cette segmentation dépend du contexte.
2. fer | merle | thé | âtre : découpage donnant des morphèmes opérants, la validation de cette segmentation dépend du contexte.
3. fermer | le | théâtre : ce découpage demeure le plus pertinent.

Cet exemple en français reflète, de manière flagrante, la complexité de la segmentation d'une langue en *scriptio continua*. Ainsi, nous pouvons projeter ce phénomène sur la langue chinoise afin de saisir la difficulté de la formation des mots et la problématique de la segmentation en mots. En effet, d'une part, la segmentation constitue une étape fondamentale et délicate de tous les traitements automatiques, d'autre part, la fiabilité et l'efficacité des différents segmenteurs automatiques sont les deux critères les plus cruciaux de tout traitement automatisé des informations de cette langue.

Choix des textes pour la comparaison

Afin de réaliser cette comparaison, nous avons choisi un extrait de deux articles[129]. Le premier article a été extrait du sous-corpus chinois de notre corpus comparable (*cf.* chapitre 4), tandis que le deuxième est issu du volet chinois du corpus parallèle (*cf.* chapitre 4).

Les deux articles traitent d'écologie, une problématique d'actualité, très débattue sur Internet et les réseaux sociaux. Le premier article aborde le sujet de la possibilité d'adopter un mode de vie ne nuisant pas à l'environnement, mode de vie en renonçant aux différentes innovations technologiques. Le second article examine la gestion des problématiques écologiques par le gouvernement chinois.

Le choix de ces deux articles s'inscrit donc dans le thème de l'analyse d'opinion de notre étude. Par ailleurs, l'emploi d'un lexique relativement moderne, à la fois technologique et écologique permet de tester à la fois l'aptitude des segmenteurs à reconnaître des formes nouvelles, c'est-à-dire, l'étendue du lexique existant. D'une manière générale, il est à noter que la presse écrite chinoise utilise le 书面语 /shū miàn yǔ/langue écrite, qui se différencie de la langue orale ou encore 口语/kǒuyǔ/langue orale par plusieurs aspects. Tout d'abord, le 书面语/shū miàn yǔ contient davantage d'expressions idiomatiques que le 口语 /kǒuyǔ/langue orale. Ces expressions seront très difficiles à analyser pour un logiciel de segmentation s'il ne dispose pas de ces expressions dans son lexique initial. Le 书面语 /shū miàn yǔ/langue écrite utilise aussi des tournures beaucoup plus brèves que le 口语 /kǒuyǔ/langue orale, ce qui rend naturellement plus difficile le travail du segmenteur.

Segmentation d'une phrase simple

A titre d'illustration, nous commençons par la segmentation d'une phrase suscitant des ambiguïtés, phrase choisie parmi l'extrait de l'article issu du corpus parallèle (discours d'une ONG, *cf.* chapitre 4). Cette phrase sera soumise à une segmentation subjective et personnelle, puis aux quatre segmenteurs retenus dans le cadre de notre étude.

La phrase retenue est la suivante :

> 中外对话：也就是说，“大气十条”在具体的政策措施上，有内在矛盾.

Traduction française[130] :

> Le site *Chinadialogue* : autrement dit, des contradictions internes se manifestent dans les applications concrètes des «Dix mesures et politiques défendant la qualité de l'air» (littéralement : grand, l'air, dix, unités).

[129] La traduction des titres des deux articles est la suivante :
1. Est-il possible de vivre « débranché » ? http://green.sina.com.cn/2012-04-18/172024292468.shtml (consulté le 23/12/2015)
2. Le plan contre la pollution de l'air est à jeter aux oubliettes. La croissance demeure la priorité en Chine (une traduction en anglais est proposée par le site). https://www.chinadialogue.net/article/show/single/ch/6462-Forget-the-new-air-pollution-plan-GDP-growth-is-still-king-in-China (consulté le 23/12/2015)

[130] Toutes les traductions en français au cours de cette thèse sont réalisées par l'auteur.

Nous proposons une segmentation subjective :

中外对话 | ：也 | 就是说，“大气十条”| 在 | 具体的 | 政策措施 | 上，有 | 内在 | 矛盾.

Cette phrase peut se présenter sous la forme d’une liste de segments de mots, respectant l’ordre de la phrase (de gauche à droite) :

- 中外对话 / zhōng wài duì huà / le site *Chinadialogue* | ：
- 也 / yě / aussi |
- 就是说 /jiù shì shuō / à savoir，|
- “大气十条”/ dà qì shí tiáo / «Dix mesures et politiques défendant la qualité de l’air» (littéralement : grand, l’air, dix, unités) |
- 在 / zài / être |
- 具体的 / jù tǐ de / concret |
- 政策措施 [131]/ zhèng cè cuò shī / les mesures et politiques |
- 上 / shàng / sur，|
- 有 / yǒu / il y avoir |
- 内在/ nèi zài / interne |
- 矛盾/ máo dùn / contradiction.

Nous allons comparer de manière automatique, cette même phrase segmentée par les quatre segmenteurs.

Comparateur automatique de la segmentation du chinois

Au cours de cette étude et plus particulièrement sur la segmentation de texte en chinois, nous avons été amenés à nous apercevoir de la complexité de cette tâche. Par conséquent, nous avons conçu un comparateur de segmentation pour la langue chinoise en C++, appelé *WordCompareTool*. L’originalité de ce logiciel réside dans la comparaison de manière automatique, deux par deux, d’un même texte segmenté par différents outils ou moyens (manuellement ou automatiquement). Il repère, recense et affiche les différences entre les résultats de la segmentation, sous la forme d’un rapport en page HTML (se reporter à l’annexe M).

[131] Il faut noter que l’expression 政策措施/zhèng cè cuò shī est composée de 政策/zhèng cè/mesure(s) politique(s) et 措施/cuò shī /mesure(s). Selon l’auteur, il est préférable de traduire cette expression par les mesures et politiques mais pas simplement par mesure(s) politique(s).

Hylanda versus Jieba

Formes segmentées différemment		
Mots repérés dans Hylanda	Mots repérés dans Jieba	Fréquence des formes différentes dans le texte
政策/ zhèng cè/ les politiques | 措施/ cuò shī / les mesures	政策措施 / zhèng cè cuò shī / les mesures et politiques	1

Hylanda	Jieba
中外 对话 也就是说 大气 十条 在 具体 的 政策 措施 上 有 内在 矛盾	中外 对话 也就是说 大气 十条 在 具体 的 政策措施 上 有 内在 矛盾

Différence principale entre Hylanda et Jieba :

Un seul mot a été segmenté différemment, 政策措施 / zhèng cè cuò shī / les mesures et politiques. Selon Jieba, les politiques et mesures prônées par les autorités compétentes forment une seule unité sémantique et lexicale tandis que selon Hylanda cette notion représente deux différents morphèmes lexicaux.

Hylanda versus ICTCLAS

Formes segmentées différemment		
Mots repérés dans Hylanda	Mots repérés dans ICTCLAS	Fréquence des formes différentes dans le texte
十条/shí tiáo/dix unités	十/shí/dix | 条/tiáo/unités	1

Hylanda	ICTCLAS
中外 对话 也就是说 大气 十条 在 具体 的 政策 措施 上 有 内在 矛盾	中外 对话 也就是说 大气 十条 在 具体 的 政策 措施 上 有 内在 矛盾

Différence principale entre Hylanda et ICTCLAS :

Le mot 十条/shí tiáo/dix unités, sous-entendu dix politiques et mesures, a été segmenté différemment par les deux logiciels.
Selon Hylanda, cette notion est un lexème individuel, mais ICTCLAS considère que le mot 条 /tiáo/unités est le classificateur du nombre dix.

Hylanda versus Stanford

Formes segmentées différemment		
Mots repérés dans Hylanda	Mots repérés dans Stanford	Fréquence des formes différentes dans le texte
十条/shí tiáo/dix unités	十/shí/dix | 条/tiáo/ unités	1

Hylanda	Stanford
中外 对话 也就是说 大气 十条 在 具体 的 政策 措施 上 有 内在 矛盾	中外 对话 也就是说 大气 十 条 在 具体 的 政策 措施 上 有 内在 矛盾

Différence principale entre Hylanda et Stanford :

Le mot 十条/shí tiáo/dix unités, sous-entendu dix politiques et mesures a été séparé en deux caractères (idem ICTCLAS).

ICTCLAS versus Stanford

Formes segmentées différemment		
Mots repérés dans ICTCLAS	Mots repérés dans Stanford	Fréquence des formes différentes dans le texte
Néant	Néant	0

ICTCLAS	Stanford
中外 对话 也就是说 大气 十 条 在 具体 的 政策 措施 上 有 内在 矛盾	中外 对话 也就是说 大气 十 条 在 具体 的 政策 措施 上 有 内在 矛盾

Aucune différence entre ICTCLAS et Stanford

ICTCLAS versus Jieba

Formes segmentées différemment		
Mots repérés dans ICTCLAS	Mots repérés dans Jieba	Fréquence des formes différentes dans le texte
政策/ zhèng cè/ les politiques | 措施/ cuò shī / les mesures	政策措施 / zhèng cè cuò shī / les mesures et politiques	1

ICTCLAS	Jieba
中外 对话 也就是说 大气 十 条 在 具体 的 政策 措施 上 有 内在 矛盾	中外 对话 也就是说 大气 十条 在 具体 的 政策措施 上 有 内在 矛盾

Différence principale entre ICTCLAS et Jieba :

L'expression 政策措施 / zhèng cè cuò shī / les mesures et politiques, est segmentée par ICTCLAS mais pas par Jieba.

Stanford versus Jieba

<table>
<tr><td colspan="2">Formes segmentées différemment</td><td></td></tr>
<tr><td>Mots repérés dans Stanford</td><td>Mots repérés dans Jieba</td><td>Fréquence des formes différentes dans le texte</td></tr>
<tr><td>十/shí/dix | 条/tiáo/ unités</td><td>十条/shí tiáo/dix unités</td><td>1</td></tr>
<tr><td>政策/ zhèng cè/ les politiques | 措施/ cuò shī / les mesures</td><td>政策措施 / zhèng cè cuò shī / les mesures et politiques</td><td>1</td></tr>
<tr><td colspan="2">Stanford

中外 对话 也就是说 大气 十 条 在 具体 的 政策 措施 上 有 内在 矛盾</td><td>Jieba

中外 对话 也就是说 大气 十条 在 具体 的 政策措施 上 有 内在 矛盾</td></tr>
<tr><td colspan="3">Différence principale entre Stanford et Jieba:

Pour le mot 十条/shí tiáo/dix unités, sous-entendu dix politiques et mesures, tout comme ICTCLAS, Stanford sépare le chiffre 十/shí/dix et le classificateur 条/tiáo/unités, en deux mots, il en est de même pour le mot 政策措施/ zhèng cè cuò shī / les mesures et les politiques. Cependant, Jieba considère ces deux mots comme des unités lexicales individuelles et opérantes.</td></tr>
</table>

En résumé, dans les lignes qui suivent, nous récapitulons les différences principales de segmentation, constatées sur cette même phrase. Nous ajoutons la segmentation de l'auteur ainsi que les interprétations de ces résultats afin d'élucider les difficultés rencontrées par les segmenteurs.

Phrase segmentée par Hylanda
中外 | 对话 |也就是说 |大气 |十条|在 |具体 |的| 政策 | 措施|上 |有| 内在| 矛盾
Phrase segmentée par ICTCLAS
中外 | 对话 |也就是说 |大气 |十 |条|在 |具体 |的| 政策 | 措施 |上 |有| 内在| 矛盾
Phrase segmentée par Stanford
中外 | 对话 |也就是说 |大气 |十 |条|在 |具体 |的| 政策 | 措施 |上 |有| 内在 |矛盾
Phrase segmentée par Jieba
中外 | 对话 |也就是说 |大气 |十条 |在 |具体 |的| 政策措施 |上 |有 |内在 |矛盾
Phrase segmentée par l'auteur (segmentation subjective)
中外对话 | 也 | 就是说 |大气十条|在 | 具体的 | 政策措施 |上|有 | 内在 | 矛盾

Les mots grisés ci-dessus, correspondent aux mots segmentés différemment. Nous constatons que les différences des quatre segmenteurs restent mineures et portent sur les mêmes segments :

- 十/shí/dix | 条/tiáo/ unités : ICTCLAS et Stanford séparent ce segment en deux mots, le chiffre 十/shí/dix et le classificateur 条/tiáo/unités. Pour les deux autres, ce segment forme une unité lexicale.
- 政策措施/ zhèng cè cuò shī / les mesures et les politiques : seul Jieba considère ces deux mots comme une unité lexicale individuelle et opérante.

Quant à la segmentation subjective, l'auteur prend en compte toutes les entités nommées et les notions et les considère comme un morphème opérant. En effet, dans la phrase ci-dessus, le segment 中外对话 / zhōng wài duì huà / le site *Chinadialogue*, forme une entité nommée à part entière, *idem* pour la notion acronymique, 大气十条/ dà qì shí tiáo / «Dix mesures et politiques défendant la qualité de l'air» (littéralement : grand, l'air, dix, unités). Le fait de les segmenter est une erreur de la part des segmenteurs, car ces nouvelles notions n'existent pas encore dans leurs dictionnaires informatiques. En ce qui concerne les autres différences par rapport à celles de l'auteur, la discussion demeure totalement ouverte :

- 也 / yě /aussi, est un adverbe,
- 就是说 /jiù shì shuō / à savoir, forme une unité sémantique, (par exemple, 这就是说 ; 那就是说/zhé jiù shì shuō ; nà jiù shì shuō /c'est-à-dire),
- 具体的/ jù tǐ de/concret, la particule 的/de, permet de déterminer un segment comme un adjectif ou un déterminant. Elle reste attachée au radical du mot.

Segmentation d'une deuxième phrase

A titre de comparaison, nous avons également segmenté une autre phrase suscitant des ambiguïtés, phrase choisie parmi l'extrait de l'article issu du sous-corpus chinois du corpus comparable (discours de la presse chinoise, *cf.* chapitre 4). Cette phrase sera soumise de la même manière que celle issue du volet chinois du corpus parallèle, c'est-à-dire, à une segmentation subjective et personnelle, puis aux quatre segmenteurs retenus dans le cadre de notre étude.

La phrase retenue est la suivante :

> 至于如何看待这种生活方式, 是否会遵循这种生活方式, 则是仁者见仁智者见智了.

Traduction française :

> En ce qui concerne les questions telles que : quelle est la vision sur ce mode de vie (écologique) ? Faut-il le respecter et le pérenniser ? Cela revient à dire : chacun a sa propre vision (sur ce mode de vie).

Nous proposons une segmentation subjective :

> 至于 | 如何 | 看待 | 这种 | 生活 | 方式 | , 是否 | 会 | 遵循 | 这种 | 生活 | 方式 |, 则是 | 仁者见仁 | 智者见智 | 了.

Cette phrase peut se présenter sous la forme d'une liste de segments de mots, respectant l'ordre de la phrase (de gauche à droite) :

- 至于 / zhìyú / quant à |
- 如何 / rúhé / comment |
- 看待 / kàndài / considérer ou voir|
- 这种 / zhè zhǒng / ce genre de |
- 生活 / shēnghuó / la vie |
- 方式 / fāngshì / le mode | ,
- 是否 / shìfǒu / oui ou non |
- 会 / huì / pouvoir|
- 遵循 / zūnxún / respecter et suivre |
- 这种 / zhè zhǒng / ce genre de |
- 生活 / shēnghuó / la vie |
- 方式 / fāngshì / mode |,
- 则是 / zé shì / alors |
- 仁者见仁 / rén zhě jiàn rén / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les bienveillants voient [au travers des yeux de] la bienveillance).
- 智者见智 / zhìzhě jiàn zhì / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les sages voient [au travers des yeux de] la sagesse).
- 了/ le / particule modale, se place en fin de phrase et exprime un changement d'état ou une actualisation.

Comparaison par comparateur automatique

Maintenant, nous allons comparer, de manière automatique, deux par deux, cette même phrase segmentée par les quatre différents outils.

Hylanda versus Jieba

<table>
<tr><td colspan="2">Formes segmentées différemment</td><td></td></tr>
<tr><td>Mots repérés dans Hylanda</td><td>Mots repérés dans Jieba</td><td>Fréquence des différentes formes dans le texte</td></tr>
<tr><td>Néant</td><td>Néant</td><td>0</td></tr>
<tr><td colspan="2">Hylanda

至于 如何 看待 这种 生活 方式 是否 会 遵循 这种 生活 方式 则 是 仁者见仁 智者见智 了</td><td>Jieba

至于 如何 看待 这种 生活 方式 是否 会 遵循 这种 生活 方式 则 是 仁者见仁 智者见智 了</td></tr>
<tr><td colspan="3">Aucune différence entre Hylanda et Jieba.</td></tr>
</table>

Hylanda versus ICTCLAS

<table>
<tr><td colspan="2">Formes segmentées différemment</td><td></td></tr>
<tr><td>Mots repérés dans Hylanda</td><td>Mots repérés dans ICTCLAS</td><td>Fréquence des différentes formes dans le texte</td></tr>
<tr><td>这种/zhè zhǒng/ ce genre de</td><td>这/zhè/ ce, 种/zhǒng/ genre</td><td>1</td></tr>
<tr><td colspan="2">Hylanda

至于 如何 看待 这种 生活 方式 是否 会 遵循 这种 生活 方式 则 是 仁者见仁 智者见智 了</td><td>ICTCLAS

至于 如何 看待 这 种 生活 方式 是否 会 遵循 这 种 生活 方式 则 是 仁者见仁 智者见智 了</td></tr>
<tr><td colspan="3">Différence principale entre Hylanda et ICTCLAS :

Le mot 这种/zhè zhǒng/ ce genre de, a été segmenté en deux mots en début de phrase mais pas au milieu (de cette même phrase) par ICTCLAS. Hylanda n’a pas segmenté ce mot.

Nota : dans le résultat d’ICTCLAS, la première occurrence de ce mot 这/zhè/ ce, est considérée par le logiciel ICTCLAS, comme un pronom verbal, annoté /rzv (figure 3.1, 4ème ligne, ci-après), et non comme un pronom démonstratif. Selon notre analyse, ceci est une erreur du logiciel. L’annotation sous l’étiquette de pronom verbal est une pratique grammaticale, souvent utilisée dans les cas suivants : 这样/zhè yàng/ comme ceci, ou 这么样/zhè me yàng /comme cela. Or, dans cette phrase, nous ne sommes pas dans ce type de cas.

Par contre, dans la deuxième occurrence, le mot 这种/zhè zhǒng/ce genre de, est un pronom démonstratif. Dans cette phrase, la distinction des deux segmentations pour le même mot n’est pas justifiée. C’est une erreur de segmentation d’ICTCLAS. En effet, dans ce cas précis, les deux mots ont une fonction grammaticale identique, à savoir, un pronom démonstratif.</td></tr>
</table>

Hylanda versus Stanford

Formes segmentées différemment		
Mots repérés dans Hylanda	Mots repérés dans Stanford	Fréquence des formes différentes dans le texte
这种/zhè zhǒng/ ce genre	这/zhè/ ce, 种/zhǒng/genre	1
仁者见仁 / rén zhě jiàn rén / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les bienveillants voient la bienveillance)	仁者/rén zhě/les bienveillants | 见仁/jiàn rén/voient bienveillance,	1
智者见智 / zhìzhě jiàn zhì / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les sages voient la sagesse)	智者 / zhìzhě / les sages | 见智 / jiàn zhì / voient sagesse	1

Hylanda	Stanford
至于 如何 看待 这种 生活 方式 是否 会 遵循 这种 生活 方式 则 是 仁者见仁 智者见智 了	至于 如何 看待 这 种 生活 方式 是否 会 遵循 这种 生活 方式 则 是 仁者 见仁 智者 见智 了

Différences principales entre Hylanda et Stanford :

- Le mot 这种/zhè zhǒng/ ce genre de, a été segmenté en deux mots (这/zhè/ ce, et 种 /zhǒng/genre), en début de phrase et au milieu (de cette même phrase) par Stanford. Hylanda n'a pas segmenté ce mot. En effet, Stanford considère que le caractère 这/zhè/ ce, est le pronom démonstratif du caractère 种/zhǒng/genre.
- Pour chacune des deux parties de l'idiotisme, 仁者见仁 / rén zhě jiàn rén / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les bienveillants voient la bienveillance) et 智者见智 / zhìzhě jiàn zhì / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les sages voient la sagesse), le segmenteur de Stanford considère qu'il y a une séparation entre le sujet, 仁者/rén zhě/les bienveillants, et le groupe verbal, 见仁/jiàn rén/voient la bienveillance, d'où la segmentation de mots dans chacune de ces deux parties.

ICTCLAS versus Stanford

Formes segmentées différemment		
Mots repérés dans ICTCLAS	Mots repérés dans Stanford	Fréquence des formes différentes dans le texte
这种/zhè zhǒng/ ce genre	这/zhè/ ce, 种/zhǒng/ genre	1
仁者见仁 / rén zhě jiàn rén / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les bienveillants voient la bienveillance)	仁者/rén zhě/les bienveillants 见仁/jiàn rén/voient bienveillance,	1
智者见智 / zhìzhě jiàn zhì / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les sages voient la sagesse)	智者 / zhìzhě / les sages 见智 / jiàn zhì / voient sagesse	1

ICTCLAS	Stanford
至于 如何 看待 这 种 生活 方式 是否 会 遵循 这种 生活 方式 则 是 仁者见仁 智者见智 了	至于 如何 看待 这 种 生活 方式 是否 会 遵循 这 种 生活 方式 则 是 仁者 见仁 智者 见智 了

Différences principales entre ICTCLAS et Stanford :

Les trois différences de segmentation relèvent strictement du même phénomène que celui constaté entre Hylanda et Stanford (se reporter aux explications ci-dessus).

ICTCLAS versus Jieba

Formes segmentées différemment		
Mots repérés dans ICTCLAS	Mots repérés dans Jieba	Fréquence des formes différentes dans le texte
这/zhè/ ce, 种/zhǒng/ genre	这种/zhè zhǒng/ ce genre	1

ICTCLAS	Jieba
至于 如何 看待 这 种 生活 方式 是否 会 遵循 这 种 生活 方式 则 是 仁者见仁 智者见智 了	至于 如何 看待 这种 生活 方式 是否 会 遵循 这种 生活 方式 则 是 仁者见仁 智者见智 了

Différence principale entre ICTCLAS et Jieba :

Dans le résultat d'ICTCLAS, la première occurrence de la forme 这种/zhè zhǒng/ ce genre, a été segmentée en deux caractères tandis que dans le résultat de Jieba, cette forme demeure inchangée et non segmentée. Quant à la deuxième occurrence de la forme 这种/zhè zhǒng/ ce genre, celle-ci n'a pas été segmentée (se reporter aux explications ci-dessus, Hylanda versus ICTCLAS et Hylanda versus Stanford).

Stanford versus Jieba

<table>
<tr><td colspan="2">Formes segmentées différemment</td><td></td></tr>
<tr><td>Mots repérés dans Stanford</td><td>Mots repérés dans Jieba</td><td>Fréquence des formes différentes dans le texte</td></tr>
<tr><td>这/zhè/ ce, 种/zhǒng/ genre</td><td>这种/zhè zhǒng/ ce genre</td><td>2</td></tr>
<tr><td>仁者/rén zhě/les bienveillants 见仁/jiàn rén/voient bienveillance</td><td>仁者见仁 / rén zhě jiàn rén / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les bienveillants voient la bienveillance)</td><td>1</td></tr>
<tr><td>智者 / zhìzhě / les sages 见智 / jiàn zhì / voient sagesse</td><td>智者见智 / zhìzhě jiàn zhì / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les sages voient la sagesse)</td><td>1</td></tr>
<tr><td colspan="2">Stanford

至于 如何 看待 这 种 生活 方式 是否 会 遵循 这 种 生活 方式 则 是 仁者 见仁 智者 见智 了</td><td>Jieba

至于 如何 看待 这种 生活 方式 是否 会 遵循 这种 生活 方式 则 是 仁者见仁 智者见智 了</td></tr>
<tr><td colspan="3">Différences principales entre Stanford et Jieba:

• Dans le résultat de Stanford, les deux occurrences de la forme 这种/zhè zhǒng/ ce genre, ont été segmentées en deux caractères, tandis que dans le résultat de Jieba, cette forme demeure inchangée et non segmentée (se reporter aux explications ci-dessus).
• Dans la locution idiomatique, 仁者见仁 / rén zhě jiàn rén / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les bienveillants voient la bienveillance), 智者见智 / zhìzhě jiàn zhì / chacun a sa vision ou chacun voit à travers ses lunettes (littéralement, les sages voient la sagesse), Jieba considère qu'il s'agit d'une expression figée ne pouvant pas en général être scindée, contrairement à Stanford (se reporter aux explications ci-dessus, Hylanda versus Stanford).</td></tr>
</table>

En résumé, nous récapitulons les différences principales de segmentation, constatées sur cette deuxième phrase issue de notre corpus comparable. Nous ajoutons la segmentation de l'auteur ainsi que les interprétations de ces résultats afin d'élucider les difficultés rencontrées par les segmenteurs.

Phrase segmentée par Hylanda
至于 |如何 |看待 |这种| 生活 |方式 | 是否| 会 |遵循 |这种 |生活 |方式 |, 则 |是 |仁者见仁 |智者见智 了
Phrase segmentée par ICTCLAS
至于 |如何 |看待 |这 |种 |生活 |方式 | 是否| 会 |遵循 |这种 |生活 |方式| , 则 |是 |仁者见仁 |智者见智 了
Phrase segmentée par Stanford
至于 |如何 |看待 |这 |种 |生活 |方式 | 是否| 会 |遵循 |这 |种 |生活 |方式| , 则 |是 |仁者 |见仁 |智者 |见智 |了
Phrase segmentée par Jieba
至于 |如何 |看待 |这种| 生活 |方式 | 是否| 会 |遵循 |这种 |生活| 方式| , 则 |是 |仁者见仁 | 智者见智 了
Phrase segmentée par l'auteur (segmentation subjective)
至于 | 如何 | 看待 | 这种 | 生活 | 方式 | 是否 | 会 | 遵循 | 这种 | 生活 | 方式 |, 则是 | 仁者见仁 | 智者见智 | 了

Les mots grisés ci-dessus, correspondent aux mots segmentés différemment dans cette deuxième phrase. Les constats sur leurs différences sont plus révélateurs que ceux de la première phrase :

- ICTCLAS segmente les phrases avec une fonction d'annotations grammaticales assez fine, distinguant les mots en fonction de leur contexte, mais une erreur a cependant été constatée (这/zhè/ce, 种/zhǒng/genre).
- Standford opère un découpage plus fin que les trois autres. D'un point de vue textométrique, un découpage fin permet de récupérer des segments répétés par les calculs statistiques.
- Hylanda et Jieba obtiennent des résultats similaires, résultats très proches de ceux de l'auteur, toutefois, Hylanda reste un logiciel payant.

Hylanda, Jieba et l'auteur prennent en compte les entités entières comme un pronom démonstratif (这种/zhè zhǒng/ ce genre), ainsi que l'idiotisme chinois (仁者见仁 / rén zhě jiàn rén, 智者见智 / zhìzhě jiàn zhì / chacun a sa vision des choses.

Quant à la seule différence par rapport à la segmentation opérée par l'auteur, la discussion demeure totalement ouverte :

- 则是 / zé shì / alors ou quant à, dans le cas de cette deuxième phrase, les deux caractères de ce mot, forment une locution adverbiale que nous pouvons traduire par *alors* ou *c'est alors que* (littéralement, quant à + être). Ce mot s'emploie assez fréquemment dans le langage soutenu. Ceci se justifie par la source de cette phrase (la presse chinoise).

Principe de comparaison pour les deux extraits d'articles

Dans cette étude, notre principe de comparaison consiste à compter le nombre de mots qui ont été segmentés différemment par les quatre segmenteurs automatiques dans les deux textes choisis. Nous allons stocker ces nombres dans un tableau afin de faciliter les analyses.

Après la comparaison de deux phrases simples, nous allons maintenant comparer les résultats de ces deux extraits dans leur totalité, résultats produits par les quatre segmenteurs.

Extrait de l'article issu du corpus comparable (la presse) non segmenté :

关注“不插电”生活，是为了让大家了解这种特殊的生活方式，如同透过万花筒一样，看到世界上原来还有一些人过着这样的生活。至于如何看待这种生活方式，是否会遵循这种生活方式，则是仁者见仁智者见智了。

他们提倡“不插电”

“放弃电动跑步机吧，你不觉得在平地上跑步更舒服吗？”

——悉尼大学“健康运动”俱乐部

“挤公交不但省钱，更是一项集散打、平衡木和瑜伽于一体的运动，自己开车哪有这种好处？”

——印度孟买“绿色环境”俱乐部

“微波炉是 20 世纪最不该出现的发明，它改变了生物分子结构，下一步就是改变人体生理结构。与其用微波炉加热，不如把食物放在蒸锅上。”

——澳大利亚“新家庭主妇”女性社团

“每天少看一小时电视，每个月就省下 30 个小时。IBM 可以用 30 小时设计一个硬件，斯皮尔伯格可以用 30 小时修改完一个剧本，Gucci 可以用 30 小时完成一张 600 万澳元订单。算算看，电视让你失去了多少财富？”

Ce premier extrait sera appelé T_Presse ci-après.

Traduction française de l'extrait ci-dessus

> Le fait de s'intéresser au sujet « vivre de manière débranchée » va permettre de comprendre ce mode de vie particulier. C'est comme si nous regardons la vie à travers un kaléidoscope, nous apercevons que dans le monde, il y a encore des gens qui vivent de cette façon. En ce qui concerne les questions telles que comment peut-on faire face à ce mode de vie ? Faut-il le respecter et le pérenniser ? Cela revient à dire : chacun a sa vision.
> Ils préconisent de vivre « débranché » :
> « Abandonnez les tapis de course, vous ne vous sentez pas mieux de courir sur le sol ? » -- Club Sports & Santé, Université de Sydney
> « Se serrer dans les bus serait une activité qui permet non seulement d'économiser de l'argent, mais aussi de pratiquer une sorte de sport, sport rassemblant les avantages du combat libre (le Sanda), de la poutre et du yoga. Conduire une voiture ne possède point ces avantages. » -- Club Environnement vert de Bombay en Inde.
> « Le four à micro-ondes est un appareil qui n'aurait pas dû être inventé au XX^{e} siècle, car il modifie la structure des biomolécules, par la suite, ces biomolécules pourraient modifier la structure physique du corps humain. Au lieu d'utiliser le chauffage par micro-ondes, il faut mieux consommer la nourriture cuite à la vapeur. » -- "Nouvelle femme au foyer" société féminine, Australie.
> « Si on regarde la télévision moins d'une heure par jour, cela permet d'atteindre une totalité de 30 heures par mois. En 30 heures, IBM concevrait un matériel, Spielberg modifierait un script, la maison Gucci réaliserait une commande de 6 millions en dollar australien. Faites votre compte, la télévision vous fait perdre combien de richesses ? ».

Extrait de l'article issu du corpus parallèle (ONG) non segmenté :

"大气十条"虽出台，GDP 增长仍是王道
中外对话：您如何评价"大气十条"作为一个全国性专项政策的力度？
李俊峰：不能说没有力度，但是还需要增加力度。这主要体现在三个方面。
一是，对严重性的认识不够。根据国家环保部公布的资料，我国的主要城市空气达标的不到 10%，京津冀、长三角尤为严重，已经影响到人民的健康、投资者的热情。 二是，突出主要矛盾不够。说是大气十条，但是洋洋洒洒，有很多具体措施，但是减少或是控制煤炭消费、提高油品质量是主要矛盾，这两个方面的力度还不够。
三是，对自身的能力的自信度不够。中国在过去很多联防联控的成功经验，也有 1 年增加 500 亿天然气、2000 万千瓦水电、1800 万千瓦风电的能力，控制阴霾主要靠优化能源结构，这个又有自信，有信心。
中外对话：也就是说，"大气十条"在具体的政策措施上，有内在矛盾？

Ce deuxième extrait sera appelé T_ONG ci-après.

Traduction française de l'extrait ci-dessus

Ce texte relate l'interview entre un journaliste du site *Chinadialogue* et Monsieur Li Junfeng, directeur du Centre national de la stratégie du changement climatique et de la coopération internationale (NCSC[132]) sur les dix (grandes) mesures promulguées par le gouvernement chinois.

> Titre de l'article : Bien que les «dix mesures et politiques défendant la qualité de l'air (de Chine)» (littéralement : l'air, dix unités/objets) soient promulguées, la croissance du PIB demeure toujours un enjeu primordial.
>
> Le site *Chinadialogue* : les «dix mesures et politiques défendant la qualité de l'air» (ci-après, Dix mesures) sont des directives spécifiques applicables à l'échelle nationale, comment évaluez-vous leurs performances ?

[132] www.ncsc.org.cn (consulté le 28/12/2015)

Li Junfeng : Je ne peux pas dire que ces mesures n'ont pas de poids, mais il faudrait intensifier les efforts. Ceux-ci s'illustrent principalement par trois aspects.

Tout d'abord, le manque de prise de conscience de la gravité (de la situation). Selon les informations publiées par le ministère de la protection de l'environnement, moins de 10% des grandes villes atteignent les normes de qualité de l'air exigées. La situation à Beijing, Tianjin, la province du Hebei et le delta du fleuve Yangtze est particulièrement grave. Cette pollution affecte la santé de la population et l'enthousiasme des investisseurs.

Le deuxième aspect est l'absence du traitement des problèmes majeurs. Dans les « Dix mesures », les grandes lignes sont tracées et beaucoup de mesures spécifiques sont également évoquées. Mais le cœur des problèmes n'est qu'effleuré : d'une part, la réduction ou le contrôle de la consommation de charbon, d'autre part, l'amélioration de la qualité des carburants. Des efforts sont à développer davantage sur ces deux derniers points.

Troisièmement, le manque de confiance sur notre propre capacité (à réaliser des projets). La Chine possède déjà beaucoup d'expériences dans la lutte contre la pollution par une coordination préventive et contrôlée à l'échelle nationale. En un an, nous avons réussi à augmenter notre capacité de production énergétique pour atteindre 50 milliards de mètres cubes de gaz naturel, 20 millions de kilowatts pour l'énergie hydroélectrique et 18 millions de kilowatts pour l'énergie éolienne. Le contrôle des particules fines (le Smog) dans l'air se réalise principalement par l'optimisation de la structure de l'énergie. Nous devons être confiants et nous sommes confiants.

Le site *Chinadialogue* : autrement dit, des contradictions internes se manifestent-elles dans les applications concrètes des «Dix mesures» ?

3.7.2.2 Résultats avec ICTCLAS

Nous utilisons ici la version en ligne [133]:

Guide de lecture pour les figures 3.1 et 3.2

La partie gauche de la figure correspond au texte segmenté par ICTCLAS avec les annotations grammaticales. Chaque annotation est représentée sous forme d'une case avec une couleur spécifique.

La partie droite de la figure se compose de trois parties :

Partie supérieure, la légende des annotations grammaticales sous forme de cases de couleur :

- Première ligne : orange : nom, bleu foncé : verbe, mauve foncé : adjectif, jaune : circonstance de temps, rose : nom d'orientation.
- Deuxième ligne : rose pâle : nombre, rose foncé : pronom, mauve : lieu d'habitation, mauve-bleu : adjectif non-prédicatif, azurin : mot descriptif.
- Troisième ligne : azur clair : classificateur/quantificateur, vert amande : adverbe, pistache : mot modal, jaune chartreuse : onomatopée, mauve : chaine de caractères.
- Quatrième ligne : jaune claire : préposition, rose dragée : conjonction, marron clair : particule, vert foncé : mot d'exclamation, gris : signe de ponctuation.
- Cinquième ligne : bleu paon : préfixe, violet gris : suffixe, rouge : mot personnalisé.

[133] http://ictclas.nlpir.org/nlpir/ (consulté le 28/12/2015)

Partie médiane : néologisme ou mots non recensés dans le dictionnaire d'ICTCLAS.

Partie inférieure : case permettant de charger un dictionnaire personnalisé.

Pour T_Presse, le résultat de segmentation est reproduit partiellement dans la figure ci-dessous. Le segmenteur remarque de manière pertinente que le thème général du texte concerne le mode de vie 生活方式/shēnghuó fāngshì/mode de vie.

Figure 3.1 Extrait du résultat de la segmentation de T_Presse produit par ICTCLAS

Le résultat de segmentation de T_ONG est reproduit partiellement dans la figure ci-dessous. Le segmenteur relève que le thème de l'article est celui de la politique.

Figure 3.2 Extrait du résultat de la segmentation de T_ONG produit par ICTCLAS

3.7.2.3 Résultats avec le segmenteur Stanford

L'exécution du segmenteur Stanford par des lignes de commande et l'environnement *Eclipse* donne les résultats de segmentation suivants :

Pour T_Presse :

关注 " 不 插 电 " 生活 , 是 为了 让 大家 了解 这 种 特殊 的 生活 方式 , 如同 透过 万花筒 一样 , 看到 世界 上 原来 还 有 一些 人 过 着 这样 的 生活 . 至于 如何 看待 这 种 生活 方式 , 是否 会 遵循 这 种 生活 方式 , 则 是 仁者 见仁 智者 见智 了 .

他们 提倡 " 不 插电 "
" 放弃 电动 跑步 机 吧 , 你 不 觉得 在 平地 上 跑步 更 舒服 吗 ? "
悉尼 大学 " 健康 运动 " 俱乐部
" 挤 公交 不但 省 钱 , 更 是 一 项 集 散打 , 平衡木 和 瑜伽 于 一体 的 运动 , 自己 开车 哪 有 这 种 好处 ? "
印度 孟买 " 绿色 环境 " 俱乐部
" 微波炉 是 20 世纪 最 不 该 出现 的 发明 , 它 改变 了 生物 分子 结构 , 下 一 步 就 是 改变 人体 生理 结构 . 与其 用 微波炉 加热 , 不如 把 食物 放在 蒸锅 上 . "
澳大利亚 " 新 家庭 主妇 " 女性 社团
" 每 天 少 看 一 小时 电视 , 每 个 月 就 省下 30 个 小时 . IBM 可以 用 30 小时 设计 一 个 硬件 , 斯皮尔伯格 可以 用 30 小时 修改 完 一 个 剧本 , Gucci 可以 用 30 小时 完成 一 张 600 万 澳元 订单 . 算算 看 , 电视 让 你 失去 了 多少 财富 ? "

Pour T_ONG :

“ 大 气 十 条 ” 虽 出 台 , GDP 增 长 仍 是 王 道
中 外 对 话 ： 您 如 何 评 价 “ 大 气 十 条 ” 作 为 一 个 全 国 性 专 项 政 策 的 力 度 ？
李 俊 峰 ： 不 能 说 没 有 力 度 ， 但 是 还 需 要 增 加 力 度 。 这 主 要 体 现 在 三 个 方 面 。
一是， 对 严重性 的 认识 不 够 。 根据 国家 环保部 公布 的 资料， 我国 的 主要 城市 空气 达标 的 不到 10%， 京 津冀 、 长三 角 尤 为 严 重 ， 已 经 影 响 到 人 民 的 健 康 、 投 资 者 的 热 情 。
二 是， 突出 主要 矛盾 不够 。 说 是 大气 十 条， 但是 洋洋洒洒， 有 很多 具体 措施， 但是 减少 或是 控制 煤炭 消费 、 提高 油 品 质 量 是 主 要 矛 盾 ， 这 两 个 方 面 的 力 度 还 不 够 。
三是， 对 自身 的 能力 的 自信度 不够 。 中国 在 过去 很多 联防 联控 的 成功 经验， 也 有 1 年 增加 500 亿 天然气 、 2000 万千 瓦 水电 、 1800 万千 瓦 风电 的 能力 ， 控制 阴霾 主要 靠 优化 能源 结构 ， 这个 又 有 自信 ， 有 信心 。
中外 对话： 也就是说， “ 大气 十 条 ” 在 具体 的 政策 措施 上， 有 内在 矛盾 ？

3.7.2.4 Résultats avec le segmenteur Jieba

Afin de lancer la procédure informatique activant le segmenteur Jieba, nous avons écrit le code Python disponible ci-après :

```
#encoding=utf-8
import jieba
content = open('texteAsegmenter.txt', 'rb').read()
print ("Input： ", content)
words = jieba.cut(content, cut_all=False)
print ("Résultat de la segmentation ： ")
for word in words:
print (word, end = '/')
```

Nous obtenons alors les résultats suivants (chaque entité de segmentation est séparée par un /) :

Pour T_Presse :

关注/"/不/插/电/"/生活/,/是/为了/让/大家/了解/这种/特殊/的/生活/方式/,/如同/透过/万花筒/一样/,/看到/世界/上/原来/还有/一些/人过/着/这样/的/生活/./至于/如何/看待/这种/生活/方式/,/是否/会/遵循/这种/生活/方式/,/则/是/仁者见仁/智者见智/了/./
/他们/提倡/"/不/插/电/"/
/"/放弃/电动/跑步机/吧/,/你/不/觉得/在/平地上/跑步/更/舒服/吗/?/"/
/悉尼大学/"/健康/运动/"/俱乐部/
/"/挤/公交/不但/省钱/,/更是/一项/集散/打/,/平衡木/和/瑜伽/于/一体/的/运动/,/自己/开车/哪有/这种/好处/?/"/
/印度/孟买/"/绿色/环境/"/俱乐部/
/"/微波炉/是/20/世纪/最/不该/出现/的/发明/,/它/改变/了/生物/分子结构/,/下/一步/就是/改变/人体生理/结构/./与其/用/微波炉/加热/,/不如/把/食物/放在/蒸锅/上/./"/
/澳大利亚/"/新/家庭主妇/"/女性/社团/
/"/每天/少看/一/小时/电视/,/每个/月/就/省/下/30/个/小时/./IBM/可以/用/30/小时/设计/一个/硬件/,/斯皮尔伯格/可以/用/30/小时/修改/完/一个/剧本/,/Gucci/可以/用/30/小时/完成/一张/600/万澳元/订单/./算算看/,/电视/让/你/失去/了/多少/财富/?/"/

Pour T_ONG :

大气/十条/”/虽/出台/，/GDP/增长/仍/是/王道/
/中外/对话/：/您/如何/评价/“/大气/十条/”/作为/一个/全国性/专项/政策/的/力度/？/
/ /李俊/峰/：/不能/说/没有/力度/，/但是/还/需要/增加/力度/。/这/主要/体现/在/三个/方面/。/
/ /一是/，/对/严重性/的/认识/不够/。/根据/国家/环保部/公布/的/资料/，/我国/的/主要/城市/空气/达标/的/不到/10/%/，/京津冀/、/长三角/尤为/严重/，/已经/影响/到/人民/的/健康/、/投资者/的/热情/。/
/ /二是/，/突出/主要矛盾/不够/。/说/是/大气/十条/，/但是/洋洋洒洒/，/有/很多/具体措施/，/但是/减少/或是/控制/煤炭/消费/、/提高/油品/质量/是/主要矛盾/，/这/两个/方面/的/力度/还/不够/。
/ /三是/，/对/自身/的/能力/的/自信/度/不够/。/中国/在/过去/很多/联防/联控/的/成功经验/，/也/有/1/年/增加/500/亿/天然气/、/2000/万千瓦/水电/、/1800/万千瓦/风电/的/能力/，/控制/阴霾/主要/靠/优化/能源/结构/，/这个/又/有/自信/，/有/信心/。
/ /中外/对话/：/也就是说/，/“/大气/十条/”/在/具体/的/政策措施/上/，/有/内在/矛盾/？

3.7.2.5 Résultats avec le segmenteur Hylanda

Les résultats suivants avec le segmenteur Hylanda ont été obtenus par le logiciel de ce même nom. Les couleurs sur les deux textes représentent les catégories grammaticales annotées par ce logiciel.

Pour T_Presse :

关注 " 不 插 电 " 生活 , 是 为了 让 大家 了解 这种 特殊 的 生活 方式 , 如同 透过 万花筒 一样 , 看到 世界 上 原来 还有 一些 人 过 着 这样 的 生活 . 至于 如何 看待 这种 生活 方式 , 是否 会 遵循 这种 生活 方式 , 则 是 仁者见仁 智者见智 了 .
他们 提倡 " 不 插 电 "
" 放弃 电动跑步机 吧 , 你 不 觉得 在 平地 上 跑步 更 舒服 吗 ? "
悉尼 大学 " 健康 运动 " 俱乐部
" 挤 公交 不但 省钱 , 更是 一项 集散 打 , 平衡木 和 瑜伽 于 一体 的 运动 , 自己 开车 哪 有 这种 好处 ? "
印度 孟买 " 绿色 环境 " 俱乐部
" 微波炉 是 20 世纪 最 不该 出现 的 发明 , 它 改变 了 生物 分子结构 , 下 一步 就是 改变 人体 生理 结构 . 与其 用 微波炉 加热 , 不如 把 食物 放在 蒸锅 上 . "
澳大利亚 " 新 家庭主妇 " 女性 社团
" 每天 少 看 一小时 电视 , 每个 月 就 省 下 30 个 小时 . IBM 可以 用 30 小时 设计 一个 硬件 , 斯皮尔伯格 可以 用 30 小时 修改 完 一个 剧本 , Gucci 可以 用 30 小时 完成 一张 600 万澳元 订单 . 算算 看 , 电视 让 你 失去 了 多少 财富 ? "

Pour T_ONG :
" 大气 十条 " 虽 出台， GDP 增长 仍 是 王道
中外 对话： 您 如何 评价 " 大气 十条 " 作为 一个 全国性 专项 政策 的 力度？
李俊峰： 不 能 说 没有 力度， 但是 还 需要 增加 力度。 这 主要 体现 在 三个 方面。
一 是， 对 严重性 的 认识 不够。 根据 国家 环保部 公布 的 资料， 我国 的 主要 城市 空气 达标 的 不到 10%， 京津冀、 长三角 尤为 严重， 已经 影响 到 人民 的 健康、 投资者 的 热情。
二 是， 突出 主要矛盾 不够。 说 是 大气 十条， 但是 洋洋洒洒， 有 很多 具体 措施， 但是 减少 或是 控制 煤炭 消费、 提高 油品 质量 是 主要矛盾， 这 两个 方面 的 力度 还 不够。
三 是， 对 自身 的 能力 的 自信度 不够。 中国 在 过去 很多 联防 联控 的 成功 经验， 也 有 1 年 增加 500 亿 天然气、 2000 万千瓦 水电、 1800 万千瓦 风电 的 能力， 控制 阴霾 主要 靠 优化 能源 结构， 这个 又 有 自信， 有 信心。
中外 对话： 也就是说， " 大气 十条 " 在 具体 的 政策 措施 上， 有 内在 矛盾？

3.7.2.6 Commentaires des résultats

Nous obtenons des résultats différents selon le segmenteur utilisé. Cependant, les différences sont relativement minimes. Le nombre de formes différentes obtenues relevées est donné dans les deux tableaux ci-dessous :

T_Presse	ICTCLAS	Stanford	Jieba	Hylanda
ICTCLAS	0	6	4	6
Stanford	6	0	5	10
Jieba	4	5	0	6
Hylanda	6	10	6	0

T_ONG	ICTCLAS	Stanford	Jieba	Hylanda
ICTCLAS	0	5	10	3
Stanford	5	0	10	5
Jieba	10	10	0	10
Hylanda	3	5	10	0

T_Presse

Pour T_Presse qui comprend 360 mots, nous obtenons au plus 10 formes différentes entre les quatre segmenteurs (ICTCLAS, Stanford, Jieba et Hylanda). Ces différences portent parfois sur des découpages de 这种/zhè zhǒng/cette sorte de. Il apparaît ainsi les deux découpages suivants :

- 这种
- 这/种

Ces deux différents découpages ne viennent pas modifier le sens de 这种 /zhè zhǒng/cette sorte de. D'un point de vue grammatical, 种 (zhǒng) est un classificateur, c'est-à-dire un caractère qui suit en principe un article ou un nombre afin de spécifier la catégorie du mot rattaché à cet article. Il existe en chinois un nombre important de classificateurs. A titre d'exemple, le classificateur 位 (wèi) est employé pour les personnes, comme le montre la phrase suivante :

这位老师很好/zhè wèi lǎoshī hěn hǎo/ Cet enseignant est sympathique.

Un autre exemple de formes différentes rencontrées n'impactant pas le sens du texte est donné par les deux découpages suivants :

- 每天 (měitiān), chaque jour
- 每/天 (měi/tiān), chaque / jour

On peut aussi relever une segmentation différente d'une expression idiomatique : 仁者 见仁 智者见智 (rénzhě jiàn rén, zhìzhě jiàn zhì) qui signifie littéralement : Le bienfaiteur voit du bienfait, le sage voit de la sagesse. Cette expression peut se rapprocher de l'idée selon laquelle différents points de vue sur une même situation sont acceptables. Nous obtenons les découpages suivants de cette expression ;

- 仁者/ 见仁/ 智者见智 (ICTCLAS)
- 仁者/ 见仁/ 智者/ 见智/ (Stanford)
- 仁者见仁/智者见智/ (Jieba)
- 仁者见仁/智者见智/ (Hylanda)

Ces différences de segmentation peuvent s'expliquer par la présence ou la non présence de ces expressions idiomatiques dans le dictionnaire utilisé par le segmenteur. Les expressions idiomatiques ne produisant un sens que globalement à partir des caractères les composant, il paraît plus pertinent dans le processus de segmentation de traiter le groupe des quatre caractères qui les composent comme une unité de segmentation, voire dans ce cas précis où une seule expression est formée de huit caractères. Seuls Jiaba et Hylanda étaient capables de segmenter cette expression correctement.

On relève aussi les formes différentes suivantes :

- 还有 (hái yǒu) , il y a aussi
- 还 | 有 (hái/yǒu) , encore/avoir

Les mots 还 et 有 ayant des natures différentes (respectivement adverbe et verbe), le découpage le plus pertinent semble être celui qui sépare les termes 还 et 有.

On relève aussi deux découpages différents : 跑步|机/pǎobù|jī/courir | machine et 跑步机 /Pǎobùjī/tapis de course. La deuxième segmentation est bien entendu la plus pertinente d'un point de vue sémantique. En revanche, si on se positionne sur le plan de la morphologie complexe ou composée, la première segmentation est exacte.

Pour T_Presse, nous constatons donc que les quatre différents segmenteurs proposent des résultats proches. Ces résultats se rapprochent du sens initial du texte.

T_ONG

Sur les 386 mots de T_ONG, on obtient au plus 10 formes différentes parmi les quatre segmentations proposées. La segmentation est donc presque identique. Il y a cependant des divergences au niveau de certains mots. Ainsi, on relève les deux segmentations suivantes :

- 十条 (shítiáo) signifiant plan, mesures
- 十 | 条 (shí/tiáo) signifiant 10 clauses (unités).

La première segmentation semble davantage proche du sens du texte qui évoque un plan de lutte contre la pollution atmosphérique.

Nous relevons également les différences de formes suivantes :

- 不能 (bù néng), ne pas pouvoir
- 不 | 能 (bù / néng), ne pas / pouvoir

renvoyant toutes deux au même sens : ne pas pouvoir. Ces deux segmentations ont le même sens. La différence se situe en fait au niveau de l'analyse grammaticale. Dans la première segmentation, la négation 不/bù et le verbe 能/néng apparaissent dans le même segment et constituent donc à eux deux une unité de segmentation. Dans la seconde segmentation, 不 (bù) et 能 (néng) constituent deux segments distincts.
Ce même type d'interprétation grammaticale se retrouve avec les formes différentes :

- 一是 (yī shì) tout d'abord, premièrement, primo
- 一 | 是 (yī/shì) un / être

Il apparaît plus pertinent, d'un point de vue sémantique, de privilégier la première segmentation 一是 (yī shì) dans la mesure où 一是 (yī shì) a ici un sens bien défini.

Des formes différentes apparaissent aussi avec les classificateurs. Ainsi, nous relevons les deux segmentations suivantes, suivant toutes deux le mot *un*:

- 一个 (yī gè) un
- 一 | 个 (yī / gè) un / classificateur

Dans la deuxième segmentation, le classificateur 个 est isolé.

D'un point de vue lexical, on note aussi les deux segmentations différentes suivantes :

- 政策/措施 (zhèngcè / cuòshī) politique / mesure
- 政策措施 (zhèngcècuòshī) mesure politique

Les deux segmentations produisent le même sens.

Il est en outre à noter qu'aucun des quatre segmenteurs n'a su reconnaître les entités nommées chinoises suivantes :

- 中外对话 (zhōngwài duìhuà), *Chinadialogue*, le site bilingue indépendant,
- 大气十条 (dàqì shítiáo), Dix grandes mesures politiques défendant la qualité de l'air.

A travers cette expérience de *benchmarking* de segmenteurs chinois, ICTCLAS, Stanford et Jieba sont des logiciels libres et commodes à manipuler. Nous avons donc choisi ICTCLAS pour la segmentation d'ENRG_CN en raison de son accessibilité, et Jieba pour celle de CLRG_CN grâce à sa nouveauté[134] et sa manipulation aisée.

Conclusion sur les segmenteurs automatiques

A l'issue de ces comparaisons, nous tirons les conclusions suivantes : les résultats fournis par Jieba et Hylanda sont similaires à seulement une différence près. Ces deux segmenteurs privilégient les segments de mots formant une unité lexicale longue et complète. Cette pratique permet de préserver la sémantique des expressions. Cependant, Stanford et ICTCLAS coupent les mots en unités fines, c'est-à-dire, en petite unité lexicale et grammaticale. Ce fin découpage facilite la recherche des formes dans la fouille de textes ou d'informations, mais il apporte souvent des difficultés d'interprétation de sens et du bruit. La segmentation fine engendre des conséquences pécuniaires et de temps qui ne sont pas négligeables.

Dans le but d'efficacité, nous préconisons l'utilisation de Jieba.

Conclusion du chapitre

Ce chapitre nous a donc permis de mieux cerner les trois langues (français, anglais et chinois) des textes de notre corpus. Nous avons ainsi vu qu'elles sont très différentes. L'anglais se caractérise en effet par une grande richesse phonologique et est souvent mieux adapté que le français aux publications techniques et scientifiques, du fait de sa structure syntaxique. Le français se caractérise par une orthographe essentiellement historique et contenant de nombreuses exceptions. Cette orthographe constitue l'une des principales difficultés dans l'apprentissage de cette langue par des locuteurs étrangers. En ce qui concerne la langue chinoise, son système d'écriture est le fruit d'une longue histoire. L'existence de mots inclus dans un autre mot (par exemple les chaînes AB et ABC sont des mots), l'invariance des verbes, et l'absence de genre et de nombre sont autant de difficultés pour réaliser une segmentation du chinois. Après avoir rappelé les différents codages des caractères chinois existants et certaines méthodes statistiques et informatiques telles que l'algorithme de recherche dichotomique, le modèle de Markov caché, nous avons fait une présentation de quelques outils de segmentation du chinois disponibles sur le marché et fondés sur ces méthodes. Nous les avons ensuite mis en oeuvre sur certains textes de notre corpus. Les résultats obtenus se sont révélés globalement similaires. Des différences apparaissent cependant entre autres du fait de la reconnaissance ou de la « non reconnaissance » des expressions idiomatiques, des entités nommées ou encore de la gestion du classificateur 个/gè/classificateur isolé ou non. Le chapitre 4 sera consacré au mode de récupération des données.

[134] Module (Jieba) de segmentation libre, disponible en 2012

4. Constitution des corpus

Le chapitre 4 aura pour objet d'expliciter le mode de constitution de nos corpus. Nous commencerons par justifier le choix des quatre principales sources de notre corpus, qui sont :

- le journal *Le Monde*
- le *New York Times*
- le site internet *qq.com*
- le site internet *sina.com.cn*

Nous vérifierons ensuite la comparabilité de ce corpus comparable ENRG qui se décline en trois sous-corpus :

- ENRG_FR : sous-corpus de textes français concernant l'énergie et l'environnement.
- ENRG_US : sous-corpus de textes américains concernant l'énergie et l'environnement.
- ENRG_CN : sous-corpus de textes chinois concernant l'énergie et l'environnement.

A la fois d'un point de vue qualitatif et quantitatif, les critères de comparaison retenus sont :

- le nombre d'occurrences,
- le nombre de formes,
- le nombre d'articles,
- le nombre de paragraphes,
- les continuités ou discontinuités sur l'axe du temps,
- les rubriques d'apparition dans le corpus,
- les séries chronologiques par année dans la poly-résonance textuelle et la résonance événementielle.

La suite du chapitre sera consacrée à la présentation du corpus parallèle bilingue chinois anglais concernant le changement climatique, corpus dénommé CLRG et pouvant se découper comme suit :

- volet CLRG_EN, EN pour le Royaume-Uni et le monde anglophone,
- volet CLRG_CN, CN pour la Chine.

Les textes de ce corpus parallèle seront issus du site *chinadialogue.net*.

La méthode de l'analyse factorielle des correspondances, dont nous rappellerons les principes fondamentaux, constituera un moyen de réaliser une partie de l'analyse quantitative de comparabilité de nos corpus.

4.1 Constitution de corpus de veille, corpus multilingues thématiques

Deux types de corpus[135] ont été réalisés pour cette recherche à savoir :

- le corpus trilingue de veille à caractère comparable : français, anglais (américain) et chinois, nous partirons du postulat que chacune des trois langues formera un sous-corpus à part entière.
- le corpus parallèle bilingue pour la veille : anglais et chinois, sera traité dans le chapitre 7. Rappelons que les différentes parties du corpus parallèle, source *versus* cible, se désignent par le terme volet à l'instar du Chapitre 1, section 1.4. Nous parlerons ci-après, du volet anglais et du volet chinois du corpus parallèle.

4.2 Corpus trilingue de veille : un corpus comparable

Les trois sous-corpus de veille à caractère comparable en trois langues différentes sont constitués à partir d'articles de journaux et de médias disponibles en ligne à savoir, la rubrique *Planète* du *Monde*, la rubrique *Environment* du *New York Times* et la rubrique *Vert* des divers médias chinois.

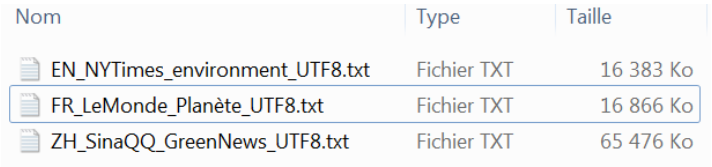


Figure 4.1 Corpus trilingue à caractère veille comparable

4.2.1 Le sous-corpus français : *Le Monde*

Le Monde est l'un des rares journaux français réputé donnant accès gratuitement à une grande partie de ses articles aux lecteurs non-abonnés, mais pas à l'intégralité.

Lors des extractions automatiques des articles français en 2012, la rubrique *Planète* du *Monde* ne distinguait pas les sujets thématiques évolutifs aussi finement qu'aujourd'hui, sujets évoluant à la cadence de l'actualité (figure 4.2).

Figure 4.2 Présentation de la rubrique Planète du Monde, consulté le 24/06/2015

4.2.2 Le sous-corpus américain : *New York Times*

La cote de popularité et la qualité rédactionnelle de ce journal dans le monde anglophone américain incitent indéniablement à le sélectionner comme échantillon de travail. Les articles du journal sont consultables en libre accès en ligne dans leurs intégralités, mais leurs extractions sont verrouillées, en particulier pour les robots d'extraction automatique[136].

[135] Ces trois sous-corpus ont été constitués à partir de programmes conçus en Perl, PHP, Ajax, Jquery, Python et C++. Les sous-corpus sont stockés dans des fichiers au format texte brut en UTF8, encodage permettant d'afficher la plupart des langues vivantes du monde, dont le chinois.

[136] En raison de cette difficulté, des programmes spécifiques en langage Perl ont été mobilisés lors de la constitution du sous-corpus anglophone.

L'homologue de la rubrique *Planète* dans le *New York Times* n'existe pas. *Environment*, une sous-catégorie de la colonne *Science* (figure 4.3) serait le thème le plus proche de la rubrique *Planète* dans nos études.

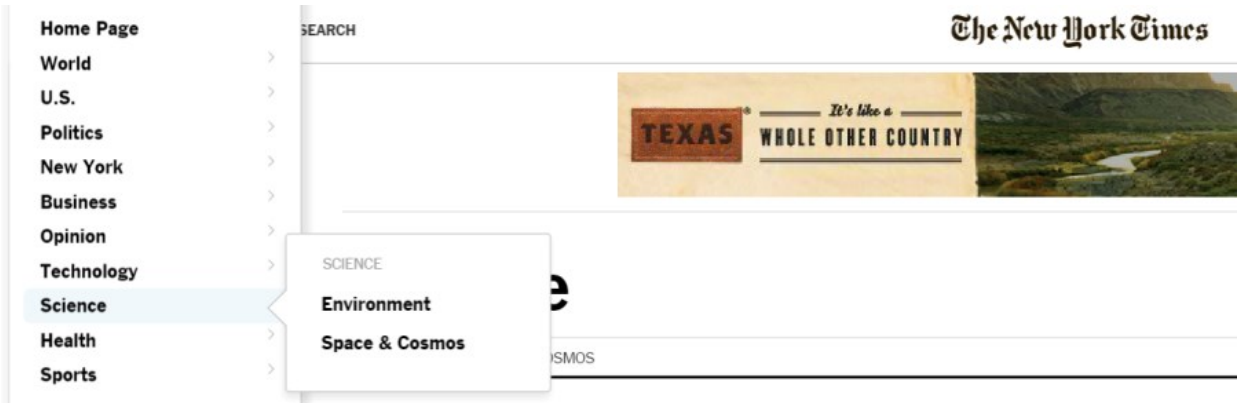


Figure 4.3 Présentation de la sous-rubrique *Environment* du *New York Times*, consulté le 24/06/2015

La classification des articles dans le journal NYT est différente du journal le Monde. Cette classification entraîne d'une part des traitements plus complexes de navigation dans l'arborescence, d'autre part, cette classification souligne la différence dans l'approche intellectuelle de la notion d'environnement entre les deux journaux.

4.2.3 Le sous-corpus chinois : *QQ* et *Sina*

Pour la plupart des journaux chinois, les sources ne sont pas en libre accès. Nous avons décidé de recourir aux sites d'informations très connus et généralistes pour nous procurer les sources de nos corpus. Ces portails informationnels participent à la vie quotidienne de la presse chinoise en publiant leurs propres articles et mettent à la disposition en parallèle les articles des différents quotidiens chinois en fonction de leurs propres choix.

Le nouveau classement d'*Alexa*[137] montre que *qq.com* est le 2ème site le plus populaire juste après *baidu.com* au sein de la population chinoise continentale et *sina.com.cn* le 4ème. Au rang mondial, des sites les plus consultés, *qq.com* se place 10ème et *sina.com.cn* le 13ème.

QQ.com

Figure 4.4 Présentation de la rubrique Vert du QQ.com, consulté le 27/06/2015

Pour la petite anecdote, le nom original de QQ provient de l'acronyme *OICQ*, signifiant *OpenICQ*. Cependant, dans les années 90, en raison de conflit possible avec une autre messagerie instantanée très populaire (*ICQ*), le nom a donc été modifié en QQ.

[137] Alexa Internet est une entreprise basée à San Francisco, fondée en avril 1996. Elle appartient au groupe Amazon. Son site Web est principalement connu pour fournir des statistiques sur le trafic du Web mondial. URL : http://www.alexa.com/topsites/ (site consulté le 23/06/2015)

La figure 4.4 présente la rubrique *Vert* du site :

1. URL du site : http://green.news.qq.com/
2. Principales rubriques du site, de gauche à droite : News (新闻/xīn wén), Vidéos (视频/shì pín), Images (图片/tú piàn), Chroniques (评论/píng lùn), Finances et Economie (财经/cái jīng), Bourse (股票/gǔ piào), etc.
3. Bandeau vert regroupant les différents sujets, de gauche à droite : Page d'accueil (首页/shǒu yè), News vertes (新闻/xīn wén), Visions vertes (绿问/lǜ wèn), Thèmes spéciaux (专题/zhuāntí), Partages verts (绿色分享会/lǜsè fēnxiǎng huì), Industries (产业/chǎnyè), Commentaires verts (绿评/lǜ píng), Palmarès des Propres et des Pollueurs (环境红黑榜/huánjìng hóng hēi bǎng), Changement climatique (气候变化/qìhòu biànhuà), Voie écologique pour les entreprises (企业绿色之道/qǐyè lǜsè zhī dào), Evénements(活动/huódòng), Recommandations (推荐/tuī jiàn), Photothèque (图库/tú kù), Groupes de chats/Twitters verts (绿色微博群/lǜsè wēi bó qún).

Figure 4.5 Extrait de la page des articles, rubrique Vert du QQ.com, consulté le 27/06/2015

En cliquant sur le sujet 新闻/xīnwén/*News vertes*, nous entrons dans la page où sont agrégés les divers articles relatifs aux *News vertes* (figure 4.5) :

1. URL de la page des articles : http://news.qq.com/l/green/list20100329114428.htm
2. Position dans le site : vous êtes dans la rubrique *Vert* du site.
3. Zone permettant la navigation dans les différentes pages où sont agrégés les articles, 10 paquets de 5 articles sont rangés par page, le nombre total des pages est 160. Cette rubrique propose 8 000 articles environ.
4. Zoom d'un paquet de 5 articles.

Nous notons que cette rubrique *Vert* du site *QQ.com* n'est plus mise à jour depuis le 22 octobre 2013. Certaines de ces données abordant le thème « Protection de l'environnement », datées du 03 août 2010 au 17 mai 2013, ont été archivées à l'adresse suivante :
http://news.qq.com/l/green/conservation/list201008393358.htm

Sina.com.cn

Figure 4.6 Présentation de la rubrique Protection de l'environnement du sina.com.cn, consulté le 27/06/2015

La figure 4.6 présente la rubrique *Protection de l'environnement* du site :

1. URL du site : http://green.sina.com.cn/
2. Principales rubriques du site (tableau 4.1 ci-dessous, à lire de gauche à droite)
3. Bandeau vert regroupant les différents sujets, de gauche à droite : News vertes (新闻/xīn wén), Zoom vert (绿镜/lǜ jìng), Interviews spéciaux (专访/zhuān fǎng), Commentaires verts (评论/píng lùn), Photothèque (图集/tú jí), Vidéos (视频/shìpín), thèmes spéciaux (专题/zhuāntí), Evènements (活动/huódòng), Industries vertes (绿色产业/lǜsè chǎnyè), Actions écologiques pour les entreprises (企业行动/qǐyè xíngdòng), Mode de vie à densité carbonique faible (低碳生活/dī tàn shēnghuó), Produits à basse consommation d'énergie (节能上品/jiénéng shàngpǐn), Micro-interviews (微访谈/Wēi fǎngtán), Actualités des ONG (NGO 动态 /NGO dòngtài), Lancer un événement (发起活动/fāqǐ huódòng).

Tableau 4.1 Extrait des principales rubriques (partie 2 de la figure 4.6) du *site sina.com.cn*

新闻 /xīn wén/ News	军事/jūnshì/ Défense	社会/Shèhuì/ Société	健康/jiànkāng / Santé	体育/tǐyù/Sport	(…)	导航/dǎoháng /Plan du site
财经 /cái jīng/ Finances et Economie	股票/gǔpiào/ Bourse	基金/jījīn/ Fonds d'investissements	理财/lǐcái/ Gestions d'épargne	科技/kējì/ Sciences et technologie	(…)	邮箱/yóuxiāng /Mail

Figure 4.7 Extrait de la page des articles, rubrique Protection de l'environnement du sina.com.cn, consulté le 27/06/2015

En cliquant sur le sujet *News vertes* (新闻/xīn wén), nous entrons dans la page où sont agrégés les divers articles relatifs aux *News vertes* (figure 4.7) :

1. URL de la page des articles : http://roll.green.sina.com.cn/green/hb_gdxw/index.shtml
2. Position dans le site : vous êtes dans la rubrique *News vertes* du site.
3. Zone permettant la navigation dans les différentes pages où sont agrégés les articles, 40 articles sont rangés par page.
4. Champ de sélection pour accéder aux archives des articles.
5. Zoom de 5 articles.

Les *news* de cette rubrique sont mises en ligne de manière extrêmement discontinue depuis le 15 janvier 2009 (selon le site), cependant le premier article disponible et détecté par le programme spécifique d'extraction automatique date du 18 juillet 2009. A ce jour, la rubrique *Protection de l'environnement* du *sina.com.cn* est toujours alimentée.

Tous les articles de ces deux sites sont accessibles dans leur intégralité (d'où ce choix comme principales sources d'extraction). La Chine a connu une croissance phénoménale de développement des ressources numériques en ligne depuis les années 2008-2010 grâce à l'organisation des Jeux Olympiques à Pékin en 2008 et l'Exposition Universelle de Shanghai en 2010. De nombreuses productions de pages web envahissent alors la toile. Ceci se traduit par la disponibilité des pages web récupérés par nos programmes d'extraction automatique[138].
Les détails des outils et programmes informatiques conçus pour les collectes des données textuelles sont expliqués dans l'annexe M.

Afin de cerner les principales caractéristiques du corpus comparable, un dépouillement général des trois sous-corpus par année a été accompli et exposé dans l'annexe B. Ce dépouillement nous révèle les informations textométriques suivantes.

4.2.4 Caractéristiques textométriques du corpus comparable trilingue ENRG

Dénomination générique du corpus comparable de veille trilingue

Pour faciliter la lecture des tableaux et des figures ci-après, la dénomination du corpus comparable trilingue contenant les trois sous-corpus, français, américain et chinois, devient la suivante :

- Nom du corpus de veille trilingue : **ENRG** pour énergies et environnement.
- Nom des trois sous-corpus[139]:

  ENRG_FR, FR pour la France
  ENRG_US, US pour les Etats-Unis
  ENRG_CN, CN pour la Chine.

[138] Ces programmes sont écrits en langage Perl, PHP, Python et C++.
[139] Codes des pays selon la norme ISO 3166, forme courte du nom en langue française ou anglaise.

Tableau 4.2 ENRG_FR, ENRG_US et ENRG_CN : analyse des caractéristiques textométriques du corpus ENRG

Sous-corpus	ENRG_FR	ENRG_US	ENRG_CN
Nombre d'occurrences	2 783 702	2 735 535	10 447 521
Nombre de formes	76 848	66 966	98 699
Nombre d'articles	4 817	3 993	14 514
Nombre de paragraphes	43 333	59 452	221 569
Période	du 24-09-1999 au 17-04-2012	du 26-01-2005 au 18-04-2012	du 23-03-2008 au 23-04-2013
Rubrique du site	Planète	Environnement (sous-rubrique)	Vert
Continuité ou discontinuité (sur l'axe du temps)	en continu sur les mois	en continu sur les mois	en discontinu sur les mois
Séries chronologiques par année dans la résonance événementielle (par AFC)	(2005, 2006) (2007, 2008) (2011, 2012)	(2005, 2006) (2007, 2008) (2011, 2012)	aucune par année
Années retenues au final	2010, 2011 et 2012 (partiel)	2010, 2011 et 2012 (partiel)	2010, 2011 et 2012

4.2.5 Comparabilité qualitative et quantitative

Selon le tableau 4.2 ci-dessus et l'annexe B, nous pouvons comparer les caractéristiques qualitatives et quantitatives entre ENRG_FR et ENRG_US, tant sur le plan chronologique que textuel, car il s'agit de deux rubriques relativement similaires, provenant de deux journaux distincts connus et reconnus par le monde occidental (critères qualitatifs) et rédigés dans des langues indo-européennes, ainsi que leurs nombres d'occurrences, de formes, et d'articles (critères quantitatifs) également proches.

En revanche, ENRG_CN s'avère plus complexe par sa composition et sa taille : d'une part les deux rubriques similaires, *Planète* et *Environnement* (sources de corpus), sont absentes dans les deux plateformes chinoises retenues. D'autre part, les deux sites *QQ.com* et *Sina.com.cn* recensent quotidiennement les articles de divers journaux chinois à travers la toile et publient leurs propres rédactions. Cela explique l'important volume de données résultantes.

Le volume d'articles des ENRG_FR et ENRG_US est relativement homogène sur l'ensemble de la période (annexe B, figure B2 et figure B17), tandis que celui d'ENRG_CN est presque quatre fois supérieur aux deux autres (chapitre 6, figure 6.1).

Typologie annuelle du sous-corpus français

Le graphique de la figure 4.8 (ci-dessous) donne la représentation des proximités de chacune des années sur le plan factoriel du sous-corpus français, entre 1999 et 2012 de la rubrique « Planète » du journal le Monde.

Selon cette figure, il y a, d'une part, une nette séparation entre les années 1999-2009, à gauche de l'axe vertical et les années 2010-2012, à droite de ce même axe, d'autre part, nous voyons apparaitre approximativement trois groupements d'années et trois années extrêmes en rupture avec ces trois groupements. Cela signifie que chacun des trois groupements partage respectivement des formes communes dans l'utilisation de leur vocabulaire. Les périodes extrêmes (1999, 2004 et 2010) seront écartées.

Après l'observation de cette Analyse Factorielle des Correspondances, plusieurs constats en résultent :

1. Trois années extrêmes : 1999, 2004 et 2010 se détachent, s'écartent des groupements et deviennent par leurs vocabulaires spécifiques, les perturbateurs de la chronologie.
2. Des chronologies ont été plus ou moins conservées dans chacun de ces trois groupements, à savoir,
   - Groupement N° 1 : 2000, 2001, 2002, 2003, 2005 et 2006
     Un constat immédiat s'impose avec une proximité plus importante entre 2003 et 2005, mais 2004 est absente et devient le perturbateur de la chronologie.
   - Groupement N° 2 : 2007, 2008 et 2009
     L'année 2009 est en rupture avec les deux autres années, séparée par l'axe horizontal.
   - Groupement N° 3 : 2011 et 2012
     Ce groupement reste dans une dimension individuelle à cause de leurs vocabulaires spécifiques.

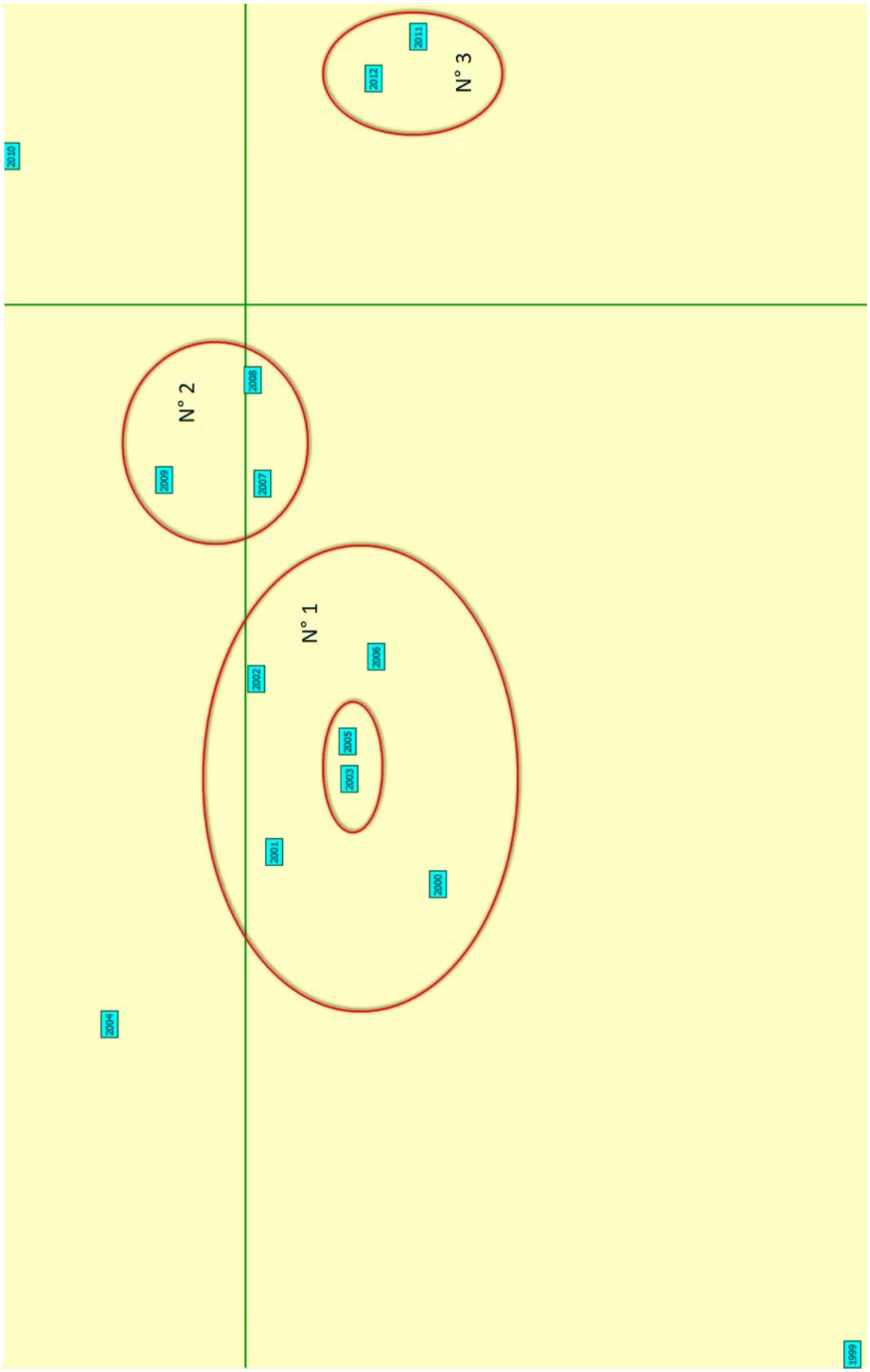


Figure 4.8 ENRG_FR de 1999 à 2012 : analyse factorielle des correspondances sur l'ensemble des années

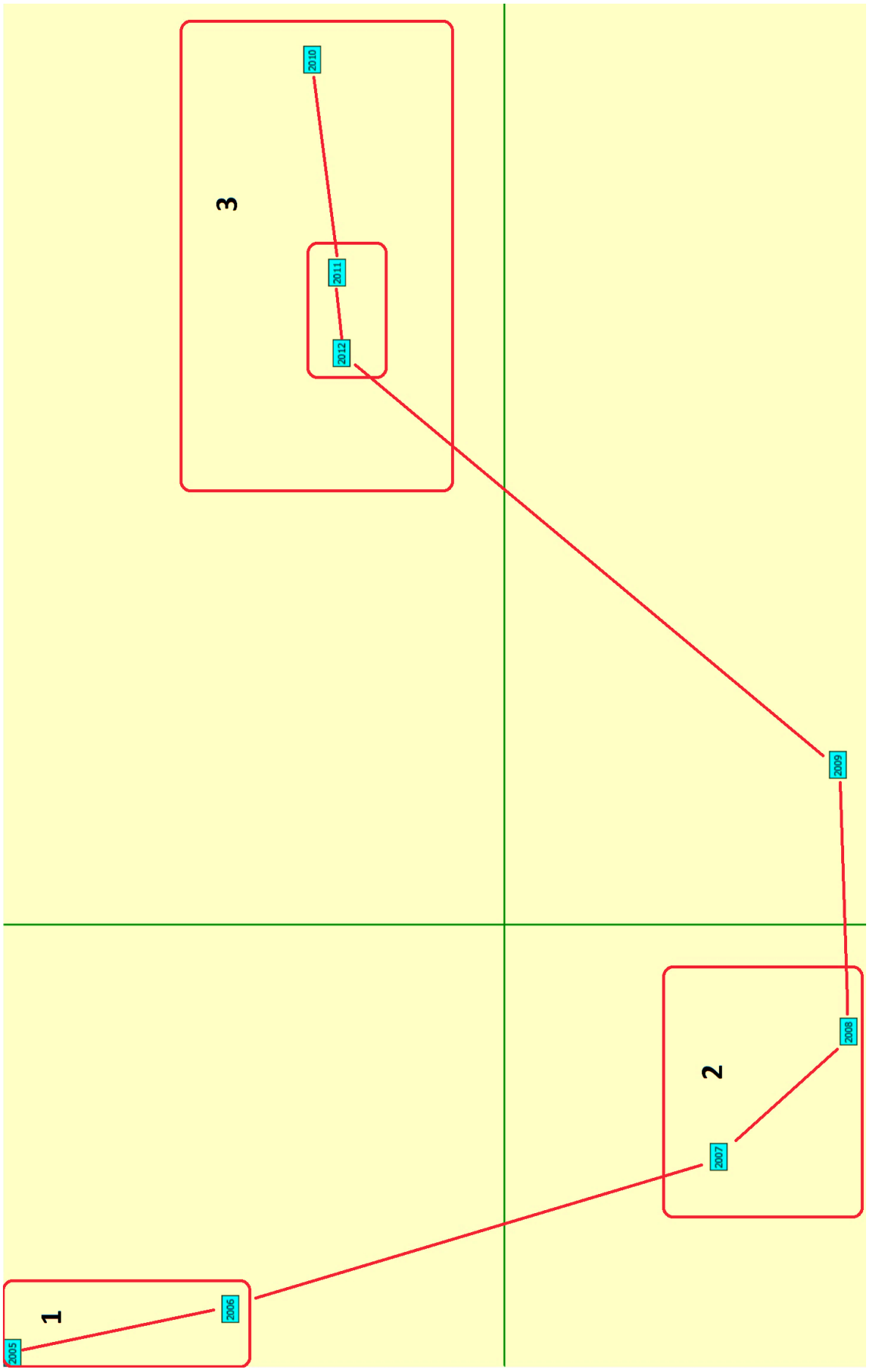


Figure 4.9 ENRG_US de 2005 à 2012 : analyse factorielle des correspondances sur l'ensemble des années

Le graphique de la figure 4.9 (ci-dessus) met clairement en évidence la proximité relative de chacune des années sur le plan factoriel, entre 2005 et 2012. Dans l'objectif de faciliter l'analyse, nous avons annoté la figure B.21 de l'annexe B (Dépouillement général du corpus comparable) par leurs groupements textuels et par la série chronologique.

D'une part, il y a une nette séparation entre les années 2005, 2006, 2007 et 2008, à gauche de l'axe vertical et les années 2009, 2010, 2011 et 2012, à droite de ce même axe. Les années sont relativement éloignées les unes des autres mises à part les années 2011 et 2012. D'autre part, comme pour le sous-corpus français, nous voyons apparaitre approximativement, 3 groupements d'années et 1 année extrême en rupture avec ces 3 groupements. Cela signifie que chacun des 3 groupements partage, *grosso modo*, respectivement des formes communes dans l'utilisation de leur vocabulaire. La période extrême (2009) écartée, se singularise dans une dimension individuelle, en rupture avec les 3 groupements.
Après l'observation de cette Analyse Factorielle des Correspondances, plusieurs constats en résultent :

1. Une année extrême : 2009 se singularise par son vocabulaire spécifique (tableau B.12) et devient l'un des perturbateurs de la série chronologique.
2. Une série chronologique apparait en 2010 se trouvant à la fin de la courbe.
3. Les 3 groupements
    - Groupement 1 : 2005 et 2006
      L'écart est relativement important entre ces deux années.
    - Groupement 2 : 2007 et 2008
      L'écart est relativement important entre ces deux années.
    - Groupement 3 : 2010, 2011 et 2012
      Ce dernier présente la particularité d'une série chronologique inversée, l'année 2010 étant à la fin de la série chronologique et éloignée du reste de ce groupement. De ce fait, 2010 est également un perturbateur de la série chronologique.

4.2.6 Perturbateurs de la chronologie pour ENRG_FR et ENRG_US

L'Analyse Factorielle des Correspondances (ci-après nommée AFC) d'ENRG_FR nous révèle que les années 2010-2012 se singularisent par leur proximité par rapport au reste des années (figure 4.8). Les perturbateurs, les années 1999 et 2004 se trouvent très éloignés des restes des groupements. Quant à ceux d'ENRG_US, une parabole d'une série chronologique sur l'ensemble de la période se forme (figure 4.9), mieux encore, les années 2010, 2011 et 2012 restent groupées.

En conclusion, l'étude comparative ne portera que sur ces trois dernières années.

Quant à la périodicité, les analyses transversales (tableau 4.2 ci-dessus et l'annexe B, dépouillement général du corpus comparable) de chaque sous-corpus et la dissymétrie chronologique montrent que :

- pour ENRG_FR et ENRG_US, la période commune la plus large recouvre les années 2005 à 2012,
- pour ENRG_CN (se reporter au chapitre 6), les articles de nos thèmes de recherche sont quasi-absents pour les années 2008 et 2009.

Donc, la seule période commune des trois sous-corpus respectant la comparabilité est de 2010 à 2012.

4.3 Corpus parallèle bilingue anglais et chinois pour la veille

Corpus parallèle : Chinadialogue.net

Le site *Chinadialogue.net* est lancé le 3 juillet 2006 par une organisation indépendante éponyme à but non lucratif basée à Londres et à Beijing, financé par une gamme de partenaires institutionnels, y compris plusieurs grandes fondations caritatives. Il apparaît comme l'un des premiers sites bilingues anglais-chinois du monde consacrés à l'environnement. Cette structure tente d'établir un dialogue et de rechercher les solutions relevant les défis environnementaux en publiant des articles bilingues entre la Chine et le monde occidental.

La voix des ONG participe indiscutablement à l'équilibre des opinions internationales aux sujets de l'environnement et favorise la mise en contraste des visions de l'ensemble des pays contribuant au développement durable de l'économie mondiale, d'où le choix de ce site.

Figure 4.10 Présentation du site www.chinadialogue.net, consulté le 07/08/2015

La figure 4.10 présente l'accès à la rubrique *Climate change & Energy*, deux thèmes classés dans le même sujet. URL : www.chinadialogue.net/topics/climate+change+&+energy

Tous les articles de ce site sont accessibles dans leurs intégralités en lecture bilingue. Cependant, certaines traductions d'articles ne sont pas accessibles en même temps que les originaux. Les pages web bilingues sont récupérées par nos programmes d'extraction automatique écrits en Python. La présentation des programmes se trouve dans l'annexe M.

Les problèmes rencontrés lors de la constitution du corpus parallèle et ses spécificités seront abordés dans le Chapitre 7.

La dénomination générique pour le corpus parallèle anglais - chinois (traité dans le chapitre 7) devient :

- Nom du corpus de veille bilingue : **CLRG** pour ***Climate change & Energy*** (Changement climatique et Energie).
  CLRG_EN, EN pour le Royaume-Uni et le monde anglophone.
  CLRG_CN, CN pour la Chine.

4.4 Évaluations des périodes de nos corpus

Rappelons que nous avons construit deux corpus à partir des articles de journaux et des ONG issus de nos trois sphères de communication, un corpus comparable, appelé ENRG et un corpus parallèle, nommé CLRG, couvrant une période de 16 années, de 1999 à 2014, ce qui nous a permis d'étudier trois notions, énergie(s), nucléaire, EPR, en trois langues, français, anglais et chinois. La figure 4.11 ci-dessous montre les différentes périodes couvertes par les médias retenus.

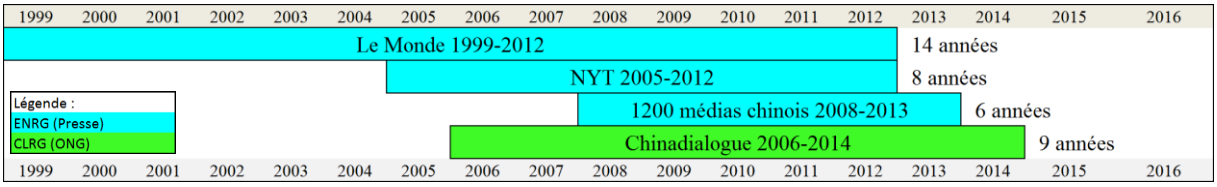


Figure 4.11 ENRG et CLRG dans la durée, de 1999 à 2014

Nous appliquons la méthode analytique d'évaluations *critériées* définie dans le chapitre 1, section 1.3.

Tableau 4.3 Évaluations et analyses « *critériées* » par domaines disciplinaires et séries chronologiques sur l'axe linguistique

Analyses *critériées*		**Axe du temps et série chronologique (par année)**									
Année		**1999-2005**			**2006-2009**			**2010-2012**			**2013-2014**	
Langues/sphère		FR	US	CN	FR	US	CN	FR	US	CN	EN	CN
Axe linguistique	Morphosyntaxique							X	X	X		X
	Sémantique							X	X	X	X	X
	Lexique et Lexicographique	X						X	X	X	X	X
	Pragmatique (deixis)							X		X		X
Intérêts manifestés	Peu intéressant, Intéressant, Très intéressant	Peu intéressant			Peu intéressant			Très intéressant				

Tableau 4.4 Évaluations et analyses « *critériées* » par domaines disciplinaires et séries chronologiques sur l'axe statistique

Analyses *critériées*		**Axe du temps et série chronologique (par année)**									
Année		**1999-2005**			**2006-2009**			**2010-2012**			**2013-2014**	
Langues/sphère		FR	US	CN	FR	US	CN	FR	US	CN	EN	CN
Axe statistique	Ventilation	X			X	X		X	X	X	X	X
	Spécificité (évolutive)							X	X	X	X	X
	AFC				X	X		X	X	X	X	X
	Carte des sections							X	X	X	X	X
	Cooc et poly-Cooc							X	X	X	X	X
	Segments répétés							X		X		
Intérêts manifestés	Peu intéressant, Intéressant, Très intéressant	Peu intéressant			Peu intéressant			Très intéressant				

Dans un premier temps, nous étudions les différents résultats linguistiques et statistiques correspondant à nos deux corpus entiers sur l'ensemble de la période. Ces résultats sont évalués en fonction de nos thèmes de recherche (énergies et environnement) afin de déterminer la validation des intérêts manifestés pour chaque critère de chacun des deux tableaux.

Pour une langue et une période données, si les résultats analytiques pour un critère d'un des deux tableaux correspondant à nos thèmes de recherche sont saillants, alors la case concernée est renseignée par une croix.

Cette méthode prospective a permis de nous procurer des brèches analytiques. Ces tableaux se manipulent comme des grilles d'évaluation appliquées sur toutes nos formes-pôles et tous les événements calculés et analysés sur les deux axes, permettant ainsi d'évaluer, critère par critère, les degrés d'importance de chaque période pour une langue sélectionnée sur une échelle de 0 à 4 pour la linguistique et une échelle de 0 à 7 pour la statistique. Si plus de la moitié des critères sont validés pour un tableau donné, alors, cette période est retenue.

Les résultats analytiques de ces deux tableaux 4.3 et 4.4, nous amènent à en déduire que les informations relatives à nos thèmes sont plus intéressantes à partir de l'année 2010.

4.5 Deux corpus de veille restreints à trois années en trois langues

Pour des raisons pratiques, le corpus parallèle CLRG sera traité dans le chapitre 7.

Au vu des résultats analytiques ci-dessus, le corpus de veille ENRG se restreint aux trois années groupées **2010, 2011 et 2012**.

Tableau 4.5 ENRG_FR, ENRG_US et ENRG_CN de 2010 à 2012 : caractéristiques textométriques du corpus ENRG

Sous-corpus	ENRG_FR	ENRG_US	ENRG_CN
Nombre d'occurrences	1 624 889	611 548	10 442 192
Nombre de formes	55 831	32 273	98 678
Nombre d'articles	2 773	1 008	14 504
Nombre de paragraphes	25 811	13 301	221 512
Période	du 03-01-2010 au 17-04-2012	du 01-01-2010 au 18-04-2012	du 12-03-2010 au 23-04-2013
Continuité ou discontinuité (sur l'axe du temps)	en continu sur les mois	en continu sur les mois	en discontinu sur les mois
Rubrique du site	Planète	Environnement (sous-rubrique)	Vert
Séries chronologiques par année dans la résonance événementielle (par AFC)	aucune par année	aucune par année	aucune par année
Années retenues	2010, 2011 et 2012 (partielle)	2010, 2011 et 2012 (partielle)	2010, 2011 et 2012

4.5.1 Traits et idées saillants de la typologie globale du corpus ENRG

Pour réaliser une étude de « *mapping* », nous allons recourir à la méthode « Analyse Factorielle des Correspondances », représentation obtenue à partir de la distance entre les différents points des nuages multidimensionnels (Benzécri, 1968, 1973, 1977, 1981).

4.5.2 Typologie textuelle sur un corpus

L'analyse factorielle des correspondances, appelée communément AFC, est une méthode statistique des analyses multidimensionnelles permettant « *d'évaluer la distance entre deux parties textuelles par leurs distributions spécifiques de types de vocabulaire* » (Lebart & Salem, 1994). Dans les lignes qui suivent, nous rappelons les principes de cette méthode très utilisée.

Cette méthode statistique vise à synthétiser les informations fournies par un ensemble de données en s'intéressant aux liens entre les différentes variables. Son emploi est particulièrement intéressant

lorsque les dimensions du tableau de données étudié sont grandes. Dans le domaine du traitement automatique des langues, elle permet de voir les différents liens entre les mots constitutifs du corpus étudié.

Pour mettre en œuvre une procédure d'analyse factorielle des correspondances, il est possible de partir d'un tableau défini comme suit : au début de chaque ligne, les 20 formes les plus fréquentes du texte sont recensées (le nombre choisi l'est à des fins d'expérimentation et permet de voir le modèle fonctionner sur de petites valeurs), en haut de chaque colonne, nous recensons les différentes parties du corpus étudié, l'intersection de la ligne i et de la colonne j contient le nombre d'occurrences de la forme i dans la partie j du corpus.

Afin d'expliciter le processus de l'AFC, nous allons considérer par exemple, un tableau mesurant le nombre de fois qu'un personnage d'un roman emploie un mot : au début de chaque ligne, nous écrivons le nom du personnage, au début de chaque colonne, nous écrivons le mot en question, à l'intersection de la ligne i et de la colonne j, nous aurons le nombre de fois que le personnage i a utilisé le mot j. Supposons qu'il y ait I personnages et J mots, nous aurons alors le tableau suivant (l'intersection de la ligne i et de la colonne j : $n_{i,j}$ représentant le nombre d'emplois du mot j par le personnage i) :

$$N = \begin{bmatrix} n_{1,1} & \cdots & n_{1,J} \\ \vdots & \ddots & \vdots \\ n_{I,1} & \cdots & n_{I,J} \end{bmatrix}$$

Par la suite, nous divisons par n le nombre total de mots pour obtenir la fréquence d'emploi du mot j par le personnage i que nous notons : $f_{ij} = \frac{n_{i,j}}{n}$. Notons $f_{i.} = \sum_j f_{i,j}$ et $f_{.j} = \sum_i f_{i,j}$. On regroupe les fréquences d'emploi f_{ij} dans un tableau F, qui est appelé le tableau des pourcentages. (les $f_{i;j}$ sont aussi les fréquences relatives du tableau)

$$F = \begin{bmatrix} f_{1,1} & \cdots & f_{1,J} \\ \vdots & \ddots & \vdots \\ f_{I,1} & \cdots & f_{I,J} \end{bmatrix}$$

Le tableau matriciel F pourra faire l'objet d'une opération de transposition matricielle (c'est-à-dire que le terme f_{ij} en ligne i et en colonne j figurera à ligne j et à colonne i de la matrice de F transposée) en fonction des objectifs de notre analyse (étude du comportement des personnages ou étude du lexique). La matrice transposée F' de F sera :

$$F' = \begin{bmatrix} f_{1,1} & \cdots & f_{J,1} \\ \vdots & \ddots & \vdots \\ f_{1,I} & \cdots & f_{J,1} \end{bmatrix}$$

Nous pouvons aussi nous intéresser aux proximités relatives de l'ensemble des points du nuage, c'est-à-dire des profils lignes, d'où une nouvelle matrice X dont les éléments sont $x_{i,j} = \frac{f_{i,j}}{f_i}$. Cette nouvelle matrice X est utilisée par la suite pour calculer les proximités relatives de chacun de ces points. Cette matrice X s'écrit comme suit :

$$X = \begin{bmatrix} \frac{f_{1,1}}{f_1} & \cdots & \frac{f_{1,J}}{f_1} \\ \vdots & \ddots & \vdots \\ \frac{f_{I,1}}{f_I} & \cdots & \frac{f_{I,J}}{f_I} \end{bmatrix}$$

Les points du nuage de points sont alors obtenus par une transformation préalable des cordonnées ; qui s'écrit :

$$z_{i,j} = \frac{1}{\sqrt{f_j}} \frac{f_{i,j}}{f_i} = p_{i,j} f_{i,j} \text{ , où } p_{i,j} = \frac{1}{f_i \sqrt{f_j}}.$$

Ces points peuvent être assimilés à des vecteurs. L'un de ces vecteurs a par exemple pour première composante la fréquence transformée d'emploi du mot 1 par le personnage i, et pour j-ème composante la fréquence transformée d'emploi du mot j par le personnage i, etc.

Dans le tableau 4.6 (ci-dessous), ce vecteur est par exemple la première ligne du tableau lexical entier (TLE).

$$z_1 = (\frac{1}{\sqrt{f_1}} \frac{f_{1,1}}{f_1}, \ldots, \frac{1}{\sqrt{f_j}} \frac{f_{1,j}}{f_1}, \ldots, \frac{1}{\sqrt{f_J}} \frac{f_{1,J}}{f_1}).$$

Le centre de gravité du nuage de points, contient les moyennes d'emploi des différents mots du texte. Il s'agit du vecteur :

$$(\sum_{i=1}^{I} f_i z_{i1}, \quad \ldots\ldots, \sum_{i=1}^{I} f_i z_{i,J})$$

Le centre de gravité du nuage représente ici un profil moyen.

L'inertie correspondant à la somme pondérée des carrés des distances entre les points du nuage et son centre de gravité se mesurera alors avec la formule :

$$\boldsymbol{d}(z_i, z_{i'}) = \sum_{j=1}^{J} (z_{i,j} - z_{i',j})^2 = \sum_{j=1}^{J} (\frac{1}{\sqrt{f_j}} \frac{f_{i,j}}{f_i} - \frac{1}{\sqrt{f_j}} \frac{f_{i',j}}{f_{i'}})^2 = \sum_{j=1}^{J} \frac{1}{f_j} (\frac{f_{i,j}}{f_i} - \frac{f_{i',j}}{f_{i'}})^2$$

Le moment total d'inertie correspond à la somme des carrés des distances des points du nuage à leur centre de gravité. Ainsi si z_G est le centre de gravité du nuage de points, le moment total d'inertie M sera donc :

$$M = \sum_{i=1}^{I} (z_i - z_G)^2$$

Une inertie nulle traduira le fait que tous les personnages de l'étude emploient chacun des mots avec la même fréquence.

Ensuite, nous calculons la matrice V de variance-covariance dont les termes sont : $V_{i,j}$ correspondant à la covariance des colonnes i et j du tableau de données initial. $V_{i,j}$ s'écrit :

$$V_{i,j} = Cov(z_i, z_j) = \sum_k (\frac{1}{\sqrt{f_i}} \frac{f_{k,i}}{f_k} - \sqrt{f_i})(\frac{1}{\sqrt{f_j}} \frac{f_{k,j}}{f_k} - \sqrt{f_j}) = \sum_k \frac{1}{\sqrt{f_i}} \frac{f_{k,i} f_{k,j}}{\sqrt{f_j} f_k} - \sqrt{f_i}\sqrt{f_j}$$

Pour i différent de j.

Si i = j, on a :

$$V_{i,i} = \sum_k \frac{{f_{k,i}}^2}{f_i f_k} - f_i = \sum_k \frac{{f_{k,i}}^2 - {f_i}^2 f_k}{f_i f_k}$$

La matrice de variance permet de quantifier la variation de chacune des colonnes par rapport aux autres et donc la dispersion des données étudiées.

Cette matrice V permet d'obtenir les valeurs propres ou encore composantes principales et vecteurs propres utiles au changement de base.

Nous obtenons la part de la variance ou dispersion expliquée par la i-ème composante principale : l_i avec la formule :

$$\frac{\mathbf{100}\, \boldsymbol{l_i}}{(\boldsymbol{tr(V)} - \mathbf{1})}$$

tr(V) étant la trace de la matrice V des covariances, c'est-à-dire la somme de ses éléments diagonaux (la somme des valeurs propres).

La formule tr(V) – 1 s'explique par le fait que nous avons exclu la valeur propre 1 (la première valeur propre), car elle définit un axe principal sur lequel nous n'avons aucune dispersion.
En retenant 2 ou 3 axes principaux, il doit être possible alors d'expliquer au moins 70 % de la variance totale.
Le calcul des coordonnées des variables sur les axes principaux permettra alors de terminer l'AFC.

En résumé, l'AFC est basée sur un tableau de correspondances, appelé également tableau de contingence, de dépendance ou encore tableau lexical entier (TLE) pour la textométrie. Ce tableau contient des unités de formes, décomptées à l'intérieur de textes du corpus : chaque ligne correspond à une forme d'un mot du corpus, et chaque colonne à la fréquence de chaque forme dans chaque partie sélectionnée du corpus (partie souvent encadrée par deux balises) ; dans nos études, ces parties sont des périodes chronologiques (par année, cf. tableau 4.6 ci-dessous). Le croisement de chaque ligne et de chaque colonne, *i.e.* le nombre d'occurrences de chaque forme, est l'association d'une forme à sa fréquence dans un empan textuel donné, autrement dit, la fréquence d'une forme dans une période donnée.

L'AFC est particulièrement efficace pour mettre en évidence les éléments du tableau de contingence de manière hiérarchisée, car la taille de ce tableau de correspondance est souvent très grande, voire gigantesque. L'AFC suggère, avec ses rôles symétriques joués par les lignes et les colonnes, des pistes de réflexion difficiles à appréhender autrement.

Nous la mettons en application, le tableau 4.6 ci-dessous présente un extrait sélectif du tableau lexical entier. La première colonne indique l'identifiant de la forme, la deuxième colonne stocke la forme des mots, la troisième et les suivantes marquent les différents nombres d'occurrences des mots par année, et la dernière colonne (ajoutée pour des raisons pratiques) est l'addition du nombre d'occurrences de chaque forme sur toute la période retenue. L'extrait sélectionné a été choisi en fonction des thèmes de nos études, à savoir, *nucléaire*, *énergie* et *environnement*. Nous constatons que le classement de trois

de nos thèmes privilégiés reste dans la partie supérieure du tableau. Rappelons que ce tableau est trié par le nombre de fréquence des formes par ordre décroissant.

Tableau 4.6 ENRG_FR : extrait sélectif du tableau lexical entier (TLE)

	Forme	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Somme
92	Etats	1	11	15	48	42	15	201	159	254	130	459	606	533	235	2709
93	nucléaire	0	0	1	3	18	3	21	112	93	116	149	175	1480	501	2672
94	avoir	1	13	1	15	38	26	140	172	193	138	305	679	602	267	2590
95	C	3	14	3	28	44	32	135	144	174	135	316	661	559	283	2531
96	encore	3	8	7	27	40	17	92	196	178	129	269	633	623	272	2494
97	sans	8	8	18	22	42	45	129	178	199	142	296	528	497	292	2404
98	personnes	0	7	7	11	29	17	123	185	112	105	394	560	638	215	2403
99	Pour	3	14	7	29	25	17	141	172	161	142	286	581	614	206	2398
100	1	4	6	0	12	17	13	102	157	216	120	270	570	617	272	2376
101	effet	7	15	6	18	47	17	130	187	320	156	299	427	501	232	2362
102	bien	4	16	9	28	35	22	110	159	186	153	265	521	570	275	2353
103	leurs	1	16	7	22	25	15	114	122	224	156	264	574	478	256	2274
104	déjà	2	4	4	24	27	15	137	174	192	128	248	525	546	220	2246
105	Ce	3	10	10	17	35	26	131	144	183	104	265	526	473	227	2154
106	rapport	0	4	14	5	26	13	93	133	295	77	230	501	518	241	2150
107	Une	3	5	2	17	35	14	96	148	134	109	264	510	511	240	2088
108	avant	2	12	9	22	31	12	94	157	155	124	231	553	457	226	2085
109	2	7	4	0	10	25	19	85	153	218	80	217	466	557	242	2083
110	Un	4	11	2	13	24	15	70	123	136	106	256	506	539	264	2069
111	soit	0	10	5	19	21	17	97	149	156	112	241	491	499	231	2048
112	monde	3	12	4	15	37	20	116	150	178	118	260	439	472	216	2040
113	gaz	4	5	2	0	19	0	124	92	284	97	199	227	665	320	2038
114	donc	3	10	3	23	23	16	77	131	138	119	223	461	587	221	2035
115	énergie	3	2	0	10	57	2	66	86	141	135	250	395	622	261	2030
116	Dans	2	3	10	16	29	15	84	134	131	115	230	496	519	243	2027
117	trois	3	10	7	12	30	6	104	207	141	96	250	465	514	175	2020
118	santé	1	9	5	13	33	38	144	179	135	118	490	353	348	149	2015
119	fois	6	12	5	12	32	19	105	146	144	104	214	485	506	216	2006
120	années	5	16	6	18	31	23	110	138	164	105	176	437	502	272	2003
121	Unis	1	11	11	38	38	11	159	106	201	100	328	408	393	184	1989
122	environnement	1	3	3	17	9	5	112	112	235	222	192	469	381	225	1986

L'analyse de ce tableau sur l'ensemble de la période nous permet d'en déduire que plus la forme est fréquente, plus elle est classée haut dans le tableau, autrement dit, la forme la plus répétée est toujours classée en première ligne. En français, le mot le plus fréquent est « de ». L'extrait de ce tableau ci-dessus montre que la forme *nucléaire* au singulier est plus présente (93ème rang avec une fréquence totale de 2 672 fois) que les mots *énergie* au singulier et *environnement* qui occupent respectivement le 115ème rang avec une fréquence de 2 030 et le 122ème rang avec une fréquence de 1 986.

4.5.3 Mise en œuvre de l'AFC sur un extrait d'un article d'ENRG_FR

Considérons l'extrait modifié de l'article du Monde : « L'EPR, chronique d'un chantier qui s'enlise »

<paragraphe=01> L'EPR de Flamanville en novembre 2009. Le temps est à l'orage au-dessus de l'EPR de Flamanville (Manche). Une fois de plus, le chantier du ***réacteur*** *****nucléaire*** *de troisième génération a été épinglé pour des défaillances. Cette fois, ce sont des malfaçons dans le gros œuvre qui ont fait l'objet d'une lettre au vitriol adressée par l'Autorité de* ***sûreté*** *****nucléaire*** *(ASN) à l'opérateur EDF, le 18 juillet, parmi d'autres courriers de réprimande révélés par Le Canard Enchaîné mercredi 31 août.*

<paragraphe=02> La semaine dernière, treize autres faiblesses avaient déjà été constatées par le gendarme du ***nucléaire****, tandis que samedi, on apprenait que le site n'était pas totalement aux normes sismiques. Et l'on ne compte plus les lettres, rapports ou documents internes égrénant les lacunes de la future centrale, tant vantée par le gouvernement, et présentée comme "la plus sûre au monde" par le fabricant Areva. Un* ***réacteur*** *du même type en Finlande et deux autres en Chine, sont en cours de construction, les chantiers accumulent d'importants retards et sont la cible de nombreuses critiques. L'EPR, d'un fleuron du* ***nucléaire*** *français, est ainsi en passe de devenir l'une des technologies les plus décriées. »*

<paragraphe=03> Sur le papier, le ***réacteur*** *pressurisé européen (European pressurized reactor), conçu par Areva et l'allemand Siemens dans les années 1990, est censé représenter, en termes de* ***sûreté****, un modèle dans le monde. Le* ***réacteur****, d'une puissance de 1 650 mégawatts, aurait été conçu pour résister à la chute d'un avion gros porteur, et ses multiples systèmes de sécurité doivent le mettre à l'abri d'un accident détruisant le cœur du* ***réacteur****. Les piscines de refroidissement des combustibles usés seront même protégées par une enceinte de confinement. Au final, le risque de prolifération des matières radioactives serait quasiment nul.*

Nous avons dans ces trois paragraphes recensés manuellement le nombre d'emploi des formes *réacteur, nucléaire et sûreté*. Nous obtenons les résultats suivants :

	réacteur	***nucléaire***	***sûreté***
Paragraphe 1	1	2	1
Paragraphe 2	1	2	0
Paragraphe 3	3	0	1

Nous appliquons le résultat ci-dessus en réalisant l'AFC au moyen du logiciel R[140], logiciel permettant de visualiser simplement les coordonnées vectorielles (figure 4.12 ci-dessous) des formes à la différence de Lexico 3, nous obtenons le résultat suivant :

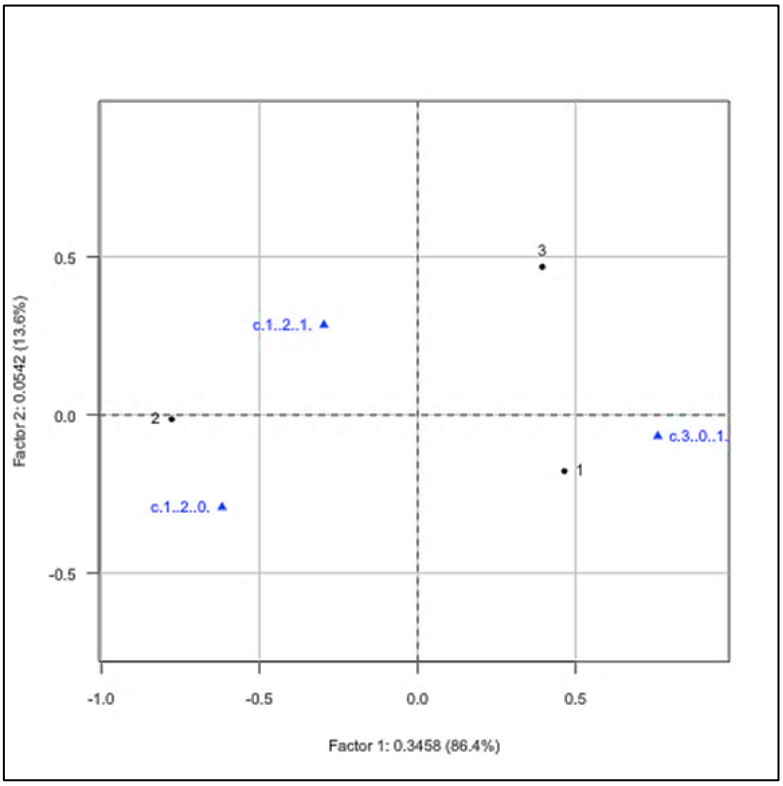


Figure 4.12 Résultat d'une démonstration d'AFC par le logiciel R

[140] https://www.r-project.org/

Nous pouvons facilement constater que les 3 formes, *réacteur, nucléaire, sûreté* sont représentées par leurs coordonnées vectorielles dans un espace : c (1, 2, 1), c (1, 2, 0) et c (3, 0, 1). L'axe 1 explique 86,4% de l'inertie, c'est-à-dire de la « variation » par rapport au profil moyen (c'est-à-dire ici au vecteur ligne (x1, x2, x3) avec x1 le nombre d'occurrence de *réacteur*, x2 le nombre d'occurrence de *nucléaire*, et x3 le nombre d'occurrence de *sûreté*. L'axe 2 explique 13,6% de l'inertie. Les deux axes permettent donc d'expliquer 100% de l'inertie.

L'axe factoriel 1 correspond à la valeur propre 0.3458.
L'axe factoriel 2 correspond à la valeur propre 0.0542.

L'axe factoriel 1 explique 0.3458/(0.3458 + 0.0542)*100 = 86,4% de l'inertie.
L'axe factoriel 2 explique 0.0542/(0.3458 + 0.0542)*100 = 13,4% de l'inertie.

Nous constatons que les profils, c'est-à-dire ici, les paragraphes que l'on caractérise par les vecteurs lignes (x1, x2, x3) avec x1 le nombre d'occurrence de *réacteur*, x2 le nombre d'occurrence de *nucléaire*, et x3 le nombre d'occurrence de *sûreté*, sont assez éloignés sur le plan de la proximité textuelle. Ainsi, il sera *a priori* peu fréquent de trouver par exemple simultanément un paragraphe ayant employé 3 fois *réacteur*, 0 fois *nucléaire*, et 1 fois le mot sûreté et un paragraphe ayant employé 1 fois *réacteur*, 2 fois *nucléaire* et 0 fois *sûreté*.

4.5.4 Typologie sur les trois sous-corpus restreints ENRG 2010, 2011 et 2012

Afin de mettre en évidence la proximité textuelle de chacune des trois années 2010, 2011, 2012, nous allons appliquer la méthode des AFC sur les trois sous-corpus.

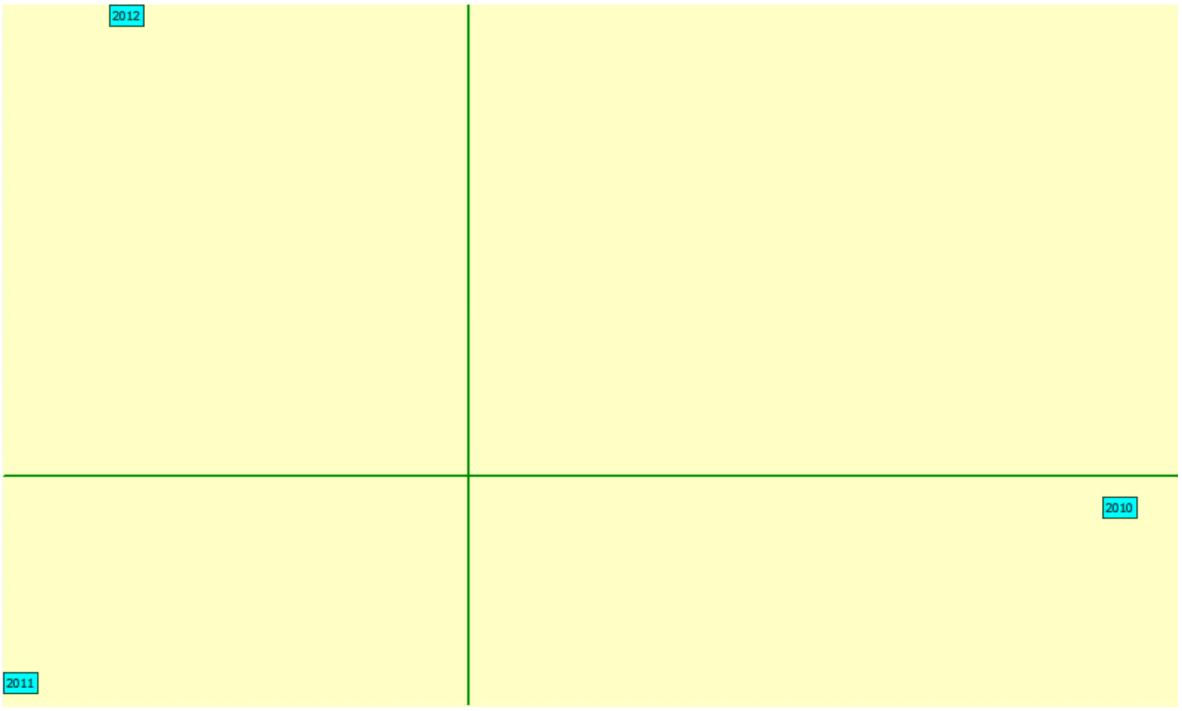


Figure 4.13 ENRG_FR de 2010 à 2012 : analyse factorielle des correspondances sur les années

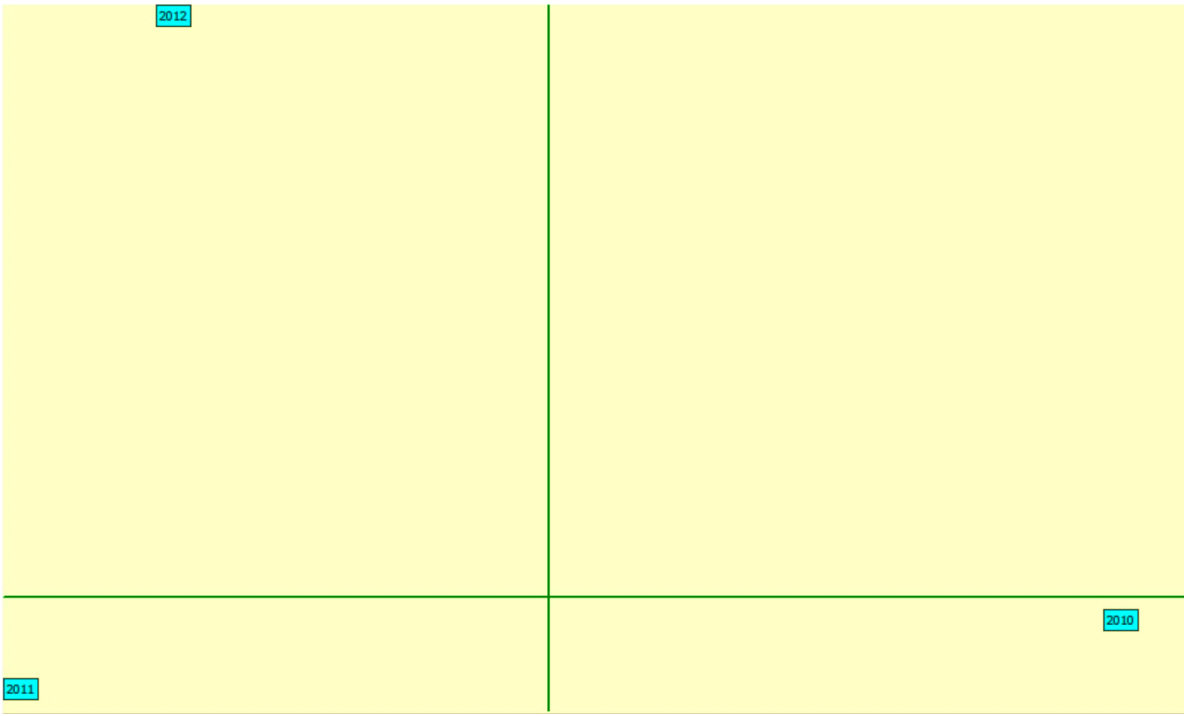


Figure 4.14 ENRG_US de 2010 à 2012 : analyse factorielle des correspondances sur les années

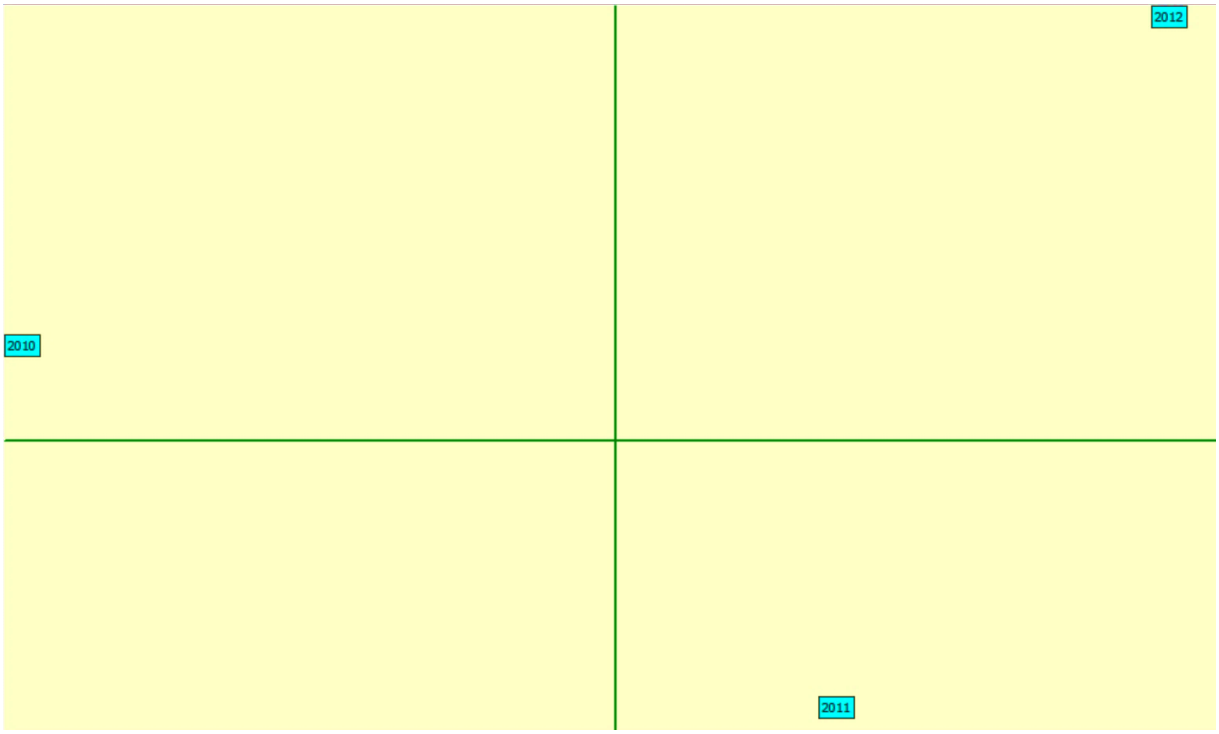


Figure 4.15 ENRG_CN de 2010 à 2012 : analyse factorielle des correspondances sur les années

Les trois calculs d'AFC (figures 4.13, 4.14 et 4.15) sur les sous-corpus amputés montrent communément que les années 2011 et 2012 restent regroupées du même côté de l'axe vertical, alors que l'année 2010 se retrouve à chaque fois isolé dans une dimension individuelle. Ceci dit qu'il y a une proximité textuelle entre 2011 et 2012.

Les trois années peuvent également former une parabole sur le plan d'AFC, ce phénomène s'explique de façon approximative, par l'effet Guttman (Flament et Milland, 2005), qui consiste à théoriser que la répétition du vocabulaire des *news*, que nous appelons parfois discursivité de vocabulaire, évolue de manière stable et chronologique et souvent sous une forme parabolique sur le plan d'AFC.

4.5.5 Points divergents de la période 2010, 2011 et 2012

A travers les analyses du dépouillement général (annexe B), nous constatons que les événements internationaux occupent davantage la une des journaux français, tandis que les Américains et les Chinois accordent plus d'attention aux événements internes du fait qu'il s'agit de deux pays géographiquement étendus avec une démographie élevée. Il faut également tenir compte que les États-Unis et la Chine forment deux grands blocs géopolitiques aux civilisations individuelles et protectionnistes par rapport à la France, l'une des deux locomotives de l'Europe.

En 2010, le Monde évoque davantage l'éruption volcanique islandaise (un fait menaçant pour l'Europe) et l'explosion de la plate-forme « *Deepwater Horizon* », quant au NYT, il se focalise principalement sur l'événement de l'explosion de la plate-forme « *Deepwater Horizon* », un fait saillant pour les Américains. En Chine, des pollutions touchant gravement à l'environnement (*cf.* chapitre 6, tableau 6.2 et annexe B, tableau B.17) ont totalement captivé l'attention des Chinois et dépassé leur préoccupation envers l'extérieur. Il s'agit principalement de failles dans les infrastructures et le non-respect des normes de sécurité.

En 2011 et 2012, la catastrophe de Fukushima occupe presque toute la presse française. En effet, l'opinion française ainsi que les associations et organisations écologistes sont particulièrement attentives à tous les événements qui touchent de près ou de loin le domaine du nucléaire. De plus, l'énergie française dépend presque entièrement du nucléaire depuis de longues années, car dès 1974 le gouvernement français lance et déploie un programme électronucléaire (*cf.* annexe D, le nucléaire et la politique française). Or, aux Etats-Unis, cette catastrophe apparait probablement secondaire dans la rubrique étudiée du NYT, car le mot nucléaire reste tabou pour les Américains (Chavardès, 2009). Les Américains demeurent très prudents vis-à-vis du nucléaire depuis l'accident en 1979 de *Three Mile Island*. Cette méfiance s'estompe peu à peu face aux contextes énergétiques et environnementaux actuels. Leur perception énergétique sera étudiée de manière détaillée dans le chapitre suivant. Les scandales liés à la santé et l'écologie ont fait couler énormément d'encre dans la presse chinoise (*cf.* chapitre 6).

ENRG_FR et ENRG_US respectent mieux leurs comparabilités et homogénéités sur les mêmes années relayant des événements relativement proches. Par ailleurs, le volume d'articles est sensiblement le même dans ces deux sous-corpus, et la collecte des données s'est arrêté le 18 avril 2012, aussi nous écartons ENRG_CN des premières analyses comparatives. Par la suite, nous nous focaliserons sur le sous-corpus chinois. Les articles de ce dernier, qui ont ponctué la période retenue, relatent constamment les nombreux scandales liés à la santé publique, aux pollutions des eaux, de l'air, de l'alimentation et à la construction d'une centrale nucléaire, ainsi que la taxe carbone imposée par l'Union Européenne en 2012 (*cf.* chapitre 6).

Conclusion du chapitre

Le chapitre 4 nous a donc permis d'expliciter le processus de récolte des données. Nous avons ainsi justifié le choix de nos sources de données. Nous avons aussi comparé notre corpus comparable trilingue : ENRG_FR : pour le sous-corpus de textes français concernant l'énergie et l'environnement, ENRG_US : pour le sous-corpus de textes américains concernant l'énergie et l'environnement, et ENRG_CN : pour le sous-corpus de textes chinois concernant l'énergie et l'environnement.

La comparabilité de ces sous-corpus a été étudiée au moyen de critères quantitatifs et qualitatifs notamment grâce aux résultats des AFC, dont nous avons rappelé les principes essentiels. La mise en œuvre de ces critères nous a conduits à retenir :

- les années 2010, 2011, et partiellement l'année 2012 pour le sous-corpus ENRG_FR,
- les années 2010, 2011, et partiellement l'année 2012 pour le sous-corpus ENRG_US,
- les années 2010, 2011 et 2012 pour le sous-corpus ENRG_CN.

La méthode de l'analyse factorielle des correspondances (AFC) a été présentée dans ce chapitre. Il convient de noter qu'il existe d'autres méthodes statistiques telles que les modèles à thème, *topic models*[141], dont l'objectif est de déterminer la structure thématique et sémantique d'un corpus de documents. Ces modèles largement utilisés en traitement automatique des langues ces dernières années prennent tout leur intérêt lorsqu'il est nécessaire de traiter un vaste ensemble de données textuelles. Les *topic models* ont rencontré un grand succès dans la détection d'ensemble de mots associés à travers des groupes de documents. Dans le cadre de notre thèse, il pourrait s'agir de déterminer la structure sémantique et thématique des thèmes liés à l'énergie dans un vaste corpus trilingue. La mise en œuvre de cet algorithme sur un corpus relativement important (plus de 20 000

[141] http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf (consulté le 01/07/2016)

documents) mené par Kevin Canini, Lei Shi et Thomas Griffiths[142] a permis globalement de retrouver les principaux thèmes du corpus et les mots les plus importants de ce jeu de données utilisé dans un temps d'exécution raisonnable. Etant donné que les principaux thèmes de nos corpus sont déjà connus, le choix du *topic model* n'aurait néanmoins pas produit de meilleurs résultats que l'AFC.

Nous avons également justifié le choix du site *chinadialogue.net* pour le corpus parallèle bilingue chinois-anglais concernant le changement climatique.

Le chapitre 5 sera l'occasion de comparer les sous-corpus français et américain.

[142] http://cocosci.berkeley.edu/tom/papers/topicpf.pdf (consulté le 01/07/2016)

Partie 3 Veille trilingue français, anglais américain et chinois

« La nature fait les hommes semblables, la vie les rend différents.»
- Confucius.

« L'ignorance coûte plus cher que l'information.»
- John Fitzgerald Kennedy

« La connaissance s'acquiert par l'expérience. Tout le reste n'est qu'information.»
- Albert Einstein

« Le savoir est la seule matière qui s'accroît quand on la partage.»
- Socrate

« Savoir pour prévoir, afin de pouvoir.»
- Auguste Comte

Dans ce chapitre, nous donnerons d'abord les caractéristiques textométriques des deux corpus comparable ENRG et parallèle CLRG notamment sur la période de 2010 à 2012 :

- répartition d'occurrences et formes par mois,
- évolution du nombre d'articles,
- évolution du nombre d'occurrences,
- accroissement du vocabulaire.

Nous rechercherons ensuite les convergences et particularités pour la typologie textuelle dans ces deux corpus. L'étude de ces deux corpus sera enfin complétée au moyen d'un calcul de spécificités et du modèle hypergéométrique dont nous décrirons les principes.

5. Essai de veille parallèle sur les sous-corpus français et américain

Définir la veille correspond à mettre en œuvre une chaîne d'activités analytiques et pratiques avec des objectifs précis qui permettent la fouille d'informations dans le passé, afin d'apporter de l'aide pour le présent mais aussi d'anticiper et de prévoir des actions pour le futur. La veille textométrique applique ce même principe : fouiller les expériences du passé dans l'objectif de prévoir et d'anticiper les informations pour le présent et l'avenir.

Une exploration textométrique sera appliquée pour la veille autour du thème « énergie(s) » avec un zoom particulier sur l'énergie nucléaire et l'EPR sur les sous-corpus français et américain en mobilisant en particulier les outils d'analyse factorielle des correspondances, spécificités, cooccurrences et poly-cooccurrences.

5.1 Caractéristiques textométriques des deux sous-corpus divisés par mois

Les dépouillements textométriques sont appliqués sur les deux sous-corpus restreints afin d'obtenir les répartitions d'occurrences et formes par mois, l'évolution du nombre d'articles et d'occurrences et l'accroissement de vocabulaire sur la période retenue.

Tableau 5.1 ENRG_FR de janvier 2010 à avril 2012 : répartition mensuelle du nombre d'occurrences, formes, et hapax

Nombre d'occurrences: 1624889 Nombre de formes: 55831
Nombre d'hapax: 22799 Fréquence maximale: 89616

Num	Partie	Occurenc	Formes	Hapax	Fmax	Forme
1	201001	32835	6431	3650	1647	de
2	201002	28628	6002	3315	1584	de
3	201003	61198	9339	4809	3215	de
4	201004	54895	7978	3925	3084	de
5	201005	59386	8605	4224	3294	de
6	201006	64678	9521	4902	3536	de
7	201007	103873	12968	6355	5512	de
8	201008	49209	8703	4783	2848	de
9	201009	32791	6741	3836	1805	de
10	201010	56703	8633	4296	3099	de
11	201011	58740	8059	3696	3201	de
12	201012	54851	8735	4554	3107	de
13	201101	35030	6436	3337	1848	de
14	201102	28011	5325	2711	1566	de
15	201103	87638	9859	4459	4764	de
16	201104	51022	7570	3842	2846	de
17	201105	49672	7898	4064	2759	de
18	201106	73521	9959	4818	3929	de
19	201107	30187	5790	3102	1732	de
20	201108	51946	8791	4792	2966	de
21	201109	52242	8580	4468	3037	de
22	201110	52573	9085	4884	2915	de
23	201111	70990	10276	5329	3951	de
24	201112	77769	11191	5801	4377	de
25	201201	82443	11902	6092	4519	de
26	201202	87440	12567	6499	4769	de
27	201203	86671	12255	6466	4936	de
28	201204	49947	8482	4608	2770	de

En analysant le tableau 5.1 ci-dessus, nous constatons des variations du nombre d'occurrences, de formes[143] et d'hapax[144] entre janvier 2010 et avril 2012, à l'exception du mois d'avril 2012 incomplet. Le nombre d'occurrences varie de 28 011 pour février 2011 à 103 873 pour juillet 2010. Le nombre de formes témoigne également d'une grande variation, variation en corrélation avec le nombre d'occurrences, à savoir, ce nombre varie de 5 325 en février 2011 à 12 968 en juillet 2010. Toutefois, à partir de novembre 2011, le nombre de formes dépasse les 10 000. Quant au nombre d'hapax, les mois de juillet 2010, février et mars 2012 dominent largement avec 6 355, 6 499 et 6 466 respectivement.

[143] Rappelons qu'il s'agit d'une extraction exhaustive de l'ensemble de formes séparées par un blanc/espace et par les signes de ponctuation suivants : .,:;!?/_-\'"()[]{}§$@#&. Il convient de noter que ce choix d'extraction est une approche empirique proposée par la textométrie.

[144] occurrences, formes, hapax, fréquence maximale, Fmax (fréquence maximale) : voir définitions dans le glossaire.

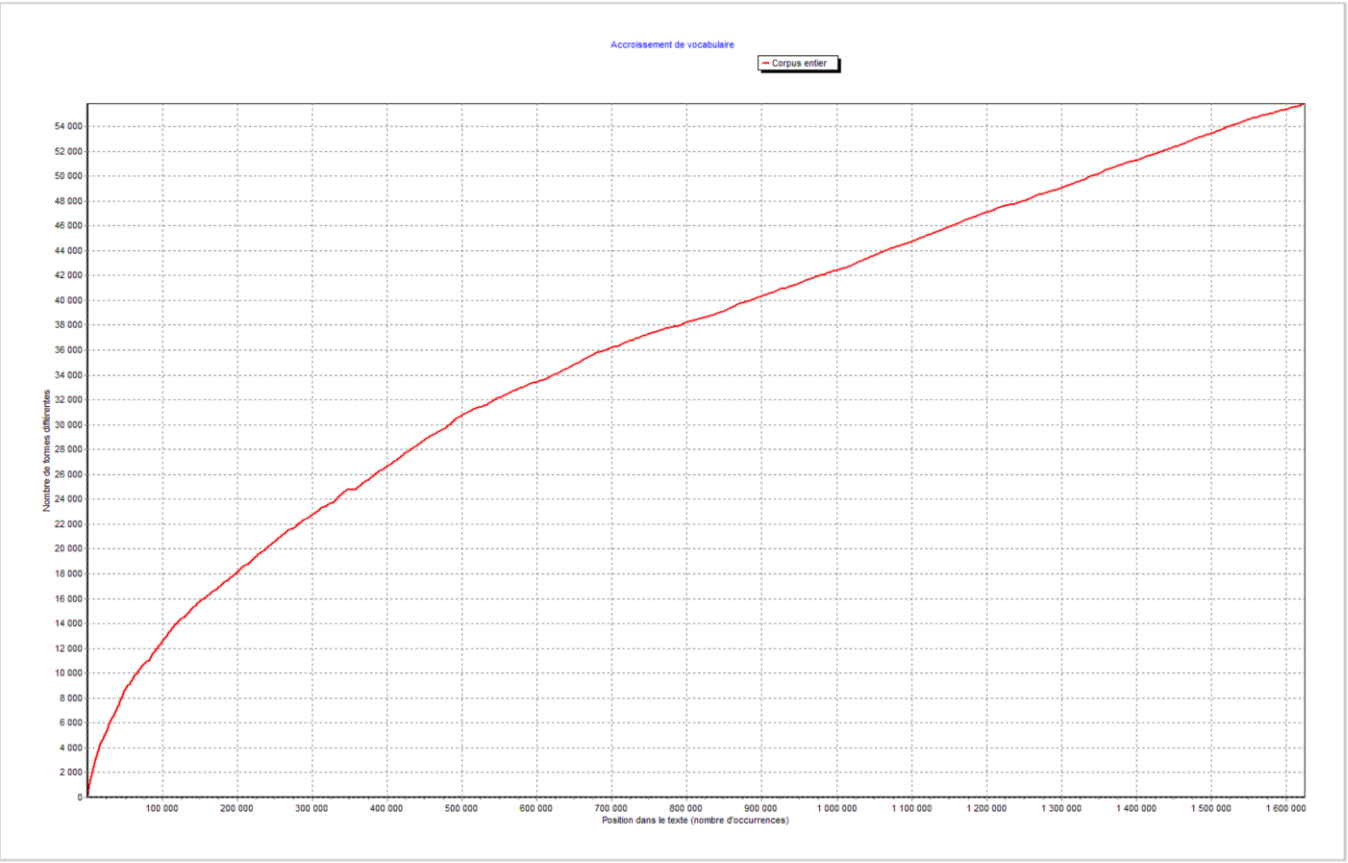


Figure 5.1 ENRG_FR de janvier 2010 à avril 2012 : accroissement de vocabulaire

Le diagramme d'accroissement de vocabulaire (figure 5.1) permet d'observer l'apparition de nouvelles formes au fur et à mesure de l'avancement dans ENRG_FR. L'ensemble du sous-corpus compte 1 624 889 occurrences et 55 831 de formes (tableau 5.1 ci-dessus), le renouvellement de formes se stabilise après 600 000 d'occurrences. Par la suite, pour chaque intervalle de 100 000 occurrences supplémentaires, le nombre de formes augmente de 2 000 environ.

Tableau 5.2 ENRG_US de janvier 2010 à avril 2012 : répartition mensuelle du nombre d'occurrences, formes, et hapax

Nombre d'occurrences: 611548 Nombre de formes: 32273
Nombre d'hapax: 13332 Fréquence maximale: 35265

Num	Partie	Occurenc	Formes	Hapax	Fmax	Forme
1	201001	24278	5599	3196	1397	the
2	201002	15223	3962	2309	824	the
3	201003	21031	4937	2827	1141	the
4	201004	34826	6874	3679	1979	the
5	201005	37513	6685	3564	2429	the
6	201006	33410	6419	3476	2006	the
7	201007	19594	4691	2671	1255	the
8	201008	26741	5522	3029	1642	the
9	201009	22003	5120	2936	1260	the
10	201010	20490	4890	2793	1129	the
11	201011	18307	4235	2345	1002	the
12	201012	22927	5046	2709	1405	the
13	201101	19372	5001	2908	1073	the
14	201102	28181	5990	3289	1572	the
15	201103	32882	6468	3485	1877	the
16	201104	31805	6812	3844	1735	the
17	201105	23381	5408	3054	1325	the
18	201106	34644	7107	3911	1881	the
19	201107	31768	6643	3717	1841	the
20	201108	15394	4193	2511	875	the
21	201109	18601	4500	2604	1017	the
22	201110	11024	3267	1985	594	the
23	201111	12765	3527	2065	827	the
24	201112	12112	3366	2019	727	the
25	201201	11136	3253	2006	645	the
26	201202	14956	3882	2189	821	the
27	201203	11014	3369	2127	650	the
28	201204	6170	2167	1414	336	the

Le tableau 5.2 ci-dessus présente des variations relativement faibles, par rapport à ENRG_FR, du nombre d'occurrences, de formes et d'hapax entre janvier 2010 et avril 2012, à l'exception du mois d'avril 2012 incomplet. Le nombre d'occurrences varie de 11 014 pour mars 2012 à 37 513 pour mai 2010. Le nombre de formes issues de l'extraction exhaustive témoigne d'une bulle plus petite de richesse par rapport à son homologue français, et d'une variation, plus ou moins, en corrélation avec le nombre d'occurrences. Toutefois, à partir d'août 2011, le nombre de formes chute avec un sursaut en septembre 2011. Quant au nombre d'hapax, nous remarquons le même phénomène que pour le nombre de formes.

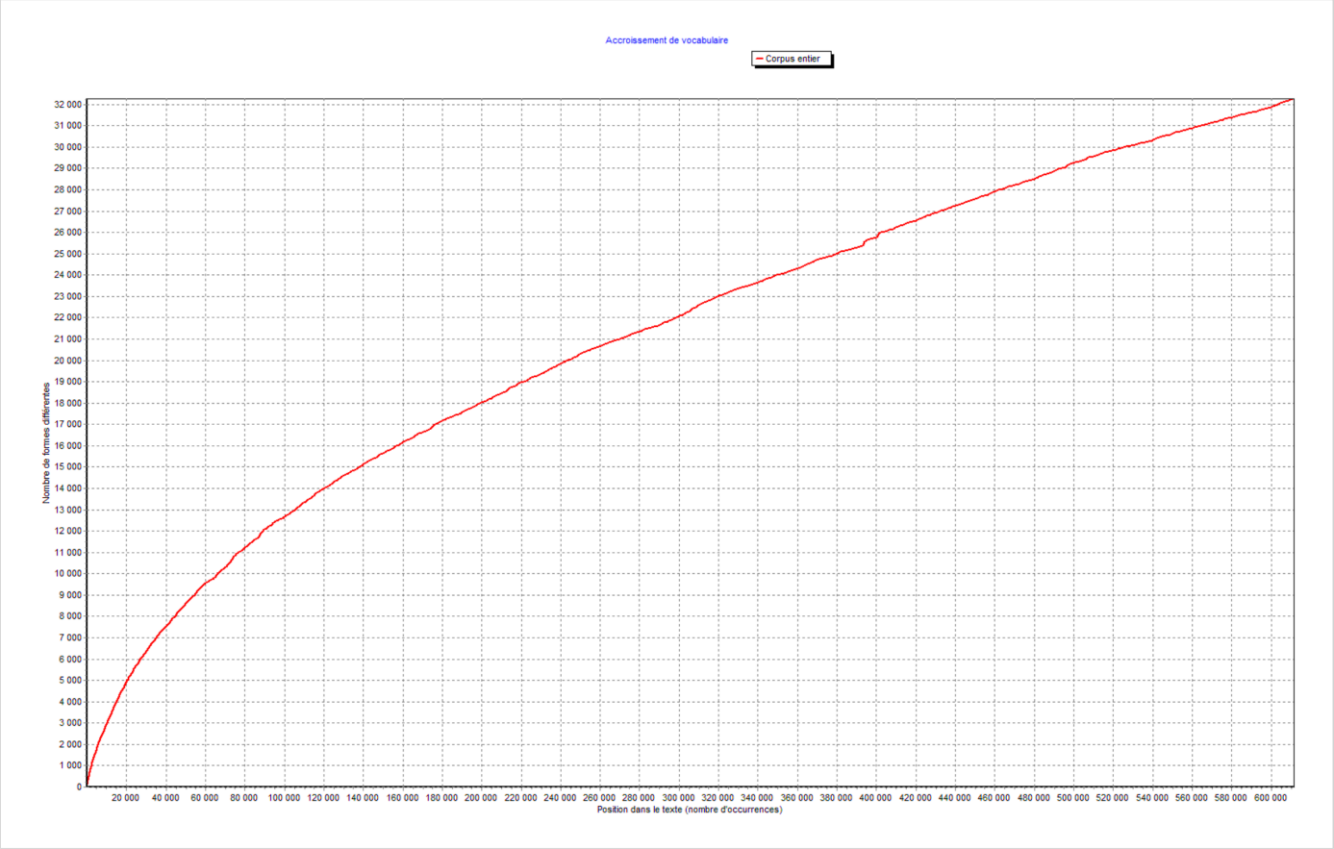
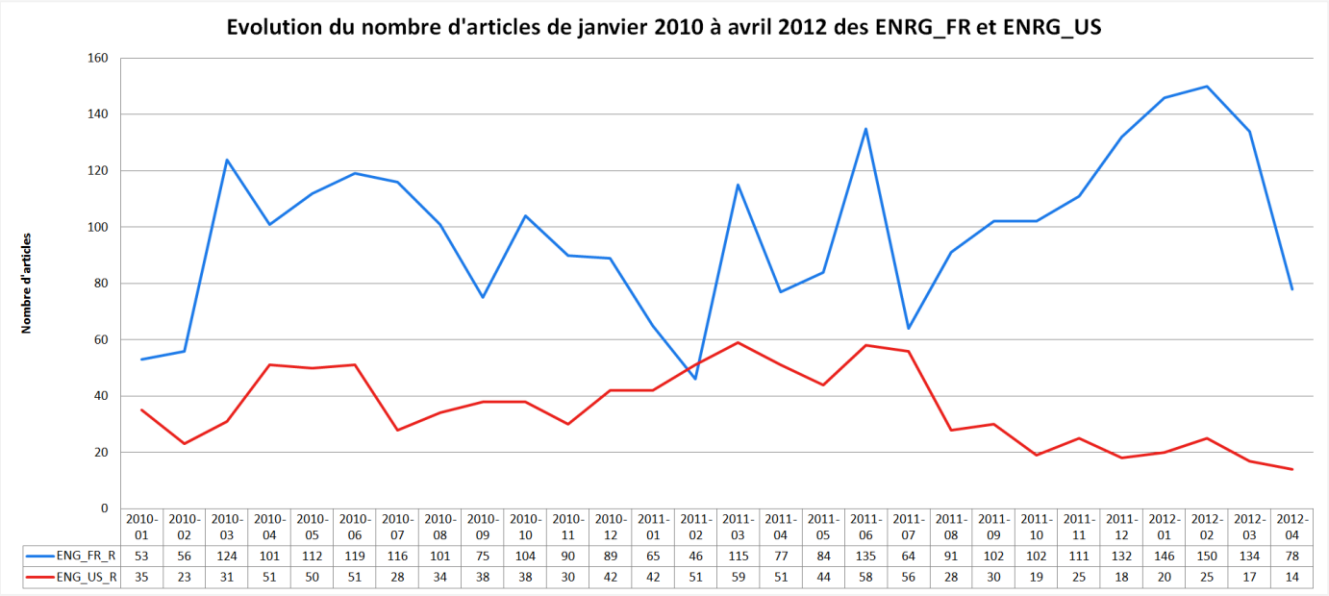


Figure 5.2 ENRG_US de janvier 2010 à avril 2012 : accroissement de vocabulaire

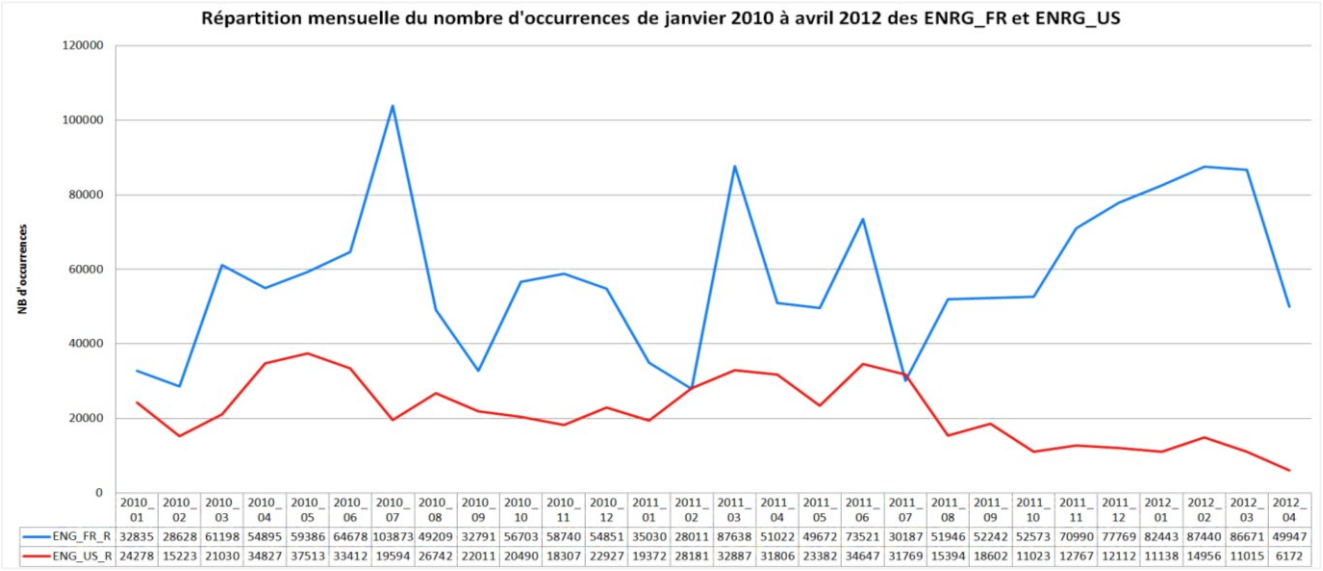
Le diagramme d'accroissement de vocabulaire (figure 5.2) permet d'observer l'apparition de nouvelles formes au fur et à mesure de l'avancement dans ENRG_US. L'ensemble du sous-corpus compte 611 548 occurrences et 32 273 formes (tableau 5.2 ci-dessus), le renouvellement de formes se stabilise après 100 000 occurrences. Par la suite, pour chaque intervalle de 20 000 occurrences supplémentaires, le nombre de formes augmente de 1 000 environ.



	2010-01	2010-02	2010-03	2010-04	2010-05	2010-06	2010-07	2010-08	2010-09	2010-10	2010-11	2010-12	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07	2011-08	2011-09	2011-10	2011-11	2011-12	2012-01	2012-02	2012-03	2012-04
ENG_FR_R	53	56	124	101	112	119	116	101	75	104	90	89	65	46	115	77	84	135	64	91	102	102	111	132	146	150	134	78
ENG_US_R	35	23	31	51	50	51	28	34	38	38	30	42	42	51	59	51	44	58	56	28	30	19	25	18	20	25	17	14

Figure 5.3 ENRG_FR et ENRG_US de janvier 2010 à avril 2012 : évolution du nombre d'articles

La figure 5.3 illustre une irrégularité de la production d'articles dans le sous-corpus français, 3 « pics » et 3 « chutes » ont ponctué la période concernée, à savoir, mars 2010, mars et juin 2011 à la hausse et septembre 2010, février et juillet 2011 à la baisse. Alors que, chez les Américains la variation est moins flagrante, mais il est à noter une baisse régulière à partir de juillet 2011, tandis que pour le Monde la courbe remonte au contraire jusqu'en février 2012 pour amorcer une décroissance en mars et avril 2012.

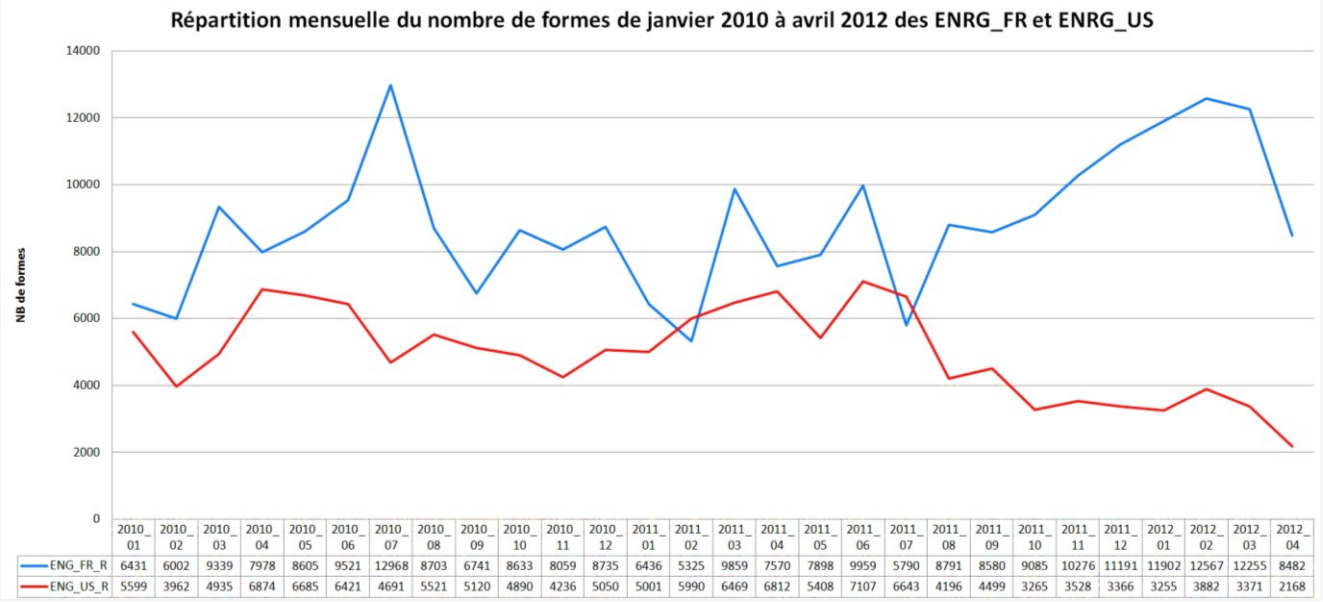


	2010_01	2010_02	2010_03	2010_04	2010_05	2010_06	2010_07	2010_08	2010_09	2010_10	2010_11	2010_12	2011_01	2011_02	2011_03	2011_04	2011_05	2011_06	2011_07	2011_08	2011_09	2011_10	2011_11	2011_12	2012_01	2012_02	2012_03	2012_04
ENG_FR_R	32835	28628	61198	54895	59386	64678	103873	49209	32791	56703	58740	54851	35030	28011	87638	51022	49672	73521	30187	51946	52242	52573	70990	77769	82443	87440	86671	49947
ENG_US_R	24278	15223	21030	34827	37513	33412	19594	26742	22011	20490	18307	22927	19372	28181	32887	31806	23382	34647	31769	15394	18602	11023	12767	12112	11138	14956	11015	6172

Figure 5.4 ENRG_FR et ENRG_US de janvier 2010 à avril 2012 : répartition mensuelle du nombre d'occurrences

Ces deux courbes (figure 5.4 ci-dessus) présentent des tendances extrêmement fluctuantes. La variation du nombre d'occurrences d'ENRG_FR varie du simple au quasi-quadruple, et du simple au triple pour ENRG_US. La cause de cette production intensifiée est étudiée ci-dessous.

Ces hausses s'expliquent par des catastrophes naturelles et humaines entraînant des conséquences sur l'environnement, à savoir, le colmatage de la plate-forme BP « *Deepwater Horizon*» le 16 juillet 2010 ; la grande vague de froid touchant 53 départements français, l'interdiction de la culture du maïs transgénique *OGM 810* de la société américaine Monsanto à Monbéqui dans le Tarn et Garonne et le redressement judiciaire de la société française Photowatt, fabricant des panneaux solaires photovoltaïques en février 2012 ; en mars 2012, la fuite de gaz de la plate-forme *Elgin Franklin* de Total en mer du Nord et l'installation des lignes THT (Très Haute Tension) Cotentin-Maine provoquant de violents remous et affrontements entre les habitants et les forces de l'ordre. En ce qui concerne les Etats-Unis, la courbe présente un « pic » en février 2012 avant sa chute. Un retour au contexte nous a permis de traquer l'un des événements principaux à savoir, la réactivation de débats sur les dons et la transparence des groupes environnementaux, suite à la révélation de l'affaire du Sierra Club, association américaine écologiste la plus ancienne, fondée à San Francisco en 1892, qui a accepté des dons de 26 millions dollars de la part d'une entreprise de gaz naturel.



	2010_01	2010_02	2010_03	2010_04	2010_05	2010_06	2010_07	2010_08	2010_09	2010_10	2010_11	2010_12	2011_01	2011_02	2011_03	2011_04	2011_05	2011_06	2011_07	2011_08	2011_09	2011_10	2011_11	2011_12	2012_01	2012_02	2012_03	2012_04
ENG_FR_R	6431	6002	9339	7978	8605	9521	12968	8703	6741	8633	8059	8735	6436	5325	9859	7570	7898	9959	5790	8791	8580	9085	10276	11191	11902	12567	12255	8482
ENG_US_R	5599	3962	4935	6874	6685	6421	4691	5521	5120	4890	4236	5050	5001	5990	6469	6812	5408	7107	6643	4196	4499	3265	3528	3366	3255	3882	3371	2168

Figure 5.5 ENRG_FR et ENRG_US de janvier 2010 à avril 2012 : répartition mensuelle du nombre de formes

Selon la figure 5.5, de manière générale, le nombre de formes en anglais est plus faible, mais plus stable que celui en français, la fluctuation du nombre de formes du Monde est supérieure à celle du NYT, sauf deux exceptions, les mois de février et juillet 2011. En ce qui concerne l'événement en février 2011, il s'agit de la plus importante amende, 9,5 milliards de dollars, de l'histoire du droit de l'environnement que devrait payer le groupe pétrolier américain Chevron pour avoir pollué pendant des années l'Amazonie. Quant au mois de juillet de cette même année, divers événements internes aux États-Unis tels que la sortie de voitures hybrides « *Lincoln MKZ Hybrid* », des commentaires des environnementalistes et des réflexions autour du projet ITER, « réacteur thermonucléaire expérimental international », ont suscité un vif intérêt médiatique du NYT.

5.2 Similitudes textuelles et contrastes pour ENRG_FR et ENRG_US

Rappelons que l'AFC décrit la répartition de toutes les formes du corpus dans les différents mois et/ou années en fonction de leurs fréquences (*cf.* chapitre 4, section 4.5.2). Dans notre étude, cette analyse croise la variable qualitative forme/mot du corpus avec la variable quantitative fréquence par année, découpée en 12 mois (fréquence par mois). L'AFC permet d'évaluer la situation d'uniformité et d'indépendance des formes du corpus, de calculer en quoi cette situation constatée en diffère, puis d'exprimer la différence d'emploi des formes sur un *mapping*. Par la suite, nous pouvons analyser, interpréter et expliquer les différents résultats de ce *mapping*.

Les deux axes unidimensionnels des AFC réalisées expriment chacun une partie de l'inertie totale ou encore de la dispersion des emplois des formes par rapport à un profil moyen d'emploi de ces mêmes formes. Le barycentre (centre de gravité) du nuage de points est le croissement des deux axes. Dans la plupart des études réalisées au moyen d'une AFC, les données sont nombreuses et comportent de multiples dimensions (par exemple : poids, âge diplôme, lieu d'habitation, etc.). L'AFC aura alors pour but de projeter ces données dans un espace à deux dimensions. Les deux axes retenus devront alors expliquer la plus grande part de l'inertie du nuage de points.

En général, les axes d'une AFC ne sont pas à interpréter, car peu de variables quantitatives interviennent dans les analyses. L'interprétation de l'AFC s'appuie généralement sur la position ou la distance relative aux différents résultats repérés. Nous avons choisi de considérer les quatre blocs formés par les deux axes comme support d'analyse. Il s'agit d'une manière empirique de procéder à nos analyses pour des raisons de lisibilité. Il aurait été sûrement possible de réaliser des groupements différents.

En effet, l'analyse des correspondances des séries textuelles chronologiques obéit à des règles d'interprétation un peu particulières. La plupart du temps, l'AFC est obtenue à partir de corpus dans lesquels la dimension chronologique constitue la principale dimension de variation des représentations un peu complexes qui alignent les périodes considérées selon des courbes en forme de paraboles (Effet Guttman). Ces représentations doivent être interprétées comme une évolution temporelle. Les périodisations que l'on peut opérer s'appuient alors sur une division des parties soumises à l'analyse en groupes formés de périodes consécutives dans le temps qui apparaissent dans une certaine proximité sur les plans factoriels.

L'AFC permet d'illustrer le clivage textuel d'un corpus par les spécificités du vocabulaire. Cette spécificité lexicale se manifeste par la périodisation de l'emploi récurrent du lexique. L'utilisation commune de formes est attestée par la proximité textuelle du plan de l'AFC. Nous pouvons regrouper les unités de l'AFC par leur proximité. Dans l'hypothèse où les informations événementielles se sont produites dans les périodes communes, nous supposons que les dimensions des deux AFC (ENRG_FR et ENRG_US) sont semblables et que les séries chronologiques communes sont comparables.

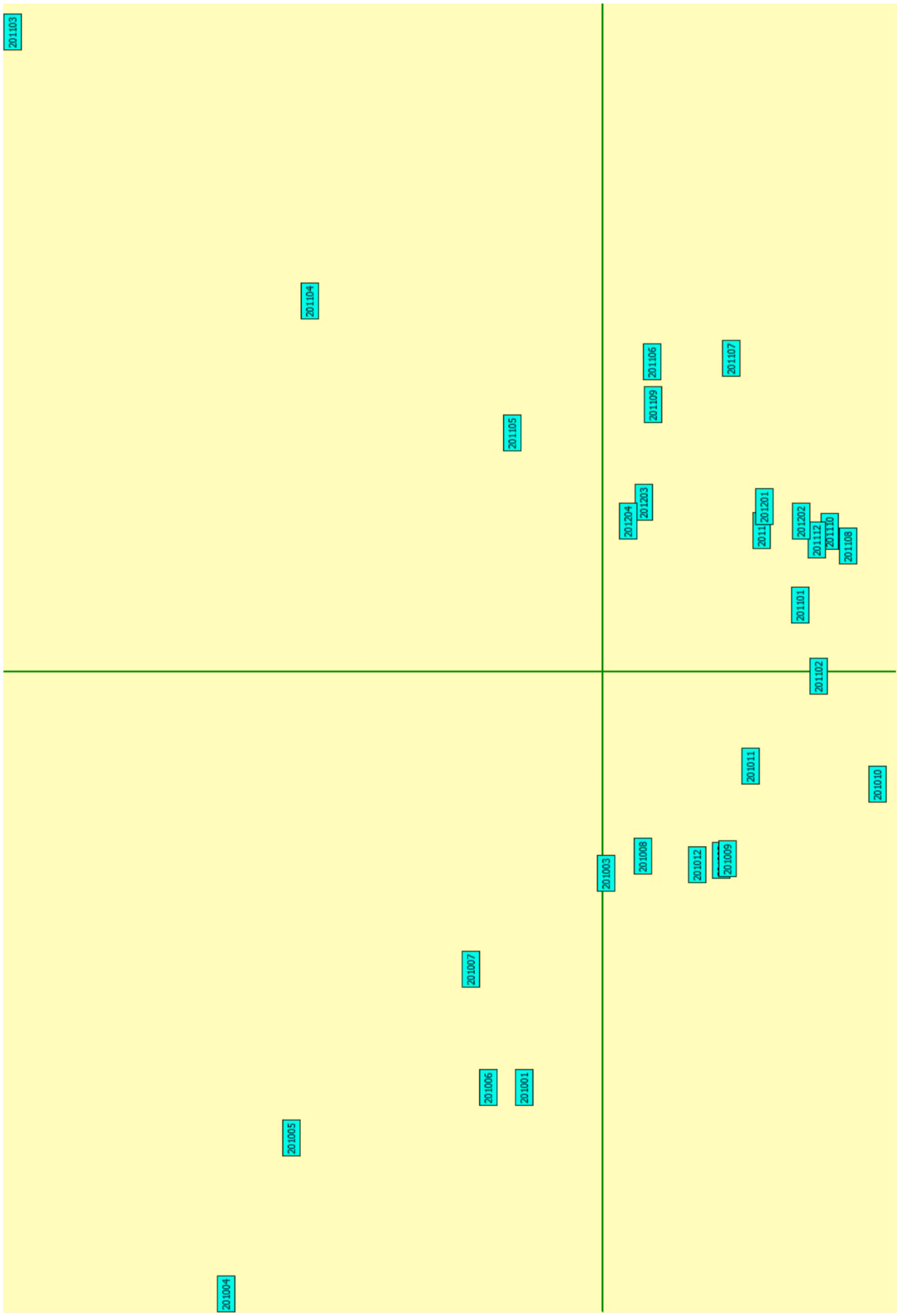


Figure 5.6 ENRG_FR de 2010 à 2012 : AFC sur l'ensemble des mois

Les articles d'ENRG_FR forment, de manière assez approximative, sur le plan AFC, figure 5.6 ci-dessus, une parabole pour la période de janvier 2010 à avril 2012, mais l'ordre chronologique général des mois n'a pas été totalement respecté. Cette parabole partiellement chronologique s'appelle l'effet Guttman : *« Le recours à des méthodes multidimensionnelles dans le traitement des données permet d'extraire des structures de données unidimensionnelles, lorsqu'elles existent, et l'on parle alors d'effet Guttman»* (Flament et Milland, 2005).

Cette AFC (figure 5.6 ci-dessus) a été réalisée à partir de l'ensemble des formes du sous-corpus comparable ENRG_FR et de leurs fréquences par mois sur la période retenue (de janvier 2010 à avril 2012).

L'évolution chronologique des mois sur la parabole témoigne la continuité mensuelle du vocabulaire récurrent, autrement dit, il y a une répétition de l'utilisation des mots entre les mois qui se suivent. L'écart entre les mois (dans la figure 5.6, chaque mois est représenté par un rectangle turquoise) s'explique de la manière suivante : plus les mois sont corrélés par leur covariance du vocabulaire, plus les mois sont proches. Rappelons que la covariance est un nombre permettant de déterminer, entre autres, les écarts de deux séries de données numériques par rapport aux moyennes (espérées).

Il y a une nette séparation, par l'axe vertical, entre les mois de l'année 2010 et ceux des années 2011 et 2012. Ceci s'explique par leurs proximités textuelles révélant ainsi la claire séparation des événements produits dans les deux groupements de périodes, à savoir, 2010 versus 2011 et 2012 (gauche et droite de l'axe vertical du plan AFC). Des séries chronologiques partielles se forment également à travers les quatre blocs du plan AFC :

- en haut à gauche : 2010-04, 05, 06, 07
- en bas à gauche : 2010-08, 09, 10, 11, 12
- en haut à droite : 2011-03, 04, 05
- en bas à droite : 2011-06, 07, 08, 09, 10, 11 (caché par 2012-01), 12 et 2012-01, 02, 03, 04
- sur l'axe horizontal de la partie gauche : 2010-03
- sur l'axe vertical de la partie basse : 2011-02

Mis à part ces séries chronologiques partielles, un certain nombre de mois non-chronologiques se groupent par leurs proximités sur le plan de l'AFC :

- en haut à gauche : 2010-01, 06
- en bas à gauche : 2010-02 (caché par 2010-09), 09, 12
- en haut à droite : aucun
- en bas à droite : 2011-08, 10, 12, 2012-02 ; 2011-11, 2012-01 ; 2012-03,04 et 2011-06, 09

En effet, l'existence de groupements de proximité textuelle est le témoin de l'emploi commun des formes du sous-corpus. Ce mécanisme langagier se manifeste par la répétition des formes dans les différentes parties du sous-corpus, soit un vocabulaire identique. Plus les unités d'AFC se rapprochent et s'entassent, plus l'emploi des formes reste commun, et vice-versa, plus elles s'éloignent, moins les formes sont communes.

Afin d'observer la relation textuelle interne des mois de 2010, nous écartons les mois des années 2011 et 2012 d'ENRG_FR dans le calcul de l'AFC, mois marqués en rose et rayé sur la figure 5.7. Nous procédons à un nouveau calcul de l'AFC (figure 5.7 ci-dessous) uniquement sur les douze mois de 2010, année riche en événements écologiques (se reporter à l'annexe B, section B.1).

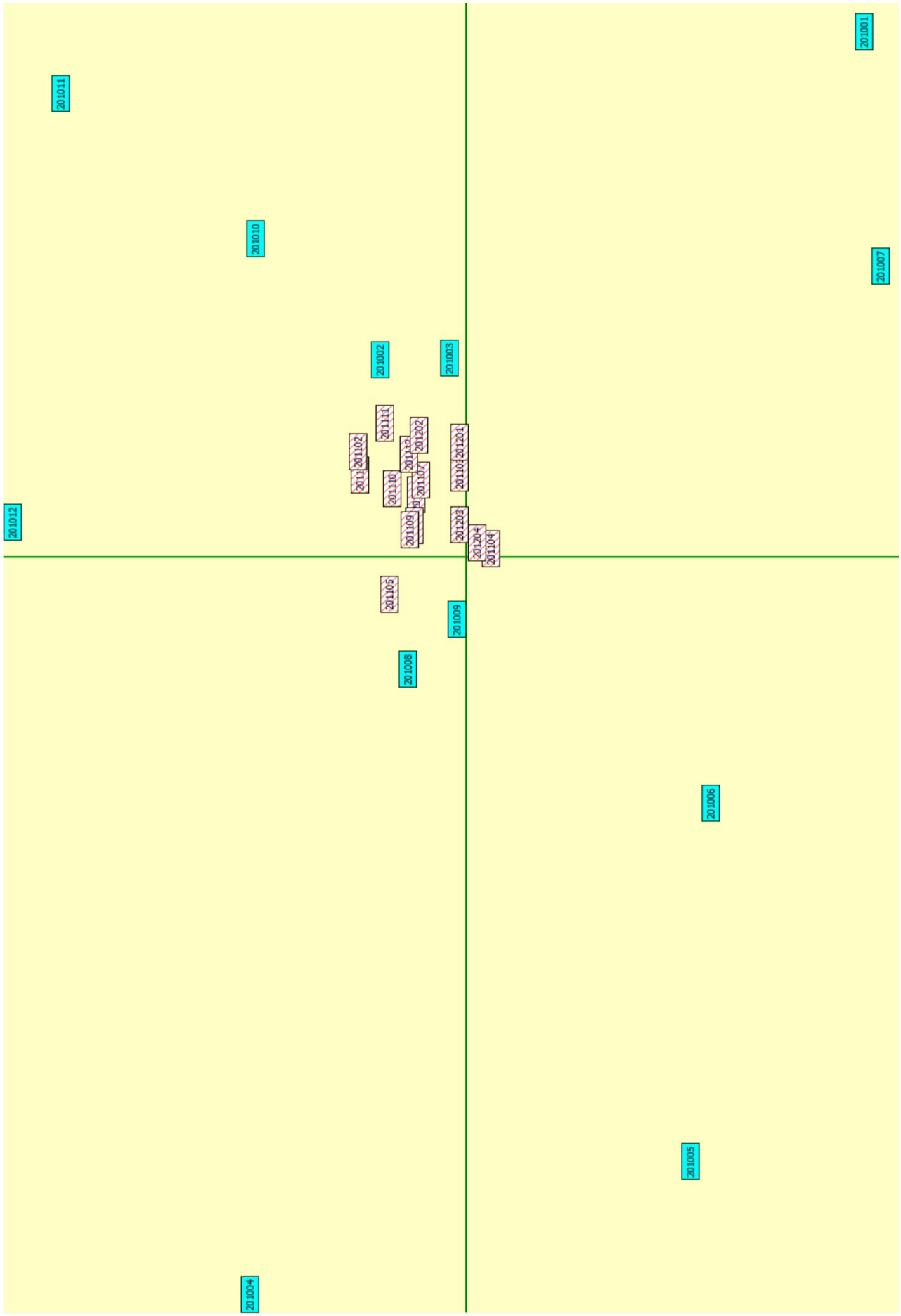


Figure 5.7 ENRG_FR 2010 : AFC sur l'ensemble des mois

Une pseudo-parabole est obtenue très approximativement par les mois de l'année 2010, figure 5.7 ci-dessus ; toutefois, l'ordre chronologique général des mois n'a pas été respecté. Il est à noter qu'il y a très peu de proximité entre ceux-ci sur le plan de l'AFC.

Groupement des mois de 2010 :

- en haut à gauche : 04, 08 et 09, avec les mois d'août et septembre relativement proches
- en bas à gauche : 05 et 06, aucune proximité constatée
- en haut à droite : 02, 03, 10, 11 et 12, avec les mois de février et mars relativement proches
- en bas à droite : 01 et 07, aucune proximité constatée.

Nous allons appliquer le même processus à ENRG_US, puis analyser les résultats des deux sous-corpus.

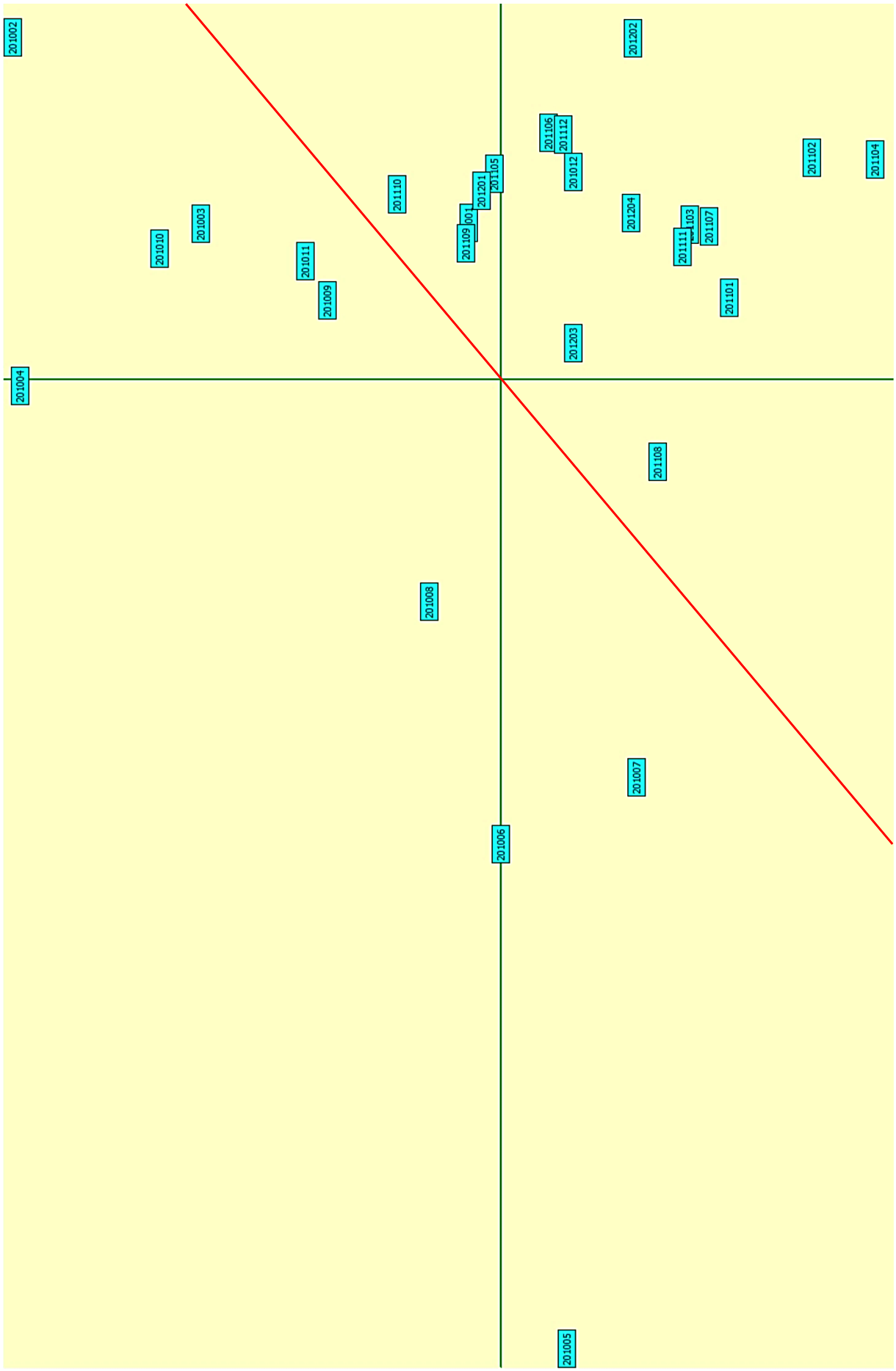


Figure 5.8 ENRG_US de 2010 à 2012 : AFC sur l'ensemble des mois

Le calcul de l'AFC sur l'ensemble des mois d'ENRG_US, figure 5.8 ci-dessus, nous montre que les mois d'avril, mai, juin juillet 2010 ainsi que le mois d'août 2011 se démarquent clairement des autres mois (gauche et droite de l'axe vertical du plan AFC).
Des séries chronologiques partielles se regroupent à travers les quatre blocs[145] du plan AFC :

- en haut à gauche : aucune
- en bas à gauche : 2010-05, 06, 07 et 08
- en haut à droite : 2010-09, 10 et 11
- en bas à droite : 2011-01, 02, 03, 04 et 2012-02, 03, 04
- sur l'axe horizontal de la partie gauche : 2010-06

Au-delà de ces séries chronologiques partielles, trois groupements de mois se manifestent :

- en haut à gauche : aucun
- en bas à gauche : aucun
- en haut à droite : 2010-01, 2011-05, 2011-09 et 2012-01
- en bas à droite : 2010-12, 2011-06, 2011-12 et 2011-03, 2011-07, 2011-11

Ces groupements reflètent leurs usages communs de lexique relayant les mêmes informations événementielles. Les calculs des spécificités nous livreront les points communs de ces groupements.

Rappelons que l'année 2010 se sépare des deux autres années sur le plan AFC par une diagonale à l'exception des mois de janvier et décembre. Une rotation des axes aurait pu être plus pertinente. Il s'agit de la causalité inverse : les usages connus de mots entraînent une représentation similaire sur l'AFC (Sharoff, Rapp, Zweigenbaum et Fung, 2013).

[145] Les quatre blocs formés par l'AFC n'ont pas de pertinence particulière, les axes peuvent changer de position en fonction des calculs. Il s'agit simplement d'un choix empirique.

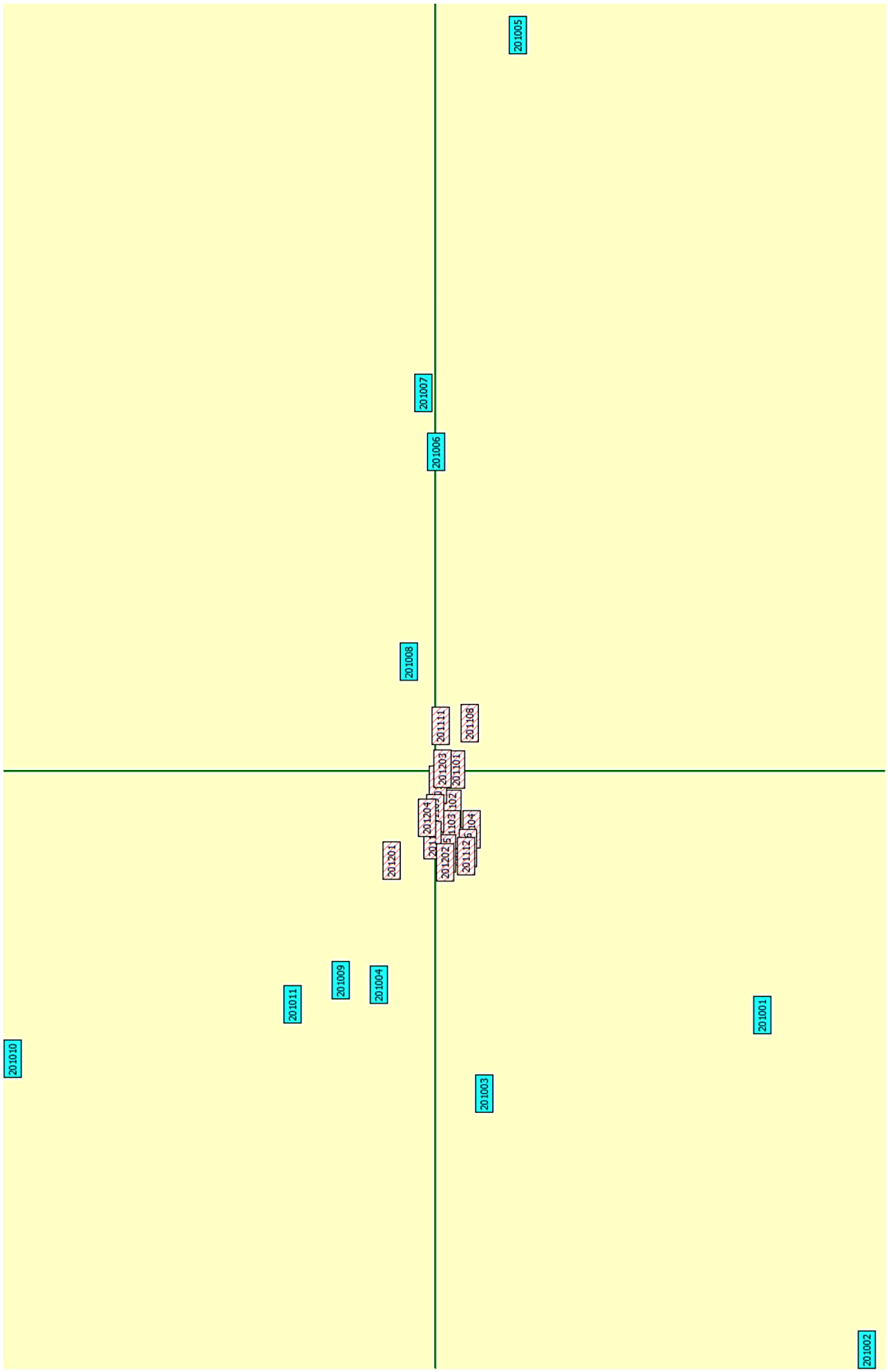


Figure 5.9 ENRG_US 2010 : AFC sur l'ensemble des mois

En écartant les mois des deux autres années d'ENRG_US, mois marqués en rose et rayés sur la figure 5.9, une parabole est obtenue très approximativement par les mois de l'année 2010, figure 5.9 ci-dessus, mais encore une fois, l'ordre chronologique général des mois n'a pas été respecté.

Groupement des mois de 2010 :

- en haut à gauche : 04, 09, 10 et 11, avec les mois d'avril, septembre et novembre relativement proches
- en bas à gauche : 01, 02 et 03, aucune proximité constatée
- en haut à droite : 06 et 08
- en bas à droite : 05
- sur l'axe horizontal dans la partie droite : 06

Selon les résultats des analyses des AFC, l'année 2010 permet de déceler une série chronologique générale commune et comparable sur ENRG_FR et ENRG_US.

Tableau 5.3 ENRG_FR et ENRG_US : récapitulatif des séries chronologiques des AFC

	Séries chronologiques dans ENRG_FR	Séries chronologiques dans ENRG_US
en haut à gauche :	2010-04, 05, 06 et 07	aucune
en bas à gauche :	2010-08, 09, 10, 11, 12	2010-05, 06, 07 et 08
en haut à droite :	2011-03, 04, 05	2010-09, 10 et 11
en bas à droite :	2011-06, 07, 08, 09, 10, 11, 12 et 2012-01, 02, 03, 04	2011-01, 02, 03, 04 et 2012-02, 03, 04

Selon le tableau 5.3 ci-dessus, les séries chronologiques partielles communes sont :

- 2010-05, 06, 07
- 2010-09, 10, 11
- 2011-03, 04
- 2012-02, 03, 04

Dans les séries ci-dessus, on observe une propriété commune dans la façon dont les deux journaux traitent les informations de ces périodes, c'est-à-dire que leurs productions de vocabulaire sont récurrentes pendant les mois qui se suivent. Nous faisons l'hypothèse que probablement ils relatent également les mêmes événements. Celle-ci sera démontrée à travers les analyses des 4 séries chronologiques communes dans les paragraphes suivants (section 5.3).

Tableau 5.4 ENRG_FR et ENRG_US : récapitulatif des groupements des AFC

	groupements dans ENRG_FR	groupements dans ENRG_US
en haut à gauche :	2010-01 et 06	aucun
en bas à gauche :	2010-02, 09 et 12	aucun
en haut à droite :	Aucun	2010-01, 2011-05, 2011-09 et 2012-01
en bas à droite :	2011-08, 10, 12, 2012-02 ; 2011-11, 2012-01, 2012-03,04 et 2011-06, 09	2010-12, 2011-06, 2011-12 et 2011-03, 2011-07, 2011-11

Dans le tableau 5.4, nous ne relevons aucun groupement commun entre les deux sous-corpus ENRG_FR et ENRG_US. Nous faisons l'hypothèse que les formes associées aux événements connus dans les deux mondes durant cette période sont absentes et que leur emploi est peu significatif, hypothèse vérifiée dans les paragraphes suivants. Les calculs des spécificités de chaque groupement nous donneront les raisons pour lesquelles les mois ci-dessus se regroupent.

5.3 Comparabilité et synchronicité : séries chronologiques, similitudes, restitutions

Les périodes communes des ENRG_FR et ENRG_US suivantes :

1. 2010 - (05, 06, 07)
2. 2010 - (09, 10, 11)
3. 2011 - (03, 04)
4. 2012 - (02, 03, 04)

forment quatre séries chronologiques mensuelles communes et comparables tout au moins au plan de la typologie textuelle pour ENRG. Les calculs de spécificités de la textométrie nous apportent une aide précieuse et efficace pour les analyses plus poussées des quatre séries, ancrées sur le thème *environnement* (se reporter à l'annexe B, spécificités de la textométrie). Nous écartons la série n° 4, puisque les analyses sur la répartition mensuelle du nombre d'occurrences des deux sous-corpus (section 5.1) par des retours aux articles, montrent déjà qu'il n'y a pas d'événements communs relatés par les deux journaux.

5.3.1 Spécificités des sous-corpus

Le calcul de spécificités est une méthode statistique permettant de mettre en évidence l'utilisation « atypique » d'une forme (ou plusieurs) dans une unité ou une quantité textuelle donnée, par rapport à sa fréquence totale. Dans notre cas, l'unité textuelle est la période chronologique. Cette utilisation atypique d'une forme (ou plusieurs) se traduit par le « sur-emploi » quand la répétition s'intensifie localement, ou le « sous-emploi » lorsque cette forme est particulièrement moins utilisée de manière locale (Lafon, 1980, 1981 ; Lebart et Salem, 1994).

Modèle hypergéométrique et calcul de spécificités de la textométrie

Le modèle hypergéométrique part d'un échantillon de N objets. Parmi ces N objets, un certain nombre que nous notons p possède une propriété notée P. Nous effectuons n tirages sans remise parmi les N objets de l'échantillon, et nous tentons de répondre à la question suivante : parmi ces n objets tirés, lesquels possèderont la propriété P ?

Par exemple, supposons qu'on dispose d'une urne de 10 boules, dont 6 sont noires et 4 rouges. On dira que les boules qui sont rouges possèdent la propriété C. On tire trois boules sans remise. On cherche alors à savoir combien de boules rouges figurent dans les trois boules tirées.

Soit X la variable aléatoire égale au nombre d'objets possédant la propriété C parmi les objets tirés. Nous notons P(X = k) la probabilité cherchée.

Nous remarquons en premier lieu que le nombre total de tirages possibles est $\binom{N}{n}$ [146](tirage de n objets parmi N). Par ailleurs, il est clair que si k est strictement plus grand que p, la probabilité cherchée est nulle, car au plus p objets possèdent la propriété P. Pour compter le nombre de tirages favorables (c'est-à-dire, le nombre de tirages de n boules où k objets ont la propriété p), nous procédons comme suit. Tirer k objets ayant la propriété p revient à tirer k objets parmi les p objets possédant la propriété P, puis à tirer les n – k objets restants parmi les N – p objets ne possédant pas la propriété p. La propriété cherchée est donc :

$$P(X = k) = \frac{\text{nombre de tirages favorables}}{\text{nombre de tirages possibles}} = \frac{\binom{p}{k}\binom{N-p}{n-k}}{\binom{N}{n}}$$

[146] Notons que $\binom{n}{p}$ est un coefficient binomial et se calcule par la formule : $\frac{n!}{(n-p)!p!}$

Applications au calcul de spécificités d'une unité textuelle :

Considérons un corpus comme un échantillon. Nous notons la fréquence absolue[147] d'une forme : n_0. Considérons le mode[148] de la distribution. Nous calculons les probabilités :

$P_{\sup n_0}$ qui est la probabilité de voir apparaître un nombre de formes supérieur ou égal (si n_0 est supérieur ou égal au mode) ou inférieur ou égal à n_0 (si n_0 est inférieur ou égal au mode).

$P_{\inf n_0}$ qui est la probabilité de voir apparaître un nombre de formes inférieur ou égal (si n_0 est supérieur ou égal au mode) ou supérieur ou égal à n_0 (si n_0 est inférieur ou égal au mode).

Nous fixons ensuite un niveau de probabilité arbitraire. Nous disons que la forme en question est :

une spécificité banale, si ni $P_{\sup n_0}$, ni $P_{\inf n_0}$ ne sont inférieures au seuil[149] de probabilité.

une spécificité positive, si $P_{\sup n_0}$ est en dessous du seuil. Dans ce cas, la forme est sur-employée dans le corpus.

une spécificité négative, si $P_{\inf n_0}$ est inférieure au seuil. Dans ce cas, la forme est sous-employée dans le corpus.

Pour tous nos calculs de spécificités, nous maintenons les paramètres « seuil » égal à 5 et « fréquence minimum » égale à 10, paramètres par défaut de Lexico 3. Afin d'optimiser la recherche de formes des noms communs dans les études de veille, qui touchent le cœur de nos thèmes, nous transformons la casse majuscule en minuscule des échantillons de mots suivants :

- Golfe, Pétrole, Marée, Energie(s), Nucléaire et Environnement pour ENRG_FR,
- *Gulf, Oil, Spill, Energies, Energy, Nuclear et Environment* pour ENRG_US.

Les outils Concordance[150] et TextPloreur (navigation textuelle) de Lexico 3 ont été mobilisés pour contextualiser les mots spécifiques sélectionnés, un autre type d'exemple est présenté également dans l'annexe B, figure B.8.

Parfois, des retours aux articles complets sont indispensables pour assurer la qualité de la restitution d'informations.

[147] Pour définir la notion de fréquence absolue, il est nécessaire de comprendre en premier lieu la notion de classe en statistiques. Dans le cas de données nombreuses, il peut être utile de partitionner l'ensemble des données en classes. Par exemple pour des mots, il est possible de faire les classes suivantes : les mots qui commencent par A, ceux qui commencent par B, etc… La fréquence absolue sera alors le rapport de l'effectif d'une classe à l'effectif de la population étudiée.
[148] Le mode de la distribution est la valeur la plus fréquente de la distribution.
[149] Le seuil de probabilité est une probabilité (nombre compris entre 0 et 1). Il est fixé par l'utilisateur, de manière arbitraire. Il sera le plus souvent choisi en fonction des hypothèses faites sur la fréquence des formes du corpus choisi (Lebart et Salem 1994, chapitre 6 : 171-198).
[150] Concordance : voir définition dans le glossaire.

5.3.2 Série n°1 : mai, juin et juillet 2010

Tableau 5.5 ENRG_FR : sélection des spécificités positives de mai, juin et juillet 2010

Forme	Frq. Tot.[151]	Fréquence[152]	Coeff.[153]
BP	864	617	***
noire	616	365	***
golfe	405	232	***
marée	597	359	***
Hayward	86	76	***
pétrole	998	361	***
puits	511	227	***
Louisiane	157	109	***
Mexique	435	208	***
(…)	(…)	(…)	(…)

Tableau 5.6 ENRG_US : sélection des spécificités positives de mai, juin et juillet 2010

Forme	Equivalent en français	Frq. Tot.	Fréquence	Coeff.
gulf	golfe	444	239	***
BP	BP	398	220	***
oil	pétrole/huile	1678	715	***
spill	renversement/marée (noire)	486	247	***
spills	déversements/marée (noire)	103	69	*33*
Louisiana	Louisiane	99	62	*28*
leak	fuite	96	60	*27*
barrels	barils	93	55	*23*
Deepwater	nom de la plate-forme	74	47	*22*
marshes	marais	52	37	*20*
(…)	(…)	(…)	(…)	(…)

Dans les tableaux 5.5 et 5.6, nous avons sélectionné les formes positives les plus spécifiques, classées dans la partie supérieure des deux tableaux avec les coefficients tendant vers l'infini. Celles-ci révèlent une information identique dans les deux sous-corpus caractérisant l'événement majeur survenu à cette période, à savoir, l'explosion de la plate-forme *Deepwater Horizon* de *BP* dans le golfe du Mexique. Concernant ENRG_FR, le tableau apporte une précision supplémentaire mentionnant le nom de l'ancien directeur général de *BP*, *Hayward* (tableau 5.5 ci-dessus). L'information principale s'avère parallèle et synchronisée dans les deux sous-corpus.

[151] Frq. Tot. : Fréquence Totale (toutes les occurrences du mot dans tout le corpus)
[152] Fréquence : toutes les occurrences du mot dans la partie sélectionnée du corpus
[153] Coeff. : Coefficient de spécificité. Le coefficient de spécificité est limité à deux chiffres. Il utilise la valeur absolue de l'exposant négatif de la probabilité exprimée en notation exponentielle et permet de rendre compte l'ordre de grandeur de cette probabilité (un indice de classement) : probabilité de 1% = 1^{E2} => coeff.2; probabilité de 1/1000 = 1^{E3} => coeff.3, etc. Plus la probabilité est petite, plus la spécificité est remarquable, en plus ou en moins (Pineira et Tournier, 2009).

5.3.3 Série n°2 : septembre, octobre et novembre 2010

Tableau 5.7 ENRG_FR : sélection des spécificités positives de septembre, octobre et novembre 2010

Forme	Frq. Tot.	Fréquence	Coeff.
Mediator	222	88	*35*
Cancun	189	78	*32*
valorisation	37	33	*31*
multinationales	74	46	*30*
médicament	209	79	*30*
Marzano	27	27	*29*
Péré	27	27	*29*
Manganella	25	24	*25*
Nagoya	43	32	*25*
climatique	785	166	*25*
climatosceptiques	46	32	*23*
Copenhague	228	74	*23*
valves	34	27	*23*
biodiversité	256	78	*22*
nature	325	88	*21*
(…)	(…)	(…)	(…)
Danube	22	19	*18*

En France, les mots portant les informations convergent davantage vers les quelques événements suivants :

- Scandale du médicament Médiator (Mediator, médicament, valves).
- Conférence sur le climat à Cancun au Mexique (Cancun, Copenhague, climatique, climatosceptiques).
- Conférence sur la biodiversité à Nagoya au Japon (biodiversité, nature).
- Réponses aux questions des internautes du Monde.fr par Péré - Marzano et Antonio Manganella (multinationales).
- Chiffrage du coût que fait peser à terme sur l'économie mondiale l'absence de politique ambitieuse de protection de la biodiversité. L'économiste de l'environnement Yann Laurans explique ce que l'on pourrait attendre de la mise en œuvre à l'échelle mondiale de sa méthode (valorisation).
- Pollution du Danube en Hongrie.

Selon le tableau 5.7 ci-dessus, le scandale médical de Médiator ainsi que les questions et risques liés au changement climatique et à l'environnement préoccupent les Français. Pour cette période, l'événement international (Danube) n'est pas leur souci prioritaire.

Au vu de cette restitution de veille, notre constat est conforme au résultat obtenu par l'enquête menée par l'Ifop en mars 2011 (annexe E, tableau E.1), illustrant les préoccupations environnementales, une priorité absolue des Français avant le 11 mars 2011, date de Fukushima. Celles-ci basculent vers les risques liés au nucléaire après la catastrophe. Malgré ces circonstances désastreuses, les risques liés au changement climatique se classent toujours en deuxième position demeurant toujours une préoccupation majeure des Français.

Tableau 5.8 ENRG_US : sélection des spécificités positives de septembre, octobre et novembre 2010

Forme	Equivalent en français	Frq. Tot.	Fréquence	Coeff.
Hungary	Hongrie	27	26	*26*
Facebook	Facebook	39	30	*23*
Microsoft	Microsoft	26	23	*21*
Bishnoi	Bishnoi	18	18	*19*
ya	Ya	16	16	*17*
Kyo	Kyo	16	16	*17*
cork	liège	18	17	*17*
sludge	boue	60	31	*16*
factory	usine	69	32	*15*
Waikiki	Waikiki	16	15	*15*
Champagne	Champagne	16	15	*15*
Danube	Danube	13	13	*14*
reservoir	réservoir	43	22	*12*

Aux États-Unis, une grande variété de mots se manifeste dans les spécificités de cette période. Il a fallu consulter attentivement les articles via les concordances de ces mots et les articles associés pour restituer les événements caractérisant cette série.

Informations restituées du tableau 5.8 :

- Les boues (*sludge*) rouges et leur réservoir de stockage en Hongrie (*Hungary*) polluent et menacent de polluer à nouveau le Danube.
- La construction d'un nouveau centre de données de *Facebook* en Caroline du Nord risque de polluer plus à cause du mode de production d'électricité.
- La Russie utilise *Microsoft* pour réprimer la dissidence de manifestations contre la réouverture d'une usine (*factory*) de papier située à proximité du Lac *Baïkal*, détenant 20% de l'eau douce de la planète.
- Présentation des *Bishnoï*, une communauté rurale hindoue du Rajasthan en Inde, ayant de fortes consciences écologiques depuis des lustres.
- La construction d'un hôtel par le groupe japonais *Kyo-Ya* menace de gâcher le paysage de *Waikiki (à Hawaï)*.
- Une campagne publicitaire prônant l'utilisation des bouchons de liège plutôt que les capsules, une action financée par l'état portugais afin de promouvoir son exportation de liège.
- Le champagne est pointé du doigt à cause du poids des bouteilles en verre épais, générant plus de production carbonique lors de leurs fabrications et transports.

Pour cette série, il est à noter qu'au regard de la comparaison des restitutions, les informations révélées par les formes hautement spécifiques de la même période ne reflètent pas d'événements identiques dans les deux mondes, hormis l'événement concernant la pollution du *Danube*. Cette forme n'apparait pas en tête du tableau. Cette faiblesse de classement pourrait être interprétée comme un événement secondaire par rapport aux autres par le Monde. Une très faible synchronicité se détecte dans cette série supposée comparable.

En dépit de l'anémie de la production de la forme *Danube*, l'approche de la textométrie permet toutefois de détecter ces signaux faibles, difficiles à percevoir et à identifier en raison de leur quantité exsangue (Ansoff, 1975 ; Hermel, 2010 : 94).

5.3.4 Série n°3 : mars et avril 2011

Les tableaux 5.9 et 5.10 vont nous faire découvrir les spécificités propres à cette troisième série.

Tableau 5.9 ENRG_FR : sélection des spécificités positives de mars et avril 2011

Forme	Frq. Tot.	Fréquence	Coeff.
Fukushima	672	265	***
réacteur	507	184	***
Tchernobyl	184	103	***
Tepco	192	105	***
accident	557	180	***
centrales	608	192	***
Japon	631	262	***
séisme	516	196	***
réacteurs	534	179	***
Tokyo	176	104	***
centrale	960	346	***
Cailletaud	47	47	***
nucléaire	2197	710	***
Japonais	105	70	*49*
catastrophe	671	185	*47*
mars	807	203	*45*
japonais	217	96	*45*
japonaise	134	73	*42*
sarcophage	39	37	*38*
combustible	156	70	*34*
confinement	122	62	*34*
radioactivité	210	81	*33*
électricité	750	169	*32*

Tableau 5.10 ENRG_US : sélection des spécificités positives de mars et avril 2011

Forme	Equivalent en français	Frq. Tot.	Fréquence	Coeff.
(...)	(...)	(…)	(…)	(…)
gas	gaz	825	175	*19*
wastewater	eaux usées	42	29	*19*
(...)	(...)	(…)	(…)	(…)
Clinton	Clinton	47	19	*8*
chromium	chrome	19	12	*8*
fracking	fracking	94	29	*8*
Japan	Japon	81	26	*8*
(...)	(...)	(…)	(…)	(…)
Merkel	Merkel	23	11	*6*

Etant donné que cette période a laissé d'importantes traces dans l'ensemble de la presse mondiale, nous tentons de restituer l'information principale avec le plus grand soin en nous focalisant sur la synchronicité de l'information événementielle.

5.3.5 Les restitutions d'informations

Pour ENRG_FR, toutes les formes de la partie supérieure du tableau 5.9 se rapportent, sans aucune exception, à la catastrophe de Fukushima.

Or, la spécificité positive de la partie supérieure du tableau 5.10 du sous-corpus américain ne comporte aucun mot se rapportant à l'événement mondialement connu. La synchronicité est quasi-nulle à l'exception de la forme *Japon*. Ceci est dû au classement initial de la rubrique *Environment*, qui est une sous-rubrique de la rubrique *Sciences,* ou aux différentes méthodes d'archivage des articles des événements internationaux du journal. En effet, nous constatons que les événements internationaux se trouvent dans la rubrique « *World* », en l'occurrence, « *World, Asia-Pacific* ».

Les mots du tableau 5.10 ne se réfèrent qu'aux événements environnementaux locaux des États-Unis.

La restitution des événements sélectionnés et relevés par la série nous apporte les informations suivantes :

- *Walmart*, entreprise américaine multinationale spécialisée dans la grande distribution, dévoile un plan rendant la chaîne d'approvisionnement plus verte (*gas*).
- Une *start-up* de *Silicon Valley* dit qu'elle a trouvé un moyen de capturer les émissions de dioxyde de carbone provenant du charbon et du gaz (*gas*) des plantes et de les verrouiller dans le ciment. Cela fonctionne sur une échelle de masse, la société *Calera* pourrait transformer ce carbone en « or ».
- L'Administration américaine a fait valoir des exemptions pour l'industrie du pétrole et du gaz (*wastewater*) malgré les pollutions qu'elle a engendrées dans les années 1980.
- Lancement d'une fondation internationale : *Clinton Climate Initiative*.
- Problèmes de traitement des déchets de *chrome* dans le comté d'Hudson du New Jersey.
- Des études suggèrent que le méthane, composant principal du gaz naturel, s'échappe dans l'atmosphère en quantités bien supérieures à celles que nous pensons lors des exploitations et transports, et forme une des causes principales du réchauffement de la planète. Les quantités

de gaz échappées sont comparables à celles émises par le *fracking* des puits de gaz de schiste (*fracking*[154]).

- Que ce soit pour des raisons électorales ou non, Madame *Merkel* a mis en garde son propre parti et l'industrie nucléaire, qu'il n'y aurait pas de retour au *statu quo* pour le nucléaire en Allemagne, décision prise bien avant la crise au *Japon*.

5.3.6 Analyses transversales des séries chronologiques entre ENRG_FR et ENRG_US

Rappelons que dans notre hypothèse, les informations événementielles se sont produites dans les périodes communes, les dimensions des deux AFC (ENRG_FR et ENRG_US) sont semblables et les séries chronologiques communes sont comparables.

Pour la première série chronologique (mai, juin et juillet 2010), les formes pour ENRG_FR et ENRG_US les plus fréquentes traitent du même sujet : l'explosion de la plate-forme *Deepwater Horizon* et la pollution de la faune et de la flore qui a suivi cette catastrophe. Bien que cet événement impacte directement les côtes américaines du golfe du Mexique, la Une de la presse française du journal *Le Monde* se fait largement l'écho de cette nouvelle alors que la France n'est pas directement concernée par cette pollution. A travers cette étude comparative que nous venons de réaliser de manière empirique, l'une des explications possibles peut être la suivante : la France a connu par le passé des pollutions similaires, notamment les naufrages de pétroliers provoquant des marées noires sur les côtes françaises, le *Torrey Canyon* en 1967 et surtout celui de l'*Amoco Cadiz* en 1978 considéré comme l'une des plus importantes catastrophes écologiques du XX^{e} siècle. La France, en matière de perception de protection environnementale, est historiquement et géographiquement plus sensibilisée que l'Amérique.

Quant aux deux autres séries, les formes entre les deux sous-corpus sont beaucoup plus éclectiques. Celles-ci peuvent s'expliquer, d'une part par les contextes politiques, économiques, et sociologiques différents entre les deux pays, d'autre part par les méthodes de classement et d'archivage des deux journaux qui ne sont pas les mêmes. Toutefois, le constat de ce différent classement nous mène à penser que la perception de l'environnement et sa protection n'est pas liée intrinsèquement à la représentation du nucléaire chez les Américains ; or le mode de vie des Français est à la fois imprégné et averti par l'esprit et les enjeux du nucléaire.

Pour la deuxième série chronologique (septembre, octobre et novembre 2010), les formes les plus spécifiques d'ENRG_FR expriment bien les préoccupations françaises, notamment les risques liés aux changements climatiques[155], car nous retrouvons les noms des conférences sur la biodiversité, le climat (*Cancun, Nagoya, Copenhague*) et leurs sujets de réflexion (*climatique, biodiversité, nature*), alors que pour ENRG_US les formes spécifiques sont diverses, mais reflètent des faits associés à la pollution aux États-Unis.

Quant à la troisième série (mars et avril 2011), le journal Le Monde laisse une grande place à la catastrophe de Fukushima. Historiquement, la France est le pays de la découverte de l'atome et ne possède pas de sources d'énergies fossiles, par conséquent, les politiques de l'époque ont misé très tôt sur le développement d'une énergie tout nucléaire[156], contrairement aux Etats-Unis, où les sources

[154] La fracturation hydraulique : fracking
http://www.futura-sciences.com/magazines/terre/infos/dico/d/geologie-fracturation-hydraulique-9048/
(consulté le 09/08/2015)
[155] Annexe E : Enquête Ifop réalisée du 7 au 10 mars 2011 et du 24 au 25 mars 2011
[156] Annexe D : Le nucléaire et la politique française

d'énergie fossile sont abondantes[157], et le nucléaire civil demeure un tabou pour les Américains (Chavardès, 2009). Les formes sont associées à des sujets qui traitent non seulement des causes de la catastrophe, mais aussi des thèmes liés au domaine nucléaire comme par exemple *sarcophage, confinement, combustible, radioactivité*. La France est l'un des rares pays où le nucléaire est la source d'énergie principale. Les Américains sont plus orientés vers la recherche de nouvelles technologies exploitant les sources d'énergie fossile ou probablement d'énergie renouvelable, en particulier la recherche en gaz (*gas*) à l'aide de techniques telles le *fracking*.

La restitution d'informations de la sélection des trois séries communes des deux sous-corpus confirme notre hypothèse définie en début de la section 5.2.

Et la Chine ? Le chapitre suivant est consacré à une étude approfondie sur le sous-corpus ENRG_CN. Les analyses comparatives trilingues nous dévoileront la face cachée des trois mondes.

5.4 Cooccurrences et poly-cooccurrences évolutives autour de la forme *EPR*

La veille active consiste à restituer le contexte énergétique dans les trois mondes à travers les sous-corpus trilingues définis à l'aide des calculs d'AFC et spécificités, puis à analyser les informations événementielles émanées des calculs de la textométrie. Ensuite, une étude de veille ciblée, épaulée par les calculs de cooccurrences et poly-cooccurrences, sera consacrée à l'énergie nucléaire en particulier à un nouveau type de réacteur qu'est l'EPR dans les trois mondes. Les interprétations de nos restitutions nous permettront de prévoir et d'anticiper des pistes de réflexion sur la politique énergétique à mener dans le domaine du nucléaire.

5.4.1 Le calcul de cooccurrences et poly-cooccurrences

Nous allons commencer par une brève présentation des notions de cooccurrences (Lafon, 1981; Martinez, 2003) et poly-cooccurrences (Martinez, 2000 ; MacMurray et Shen, 2010).

La cooccurrence se réfère à l'emploi simultané de deux formes différentes dans le même contexte. On dit que deux formes sont cooccurrentes lorsque la présence d'une de ces deux formes dans un corpus implique la présence de l'autre forme. Par exemple :

lampe/lumière ; plage/mer ; patron/employé

Il est probable que chacun des trois couples de formes citées ci-dessus peut apparaitre dans une même séquence textuelle. Or, tout couple de mots n'est pas obligatoirement une paire de cooccurrences. Il est donc nécessaire de s'assurer que la fréquence de la co-présence des deux formes recherchées dans le corpus n'est pas due au hasard. Ainsi, une importante co-répétition de deux formes dans une même séquence textuelle peut révéler des informations intéressantes voire stratégiques. Le degré de cooccurrences d'une paire de formes se mesure au moyen de certaines techniques statistiques telles que le modèle hypergéométrique.

Nous calculons la cooccurrence de deux formes sur une unité textuelle du corpus en faisant le produit de deux probabilités : la probabilité attachée à la première forme et la probabilité attachée à la seconde forme. Ces deux probabilités s'estiment au moyen d'une loi hypergéométrique décrite plus haut (section 5.3) et de paramètres :

[157] Se reporter au chapitre 2, section Des énergies fossiles.

- N : le nombre total de formes dans le corpus.
- n : le nombre total de formes dans une unité textuelle du corpus choisi.
- f : la fréquence de la forme considérée dans le corpus.

Le produit de ces deux probabilités donne la probabilité de rencontrer simultanément ces deux formes dans la même unité textuelle du corpus. En segmentant le corpus en unité textuelle (phrase, paragraphe, article, etc.), nous obtenons le nombre attendu d'apparitions simultanées de ces deux formes dans le corpus.

Par exemple, imaginons un corpus d'un certain nombre de mots, et cherchons à savoir si les formes *père* et *mère* sont cooccurrentes. Pour répondre à cette question, il faudra comparer le nombre d'occurrences simultanées des formes *père* et *mère* au nombre attendu d'apparition simultanée de ces deux formes dans le corpus. Si le nombre d'occurrences simultanées des formes *père* et *mère* est très supérieur au nombre attendu d'apparition simultanée de ces deux formes dans le corpus, il sera alors raisonnable de penser qu'il existe une relation de cooccurrences entre les formes *père* et *mère*.

La notion de poly-cooccurrences est similaire à celle de cooccurrences à la différence que nous ne considérons plus une forme associée à une autre mais à plusieurs. Nous obtenons une fois que nous avons calculé les relations de cooccurrences, un graphe où chacun des nœuds représente les différentes formes du corpus, et où chacune des arêtes du graphe donne les liens de cooccurrences entre les différentes formes.

5.4.2 Un exemple d'application de la notion de cooccurrences

Considérons le corpus suivant de 217 mots (deux paragraphes entiers issus d'un article d'ENRG_FR):

« La semaine dernière, treize autres faiblesses avaient déjà été constatées par le gendarme du nucléaire, tandis que samedi, on apprenait que le site n'était pas totalement aux normes sismiques. Et l'on ne compte plus les lettres, rapports ou documents internes égrénant les lacunes de la future centrale, tant vantée par le gouvernement, et présentée comme "la plus sûre au monde" par le fabricant Areva. En Finlande et en Chine aussi, où sont construits trois autres réacteurs du même type, les chantiers accumulent d'importants retards et sont la cible de nombreuses critiques. L'EPR, d'un fleuron du nucléaire français, est ainsi en passe de devenir l'une des technologies les plus décriées. Sur le papier, le réacteur pressurisé européen (European pressurized reactor), conçu par Areva et l'allemand Siemens dans les années 1990, est censé représenter, en termes de sûreté, un modèle dans le monde. Le réacteur, d'une puissance de 1 650 mégawatts, aurait été conçu pour résister à la chute d'un avion gros porteur, et ses multiples systèmes de sécurité doivent le mettre à l'abri d'un accident détruisant le cœur du réacteur. Les piscines de refroidissement des combustibles usés seront même protégées par une enceinte de confinement. Au final, le risque de prolifération des matières radioactives serait quasiment nul. »

Considérons la paire de mots *réacteur* et *nucléaire*. Dans le présent texte nous avons :
-3 occurrences de *réacteur*
-2 occurrences de *nucléaire*
Considérons la probabilité de trouver k occurrences de la forme *réacteur* :
Notons X la variable aléatoire qui donne le nombre d'occurrences de *réacteur* dans le corpus. Cette probabilité est nulle si k est strictement supérieur à 3.

Dans notre exemple, N = 217 car il y a 217 mots dans le corpus.
Nous choisissons ici n = 10 mots tirés dans le corpus sans remise.

Si k est compris entre 0 et 3, nous avons :

$$P(X = k) = \frac{\binom{3}{k}\binom{217-3}{10-k}}{\binom{217}{10}}$$

choix de k formes *réacteur* parmi les 3 formes du corpus puis choix des 10 – k formes restantes parmi les formes du corpus différentes de *réacteur*.

Rappelons que $\binom{n}{p}$ est un coefficient binomial et se calcule par la formule : $\frac{n!}{(n-p)!p!}$

Par exemple, $\binom{7}{3}$ se calcule : $\frac{7!}{(7-3)!3!}$ avec par exemple, 7! = 7x6x5x4x3x2x1 = 5040

Nous avons donc :

$$P(X = k) = \frac{3!214!(197)!10!}{(k-3)!k!(10-k)!(204+k)!(217)!}$$

Nous obtenons alors pour k = 0, 1, 2 ou 3 les valeurs suivantes de P(X = k) :

k	0	1	2	3
P(X = k)	0,87	0,12	0,005	0,005

De même en notant Y la variable aléatoire donnant le nombre k d'occurrences de la forme *nucléaire* nous avons :

$$P(Y = k) = \frac{\binom{2}{k}\binom{217-2}{10-k}}{\binom{217}{10}}$$

$$= \frac{2!215!(197)!10!}{(k-2)!k!(10-k)!(205+k)!(217)!}$$

Nous obtenons alors pour après calculs pour k = 0, 1 ou 2 les valeurs de P(Y = k) résumées dans le tableau ci-dessous.

k	0	1	2
P(Y = k)	0,91	0,08	0.01

En cas d'indépendance, nous avons donc les probabilités jointes suivantes :

Le tableau se lit comme suit à l'intersection de la ligne i (i = 0, 1, 2 ou 3) et j (j = 0, 1 ou 2) la probabilité d'avoir X = i et Y = j est la valeur donnée à l'intersection des lignes i et j. Par exemple, à l'intersection de la ligne 1 et de la colonne 1 nous avons la probabilité que X = 0 et Y = 0, c'est-à-dire, la probabilité qu'il n'y ait aucune occurrence de *réacteur* ni de *nucléaire*, qui vaut 0,79. En effet, les variables aléatoires X et Y sont calculées dans l'hypothèse où elles sont indépendantes. On a donc par exemple P(X = 0 et Y = 0) = P(X = 0)xP(Y = 0) = 0,91x0,87 = 0,79. On obtient de la même façon les autres résultats du tableau ci-dessous.

Y \ X	0	1	2	3
0	0,79	0,11	0,005	0,005
1	0,07	0,009	0,0004	0,0006
2	0,008	0,001	0,00005	0,0005

On somme ensuite les probabilités où il y a 0 occurrence de *réacteur* et 1, 2 ou 3 occurrences de *nucléaire* (cas où seule la forme nucléaire apparaît) et les probabilités où il y a 0 occurrence de *nucléaire* et 1, 2 ou 3 occurrences de *réacteur* (cas où seule la forme réacteur apparaît).

Nous obtenons alors :

P(X = 0 et Y = 0) + P(X = 0 et Y = 1) + P(X = 0 et Y = 2) + P(Y = 0 et X = 1) + P(Y = 0 et X = 2) = 0.983

Nous en déduisons la probabilité de trouver ces deux formes simultanément par complémentaire car la somme des probabilités vaut 1 : 1 - 0.98 = 0.02.

Les résultats ci-dessus montrent qu'il y a environ 98 % de chances de ne pas trouver simultanément les mots *réacteur* et *nucléaire* dans ce petit texte. Il y a donc 2 % de chances de trouver simultanément les deux mots dans ces deux paragraphes lorsque l'on y tire 10 mots au hasard.

Nous notons l'espérance[158] respectivement de X et Y par E(X) et E(Y). Ici nous avons :

E(X) = 0x0.87 + 1x0.12 + 2x0.005 + 3x0.005 = 0.145
E(Y) = 0x0.91 + 1x0.08 + 2x0.01 = 0.1

Les probabilités calculées ci-dessus montrent que dans un segment non contigu de 10 mots de notre texte d'étude où le tirage s'effectue au hassard, nous nous attendons à avoir :

10*E(X) = 10x0,145 = 1,45 occurrences de *réacteur*
10*E(Y) = 10x0,1 = 1 occurrence de *nucléaire*.

Dans le cas où le segment est contigu, le dénominateur n'est plus $\binom{217}{10}$, mais simplement 217 – 10 + 1.

Si on trouve plus d'occurrences que le nombre d'occurrences attendu dans un segment de 10 mots, la relation de cooccurrences entre les mots *réacteur* et *nucléaire* sera probable.

Fin de l'exemple.

En dépit de la complexité de la restitution des informations, une veille générale par poly-cooccurrences dite « aveugle », c'est-à-dire, une restitution de contexte sans retour systématique au corpus, sur l'ensemble des deux sous-corpus ENRG_FR et ENRG_US, mais ciblée sur les thèmes nucléaire et énergie, a été exécutée bien avant toutes les analyses poussées de chaque période. Cette veille générale a permis de concevoir deux dictionnaires d'événements autour de nos thèmes de recherche. Les résultats et les figures sont consultables dans l'annexe H pour la France et l'annexe I pour les Etats-Unis. Quant à ENRG_CN, la restitution sera traitée dans le chapitre 6.

La restitution d'informations des résultats des deux sous-corpus nous révèle les informations générales et non exhaustives suivantes. Un exemple de retour au contexte sera exposé au cours de ces constats.

[158] L'espérance d'une variable aléatoire Z discrète (c'est-à-dire à valeurs dans un ensemble E fini ou infini dénombrable comme par exemple l'ensemble des entiers naturels) se calcule par la formule : $\sum_{k \in E} k\, P(Z = k)$ Ici X est ici à valeurs dans l'ensemble [0,1,2,3] et Y est à valeurs dans l'ensemble [0,1,2] et on a donc l'espérance par exemple de X : E(X) qui vaut E(X) = 0x0,87 + 1x0,12 + 2x0,005 + 3x0,005.

5.4.3 Poly-cooccurrences autour d'ENRG_FR

Il s'avère que les calculs individuels des deux formes *nucléaire* et *énergie* produisent les mêmes résultats avec les mêmes paramètres (annexe H). Cela s'explique par l'omni-coprésence des deux formes qui s'auto-transposent entre elles par la linéarité phraséologique poly-cooccurrente dans ce sous-corpus, à savoir que, la source d'énergie prédominante est le nucléaire.

Formes fédérées à la forme-pôle : poly-cooccurrences

De plus, nous constatons qu'il y a une absence totale de consécutivité des formes poly-cooccurrentes, formes récupérées à partir des formes-pôles *nucléaire* et *énergie* (*cf.* annexe H, figure H.1 et figure H.3). On n'obtient que des mots individuels, noms communs ou noms propres dans les résultats, rarement un enchainement de mots tel que l'on peut le trouver dans les résultats d'ENRG_US (par exemple, *Accident>Mile>1979>Three*). Dans ce travail, nous appelons ce type de résultats, une suite de mots fédérés à la forme-pôle, contrairement aux mots individuels que nous nommons informations fragmentaires (à la forme-pôle).

Le mécanisme de l'attraction lexicale des poly-cooccurrences n'a pas donné davantage de résultats sur le sous-corpus français avec les paramètres par défaut du logiciel Trameur[159] (co-fréquence 2, seuil 10, contexte . !?). Le paramètre « contexte . !? » signifie que l'unité de calcul des poly-cooccurrences se base approximativement sur chaque phrase du corpus. En effet, la poly-cooccurrence fonctionne par l'attraction lexicale du renouvellement de formes dans chacune des unités phraséologiques, elle récupère chaque nouvelle forme répétée le même nombre de fois que la forme-pôle choisie, dans la phrase (ou le segment) qui suit.

L'absence des enchainements de mots dans les résultats français s'explique par l'absence de répétition de nouvelle forme dans la phrase suivante. Ce phénomène est dû au mécanisme anaphorique de cette langue. L'utilisation des anaphores, appelée en pragmatique linguistique, la deixis, le déictique ou l'emploi déictique, due au bon usage du français serait la véritable cause de ce phénomène dans les articles du Monde. Par exemple, *nucléaire* serait remplacé par *atome*, *François Hollande* par *le Président de la République* ou *le Chef de l'Etat*, etc. Ce genre de pratique ferait tomber la linéarité transposée[160] des poly-cooccurrences.
Une grande variété de formes a été repérée par rapport à ENRG_US, en particulier, un grand nombre de toponymes indiquant le nom géographique des sites liés au domaine nucléaire :

- Sites de centrales nucléaires : *Flamanville* (*Manche*), *Fessenheim* (*Haut-Rhin*), *Tricastin* (Drôme), *Dampierre-en-Burly* (Loiret) et *Penly* (Le pays de Caux, Seine-Maritime),
- Lieux de centres de recherches : *Cadarache*, *Marcoule*,
- Noms de sites étrangers associés aux catastrophes nucléaires : *Three Mile Island, Tchernobyl, Fukushima.*

Les termes liés aux activités nucléaires : *coût* (*EPR*, *électricité*), *sécurité*, *convoi*, *retraitement*, *radioactifs, stockage, déchets, fusion, fission, uranium, radioactivité,* etc.
Les noms d'instances écologistes : Les *Verts* (*EELV*), *Greenpeace*.
Les acteurs économiques du nucléaire : *Daiichi* (Japon).
Les termes polysémiques : *Sortir* du *nucléaire* (actions ou Association indépendante), *Hollande* (le Président ou une région et une ancienne province des Pays-Bas), *gendarme* (*AIEA* ou la Gendarmerie).
Des opérateurs nucléaires : *EDF, Areva*, *Tepco* (*Tokyo Electric Power*)

[159] Le logiciel Trameur a été conçu par Serge Fleury à l'Université Sorbonne Nouvelle - Paris 3, le manuel est disponible sur le site dédié : http://www.tal.univ-paris3.fr/trameur/.
[160] La linéarité de calcul des poly-cooccurrences s'explique par des propriétés systématiques associées à la dépendance linéaire en mathématiques, c'est-à-dire, le captage des poly-cooccurrences dans les textes avance de manière linéaire.

Des organismes de surveillance et de recherche du nucléaire : *autorité, ASN, IRSN, CERN, CEA*
Des hommes ou femmes politiques : *Nicolas, Sarkozy, Hollande, Besson, Angela, Merkel*

5.4.4 Poly-cooccurrences autour d'ENRG_US

Les figures se rapportant à ces constats sont disponibles dans l'annexe I (Dictionnaire d'événements et restitutions par poly-cooccurrence des formes-pôles *nuclear* et *energy* pour le sous-corpus ENRG_US pour la période du 26 janvier 2005 au 18 avril 2012).

Contrairement à ce qu'il a été restitué en français, les constats des résultats des calculs des deux formes chez les Américains sont différents. D'une part, peu de formes communes sont repérées dans les résultats des deux formes *energy* et *nuclear* en anglais, mis à part, *electricity, sources* etc. D'autre part le nombre total de formes poly-cooccurrentes est moins important et moins varié que celui en français. Le nombre de poly-cooccurrences trouvé par la forme *nuclear* est moindre que celui d'*energy*. Cela indique étroitement que les informations sur le nucléaire sont plus « sensibles » ou tout au moins, moins accessibles dans le NYT. Cependant, des renseignements sur les lieux des sites de centrales, de stockage ainsi que sur les autres types de sources d'énergies ont été mis en exergue par les poly-cooccurrences de la forme *nuclear*.

5.4.5 Informations révélées par la forme-pôle *nuclear*

Dans les résultats illustrés des poly-cooccurrences, deux types d'informations sont à noter :

- informations fragmentaires,
- suites de mots fédérés indiquant une unité d'informations.

5.4.5.1 Informations fragmentaires

De petites quantités d'informations ont été retrouvées par la forme-pôle *nuclear,* seuls des toponymes relatifs aux lieux et noms de centrales ont été repérés, comme par exemple, *Yucca Mountain* (nom de centrale) dans le *Nevada*, *Three Mile Island* (en Pennsylvanie)*, Indian Point à Buchanan* (ville), *Oyster Creek* (la plus vieille centrale des USA), *Yankee* dans le *Vermont*.

Toutefois, les mots concernant d'autres sources d'énergies que le nucléaire restent poly-cooccurrents avec cette forme-pôle, tels que, *coal* (charbon), *wind* (vent), *clean technologies* (technologies propres), *plants* (centrales), *gas* (gaz), *alternative sources* (sources alternatives[161]). Un nom d'un laboratoire nucléaire *Flats Rocky* (laboratoire de production d'armes nucléaires) est sorti de la masse de données. Les explications relatives à d'autres formes sont disponibles dans l'annexe I.

Des mots à forte connotation nucléaire ont été également recensés : *energy*, *electricity*, *atomic*, *uranium*, *plutonium*, *power*, *reactor*, etc.

5.4.5.2 Suites de mots fédérées indiquant une unité d'informations

Accident > Mile > 1979 > Three

Le nom d'un accident nucléaire partage les mêmes formes *Three Mile* que celles du nom de la centrale où s'est produit l'*accident*, accident qui a eu lieu en *1979*.
Radioactive > weapons > tests > 1951
La forme *1951* marque le début d'une série de cinq essais atomiques atmosphériques sur le site d'essais du Nevada.

[161] Les énergies alternatives sont des énergies susceptibles de remplacer le pétrole en matière de carburant.

5.4.6 Informations révélées par la forme *energy*

5.4.6.1 Informations fragmentaires

Une bonne quantité d'entités nommées, toponymes, noms propres et instances se rapportent à la vie politique américaine tels que *President*, *Barak*, *Obama*, *Vice*, *Dick*, *Cheney*, *Bush*, *signed*, *Bill*, *White House*, *Senate*, *Washington*, *Congress*, etc., à la législation (*legislation*) américaine sur le climat, aux énergies fossiles (*coal*, *gas*) et renouvelables (*renewables*, *solar*, *clean*, *cleaner* etc.).

5.4.6.2 Suites de mots fédérées indiquant une unité d'informations

Energy > *Electricity* > *solar* > *sources* > *renewable*

Cette suite peut s'interpréter de la manière suivante : l'énergie solaire serait l'une des sources à promouvoir comme énergie renouvelable pour la production d'électricité. Un retour au contexte par le module de la carte des sections a été effectué afin d'attester cette anticipation du développement de l'énergie photovoltaïque aux USA.

Guide de lecture des cartes des sections

La carte des sections est une représentation du corpus sous forme de petites cases[162], sectionné à l'aide d'un délimiteur par phrase, paragraphe, article, mois, année ou autre découpage défini par l'utilisateur. Ainsi, la répartition des formes ou groupes de formes (Lamalle et Salem, 2002 : 404) choisies pourra se projeter sur une carte en fonction de leur(s) nombre(s) d'occurrences calculé(s). Un carré correspond à un article dans notre cas, mais peut être également une phrase ou un paragraphe ou un mois ou une année, etc. Lexico 3 permet un affichage par seuil rendant plus lisible la densité des formes choisies en 3 niveaux de couleur, plus la forme est répétée dans le carré, plus la couleur est intense, figure 5.10, ci-dessous :

1. faible intensité : présence d'une occurrence d'une forme dans un carré,
2. moyenne intensité : présence de 2 à 4 occurrences dans un carré,
3. forte intensité : présence de 5 occurrences ou plus dans un carré.

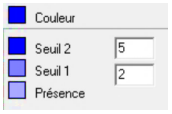


Figure 5.10 Intensité de couleur de la carte des sections de Lexico 3

Guide de lecture de la carte des sections : nombre d'occurrences par intensité de couleur.

Cette carte des sections ainsi définie expose ci-dessous la répartition des formes, *Energy* > *Electricity* > *solar* > *sources* > *renewable* dans le sous-corpus. Cela signifie que toutes ces formes doivent être présentes ensemble dans un texte. Les articles du sous-corpus sont délimités par des signes spéciaux tels que « § » pour la fin d'un article. Ces signes permettent de cerner un article comme étant une unité textuelle. L'un des articles sélectionnés en bleu foncé, pointé par une flèche rouge, atteste la forte présence des formes choisies.

[162] Chaque petite case de la carte des sections correspond à une unité textuelle créée et choisie par l'utilisateur telle qu'une phrase, un paragraphe, un article, etc.

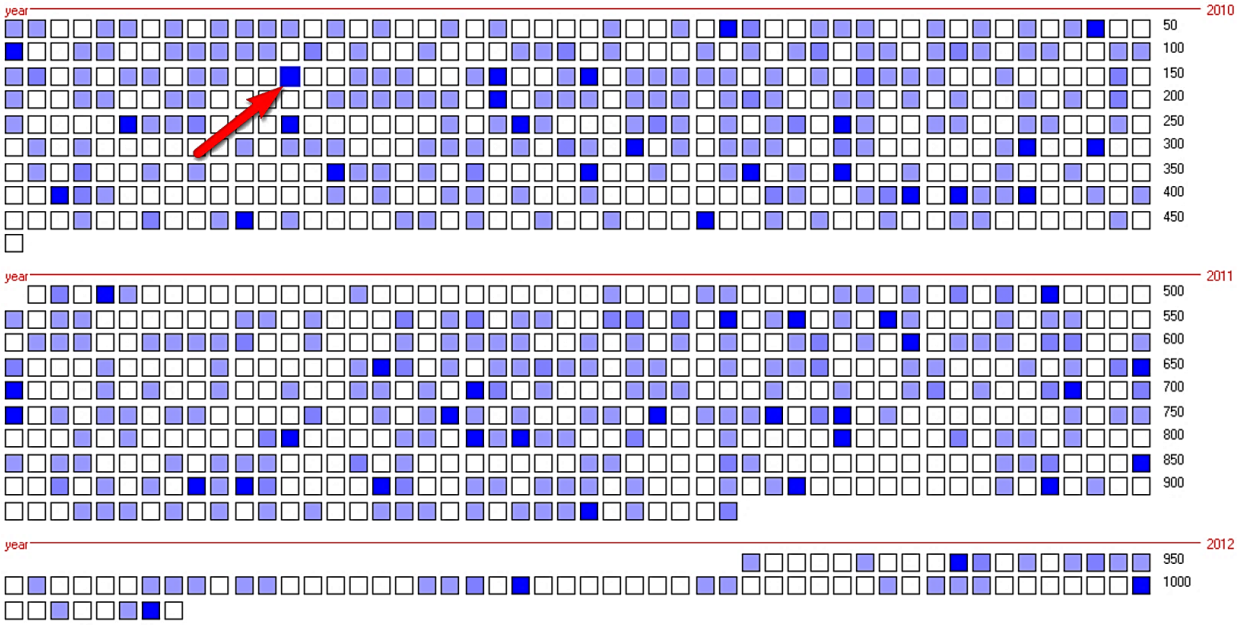


Figure 5.11 ENRG_US de 2010 à 2012 : carte des sections pour le groupe de formes *energy*, *electricity*, *solar*, *sources*, *renewable*, un carré bleu = un article

<day=20100415><article=3126>title : Solar Growth Slows , With Homes a Glaring Exception # A new report from a solar industry group found that the pace of solar installations slowed last year amid the economic downturn . # Total capacity installed for all types of solar energy grew by 5 . 2 percent in 2009 , compared with 9 . 6 percent the previous year . But Rhone Resch , the chief executive of the Solar energy Industries Association , which released the report Thursday , said that the overall number hid tremendous variation within the industry . # For example , he said , the residential market for photovoltaic panels (the type used on rooftops) grew at its fastest pace ever in 2009 , and utilities ' demand for these panels also stayed strong . On the other hand , the large commercial market - companies putting solar panels on their rooftops - lagged . As a result , overall growth in capacity for photovoltaic panels stood at 38 percent last year , down significantly from 84 percent growth a year earlier . # Demand for solar pool heating equipment fell last year , with new installations down by 10 percent compared with 2008 . " It ' s pretty clear that the solar pool heating was hit hardest by the recession , " Mr . Resch said , noting that unlike other types of solar energy , pool heating does not benefit from federal tax incentives . # On a worldwide scale , the United States ranked fourth in solar - electric installations last year , the report said , after Germany , Italy and Japan . In overall capacity , the United States is also fourth , behind Germany , Spain and Japan . Solar power accounts for less than 1 percent of the electricity supply in the United States . # The solar industry has undergone significant changes in the past few years . The price of photovoltaic panels has fallen by over 40 percent since mid - 2008 , due to a combination of factors such as reduced demand in Spain and increased supply of the polysilicon material used to make the panels . However , homeowners have not reaped the full rewards of that price drop : the cost of panels and installation together fell by only 10 percent , the report said , reflecting that labor still accounts for a hefty part of the overall bill . # The industry has also benefited from substantial new federal incentives , aimed at helping homeowners , businesses and utilities afford solar (which is more expensive than most other sources of electric generation) . # Beginning in January 2009 , the federal government lifted a $ 2 , 000 cap on the 30 percent tax credit homeowners receive for solar - electric installations . Larger solar projects also received assistance from a provision in last year ' s stimulus bill that turned a tax credit into direct grants . As of February this year , the industry had gotten Treasury grants worth $ 81 million , the report said . That grant program is scheduled to end Dec . 31 . # States have also helped . California is by far the leading state for solar - electric installations , followed by New Jersey . That is because of the state ' s aggressive solar - rebate program and other environmental requirements . According to the report , this year California will also begin " the most ambitious " program of any state to encourage installation of solar water heaters - with the aim of adding 200 , 000 such systems across the state . # (Mr . Resch also noted that Florida " really leapfrogged " in the state rankings last year ; the Sunshine State came in third in solar - electric capacity additions .) # The industry is still hoping that Congress will approve further policies to aid solar . On its wish list is a national renewable electricity requirement , with special requirements for solar power (18 states already have a solar electricity requirement) . # Mr . Resch said he was optimistic that Congress would pass an energy bill this year , but he noted that the requirements must be structured correctly to make a difference . The current proposal in Congress , he said , " needs a lot of work . " §

Figure 5.12 ENRG_US : article retenu contenant le groupe de formes *energy*, *electricity*, *solar*, *sources*, *renewable* publié le 15 avril 2010

Retranscription du titre et du passage sélectionnés (figure 5.12)

<day=20100415><article=3126> title: Solar Growth Slows, With Homes a Glaring Exception (……)
States have also helped. California is by far the leading state for solar - electric installations, followed by New Jersey. That is because of the state's aggressive solar - rebate program and other environmental requirements. According to the report, this year California will also begin "the most ambitious" program of any state to encourage installation of solar water heaters - with the aim of adding 200, 000 such systems across the state. (Mr. Resch also noted that Florida "really leapfrogged" in the state rankings last year; the Sunshine State came in third in solar - electric capacity additions.)
The industry is still hoping that Congress will approve further policies to aid solar. On its wish list is a national renewable electricity requirement, with special requirements for solar power (18 states already have a solar electricity requirement).

Traduction française

<day=20100415><article=3126> Titre: La croissance solaire ralentit malgré une exception flagrante avec les maisons à panneaux solaires

(……)Les Etats ont également contribué (à cette croissance). La Californie est de loin le premier Etat au plan des installations électriques solaires, suivie par le New Jersey. Cela fait suite à la politique agressive de l'Etat en matière d'énergie solaire, programmes de remboursements et autres exigences environnementales. Selon le rapport (plus haut dans l'article), cette année en Californie va commencer le programme le plus ambitieux de tous les Etats-Unis pour encourager l'installation de chauffe-eau solaire, avec l'objectif d'ajouter 200 000 de ces systèmes à travers l'Etat. (M. Resch a également noté que la Floride a vraiment bondi dans les classements des Etats l'année dernière ; la Floride, l'Etat du soleil (Sunshine State) est arrivé troisième du palmarès par son ajout de capacité électrique en énergie solaire).

L'industrie (solaire) espère toujours que le Congrès approuvera de nouvelles politiques pour aider le secteur de l'énergie solaire. Une exigence nationale concernant la production d'électricité de nature renouvelable est sur la liste de souhaits du Congrès américain, en particulier, pour l'énergie solaire (18 États ont déjà cette exigence d'électricité solaire).

La carte des sections (figure 5.11) ci-dessus, illustre la répartition des mots du groupe de formes retenues dans les trois années d'ENRG_US. L'accès à l'article en anglais montre que l'énergie solaire progresse lentement mais sûrement aux États-Unis, malgré la complexité de son contexte politique.

5.4.7 Comparaisons analytiques des poly-cooccurrences français-anglais

Après l'analyse des poly-cooccurrences tirées des formes *énergie*, *nucléaire* en français, *energy* et *nuclear* en anglais, des comparaisons transversales seront nécessaires pour la restitution, la prévision et l'anticipation d'informations des constats bilingues.

5.4.7.1 Points communs

A l'exception de quelques formes dont *électricité* (*electricity*), *charbon* (*coal*), hommes et femmes politiques, instances politiques, etc., peu de points communs comparables ont été relevés dans les constats. En revanche les formes liées aux sources d'énergies renouvelables et alternatives sont communément présentes dans les deux sous-corpus.

5.4.7.2 Points divergents

ENRG_FR rapporte des événements tant nationaux qu'internationaux. Par contre ENRG_US relate davantage les événements nationaux. Par exemple, nous retrouvons les mots correspondant à des accidents nucléaires majeurs, celui de *Three Mile Island* survenu aux Etats-Unis ou celui de *Fukushima*, mais le mot *Tchernobyl* est absent des poly-cooccurrences américaines, les USA n'ayant pas été touchés par les conséquences de Tchernobyl.

Des formes liées aux mouvements écologistes nationaux et internationaux, (*EELV*, *Sortir* du *nucléaire*, *Greenpeace*, etc.), aux instances de recherche, de surveillance du nucléaire, d'opérateurs (*CERN, ASN*, *AIEA, EDF, AREVA*, etc.) sont présentes uniquement dans ENRG_FR.

Par ailleurs, les implantations de centrales nucléaires sont rarement mentionnées dans ENRG_US, seuls quelques sites sont détectés, par exemple des centrales nucléaires telles que *Yucca Mountain*, *Indian Point*, *Oyster Creek*, *Yankee Vermont* et un laboratoire de production *Flats Rocky*, sites où les incidents nucléaires sont relativement récurrents.

5.4.8 Point d'entrée pour la veille bilingue français-anglais

L'absence des mots *incident(s)* des poly-cooccurrences américaines nous a surpris, au vu de l'ancienneté des centrales et de leurs réacteurs, néanmoins, la forme *reactor* (réacteur) a été repérée plusieurs fois. Une recherche plus approfondie a été nécessaire afin de comprendre le mécanisme linguistique de l'anglais dans la narration des faits liés aux incidents ou accidents. L'association des quatre formes, *nuclear*, *Vermont, Yankee* et *reactor* constitue une unité d'information plausible, car le champ sémantique formé par les quatre formes crée un espace physique. L'objectif est d'en déduire s'il y a une action ou tout au moins une forme quelconque d'action qui s'est produite dans l'espace-temps de notre événement. Par déduction, la forme *leak* (fuite) serait une piste de recherche. En effet, un incident nucléaire est très souvent lié au fait qu'il y ait une fuite radioactive. Ainsi, cette forme a été projetée sur la carte des sections d'ENRG_US (figure 5.13, ci-dessous).

Sa projection sur la carte des sections avec le délimiteur § pour les articles a permis le retour des 41 articles concernés, contenant la forme *leak* (fuite) en bleu. Afin d'affiner la recherche, la forme qui incarne le champ sémantique de notre objet, *nuclear* (nucléaire) en rouge a été, par la suite, projetée sur cette même carte. Une coprésence des deux formes est à noter, seulement pour l'année 2010 avec 7 articles concernés sur les 41 présents dans le sous-corpus.

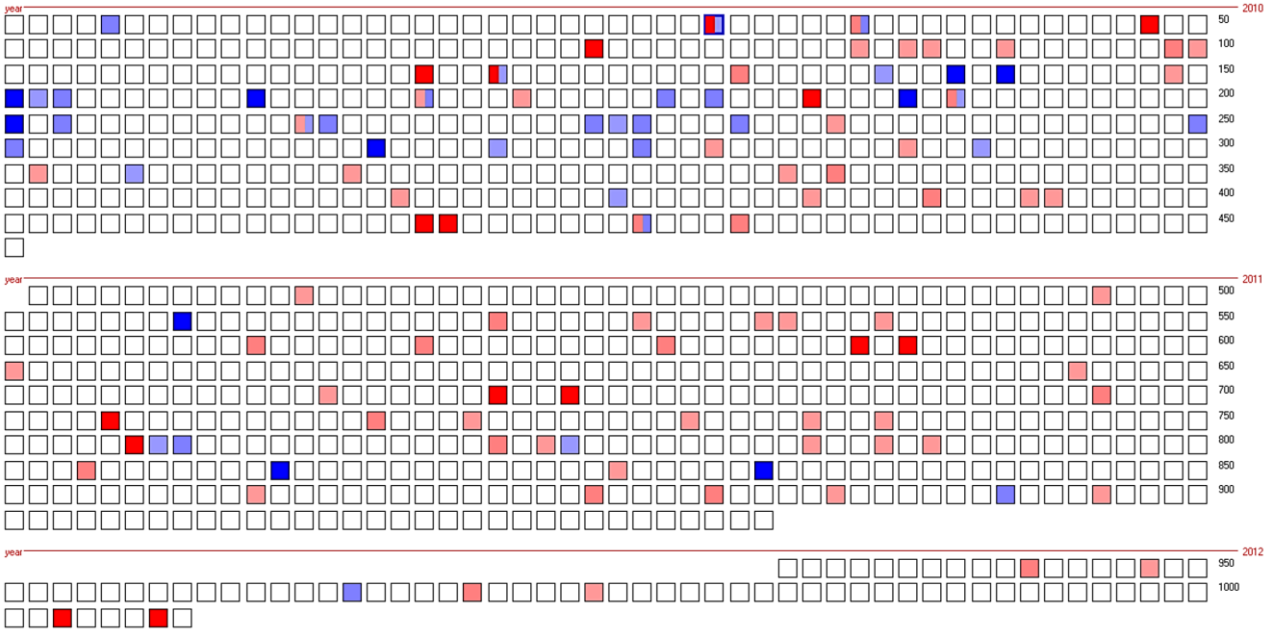


Figure 5.13 ENRG_US de 2010 à 2012 : carte des sections pour les formes *leak* (fuite) en bleu et *nuclear* (nucléaire) en rouge, un carré = un article

La consultation de chacun de ces 7 articles contenant la forme *leak* (fuite) et *nuclear* (nucléaire) révèle que 2 articles relatent des incidents nucléaires attachés à la centrale nucléaire Vermont Yankee, 1 évoque l'incident de cette centrale, et les 4 autres des rappels d'événements nucléaires à l'occasion de fuites de pétrole en particulier l'accident de BP dans le golfe du Mexique.

Extraits des 2 articles relatant des incidents nucléaires à la centrale nucléaire Vermont Yankee :

> *<article=3043> title : Radiation Levels Cloud Vermont Reactor's Fate # Levels of radioactive tritium have risen rapidly in recent weeks in the groundwater surrounding Vermont's sole nuclear power plant , leading both longtime supporters and foes of the reactor to question whether it will be allowed to keep operating . # Marchers gathered in Montpelier this month to push for the closing of the Vermont Yankee nuclear plant near Brattleboro. (...)*
>
> *<day=20100202><article=3049> title : Vermont Power Plant Continues to Leak Radiation # Technicians seeking the source of a leak of radioactive tritium at the Vermont Yankee nuclear plant have found concentrations in groundwater there that were three times higher than what was discovered last week , a plant spokesman said Monday . (...)*

Traduction française

> <article = 3043> Titre: Les niveaux de rayonnement embrument le destin des réacteurs du Vermont # Les niveaux de tritium radioactif ont augmenté rapidement au cours des dernières semaines dans les eaux souterraines environnantes de l'unique centrale nucléaire du Vermont, conduisant les partisans et adversaires du réacteur de longue date à la question de savoir s'il sera autorisé à continuer de fonctionner. # Les manifestants se sont réunis à Montpelier ce mois-ci pour faire pression sur la fermeture de la centrale nucléaire Vermont Yankee, près de Brattleboro. (...)
>
> <= 20100202 jour> <article = 3049> Titre: La centrale Vermont fuit toujours # Les techniciens cherchent la source de la fuite de tritium radioactif dans la centrale nucléaire Vermont Yankee. Il s'avère que la radioactivité des eaux souterraines est trois fois plus élevée que celle découverte la semaine dernière, a déclaré un porte-parole de l'usine lundi. (...)

Champ sémantique pour la veille informationnelle

L'expérience de cette restitution démontre que les formes verbales (au sens linguistique du verbe) et le gérondif en anglais désignent généralement des événements ponctuels, accidentels, circonstanciels ou encore exprimant un but ou un constat, comme par exemple fuite (*leak*) dans notre cas. Ces mots sont souvent sous forme de noms, désignant un accident ou une forme d'action accidentelle dans un espace-temps. Ces formes s'avèrent très révélatrices pour la veille informationnelle et créent des points d'entrée extrêmement efficaces. Le champ sémantique à caractère actionnel généré par ces genres de formes, tels que fuite (*leak* ou *leaking*), protection, frappe (*hit ou hitting*), renversement ou se répandre ou nappe (*spill, oil spill* pour marée noire), effusion ou écoulement (*spilling*), etc. peut constituer un outil de prospection pour l'exploration des informations. Lorsqu'on ajoute un groupe de formes d'entités nommées à ce champ sémantique, ce couple va permettre par la suite d'accéder à des informations monolingues, bilingues puis multilingues dans les corpus. Dans notre cas, nous pouvons reconstituer un fragment d'information à partir du couple de formes *Vermont* et *leaking*. Ensuite, des explorations orientées par ce couple pourront apporter davantage d'éléments à nos recherches.

5.4.9 Cooccurrences évolutives autour des formes *énergie(s)* sur le sous-corpus français

Les calculs de cooccurrences évolutives ont été effectués à partir des formes *énergie* et *énergies* sur ENRG_FR afin d'obtenir un réseau de mots cooccurrents évolutifs (figures disponibles dans l'annexe N). Des restitutions d'informations autour des formes *énergie(s)* sont réalisées, cependant des retours aux textes des articles sont appliqués uniquement sur la forme *EPR*. Rappelons que les calculs de cooccurrences évolutives consistent à procéder aux mesures des cooccurrences période par période, afin de comparer chronologiquement l'évolution des formes employées dans chaque période. Dans notre cas, il s'agit de scinder le sous-corpus par année, puis de calculer les cooccurrences année par année.

Tableau 5.11 ENRG_FR de 2010 à 2012 : synthèse des résultats des cooccurrences évolutives pour les formes *énergie* et *énergies*

énergie	2010	2011	2012	énergies	2010	2011	2012
Ademe	x			2020		x	
Agence	x	x	x	alternatives		x	
AIE	x	x		biomasse		x	
AIEA		x		Chine	x		
atomique	x	x	x	développement		x	
Besson		x		développer	x		
CEA		x		électricité		x	x
charbon		x		énergétique		x	
Commissariat		x		énergie	x	x	
consommation	x	x		éolien	x	x	
consommé	x			fossiles	x	x	x
économies	x	x		hydraulique	x		
électricité	x	x		investissements		x	
électrique	x			nucléaire		x	
énergétique		x		part		x	
énergies	x	x		photovoltaïque		x	
environnement	x			renouvelables	x	x	x
éolien	x			solaire	x		
éolienne	x	x		Syndicat[163]		x	
Eric		x					
gourmands	x						
internationale		x	x				
kWh	x						
maitrise	x						
nucléaire	x	x	x				
production	x	x					
projet	x						
renouvelable	x	x	x				
renouvelables		x					
solaire	x	x	x				
source		x					
sources		x					

Les formes communes obtenues par les calculs de cooccurrences entre *énergie* et *énergies* ont été colorées en rouge dans le tableau 5.11. Il convient de noter que l'année 2012 n'étant pas complète, sa colonne ne peut être interprétée à égalité avec les années 2010 et 2011.

[163] Créé en 1993, le Syndicat des Énergies Renouvelables regroupe, directement ou indirectement, plusieurs milliers d'entreprises, concepteurs, industriels et installateurs, associations professionnelles spécialisées, représentant les différentes filières. Parmi ses adhérents figurent les plus grands énergéticiens mondiaux ou nationaux comme des groupes ou acteurs locaux des énergies renouvelables, http://www.enr.fr/histoire-et-missions (consulté le 20/08/2015).

Synthèse d'information : *énergie* versus *énergies* dans ENRG_FR

Par analyse transversale, un constat immédiat des réflexions sur la synthèse des résultats en français (tableau 5.11) démontre que le champ sémantique de la forme *énergie* présente une très forte connotation en relation avec l'énergie atomique, c'est-à-dire le nucléaire (*atomique*), ainsi que toutes les entités nommées côtoyant les organismes et instances en relation avec le nucléaire (*Ademe, Agence, AIE, AIEA, CEA, Commissariat*), les hommes politiques (*Besson, Eric*), les sources (*charbon, éolien, éolienne, renouvelable, renouvelables, solaire, source, sources*), les relations économiques, techniques et d'exploitation (*consommation, consommé, économies, électricité, électrique, énergétique, énergies, environnement, gourmands, internationale, kWh, maîtrise, production, projet*).

La forme *énergie* est cooccurrente avec *Agence, atomique, nucléaire, renouvelable et solaire* dans les trois années du sous-corpus français.

Tableau 5.12 : ENRG_FR : extrait de l'inventaire distributionnel trié après la forme *énergie*

```
    2 ---- ---- ---- ---- énergie Eric Besson
 62 ---- ---- ---- ---- ---- énergie atomique
          2 ---- énergie atomique et aux énergies alternatives
 11 ---- ---- ---- ---- ---- énergie consommée
 14 ---- ---- ---- ---- ---- énergie électrique
 33 ---- ---- ---- ---- ---- énergie éolienne
      3 ---- ---- ---- énergie éolienne et solaire
  2 ---- ---- ---- ---- ---- énergie géothermique
  2 ---- ---- ---- ---- ---- énergie hydraulique
  3 ---- ---- ---- ---- ---- énergie hydroélectrique
 94 ---- ---- ---- ---- ---- énergie nucléaire
  4 ---- ---- ---- ---- ---- énergie photovoltaïque
 26 ---- ---- ---- ---- ---- énergie renouvelable
 51 ---- ---- ---- ---- ---- énergie solaire
```

L'inventaire distributionnel, tableau 5.12, réalisé avec les contextes après la forme *énergie* permet de hiérarchiser les entités adjacentes de cette même forme et de visualiser leur nombre de répétitions. L'extrait montre que la part du nucléaire (94 fois et 62 fois pour *atomique* dont 2 fois avec *énergies alternatives*) domine incontestablement les autres formes d'énergies. Il faut noter que la séquence *énergies alternatives* est aussi un composant du nouveau nom du CEA (Commissariat à l'énergie atomique et aux énergies alternatives). Les autres entités concernent principalement les différents types d'énergies et leurs transformations telles que solaire, éolienne, hydroélectrique, géothermique etc.

Tandis que sa forme plurielle, *énergies,* selon le tableau 5.13, évoque davantage les éventuelles substitutions de l'énergie atomique par les nouveaux types de sources alternatives (*alternatives, biomasse, énergie, énergétique, éolien, fossiles, hydraulique, photovoltaïque, nucléaire, solaire, renouvelables*), les objectifs et participations nationales et internationales (*2020, Chine, Syndicat*).

Les formes *fossiles* et *renouvelables* restent cooccurrentes triennales avec *énergies*.

Par analyse chronologique, l'année 2011 a été très marquée par l'événement de Fukushima, ceci a déclenché toute une série de remises en questions de l'emploi du nucléaire, et explique l'apparition des formes d'organismes de contrôle et de sûreté nucléaire, et des hommes politiques prenant position pour la poursuite du nucléaire[164] (*Agence, AIE, AIEA, atomique, Besson, CEA, Commissariat*) par le

[164] Se reporter au document dans le dossier Doc_Annexes, Doc_Annexe-D-nucléaire-France : interview de Monsieur Éric Besson et commentaires d'experts, document disponible sur : https://drive.google.com/drive/folders/0B8XHfHwNzWAAclBOOExJTUw5aTA

fait qu'ils envisagent de mettre en place de nouvelles mesures et normes de sécurité dans les centrales. Dans ce contexte, le *charbon* n'a plus jamais été considéré comme la principale alternative du nucléaire. Quant aux *investissements*, forme apparue en 2011, sur la recherche de nouvelles énergies (*biomasse, part, photovoltaïque, renouvelables*, *Syndicat*), formes apparues également en 2011 et sur les renforcements pour améliorer la sécurité et la sûreté du nucléaire, ceux-ci ne font que consolider la peur du nucléaire dans le monde.

Tableau 5.13 : ENRG_FR : extrait de l'inventaire distributionnel trié après la forme *énergies*

```
  3 ---- ---- ---- ---- ---- énergies 2050
 12 ---- ---- ---- ---- ---- énergies alternatives
    2 ---- ---- ---- ---- énergies alternatives renouvelables
  2 ---- ---- ---- ---- ---- énergies décarbonées
    2 ---- ---- ---- ---- énergies en 2020
    2 ---- ---- ---- ---- énergies faiblement carbonées
 60 ---- ---- ---- ---- ---- énergies fossiles
  3 ---- ---- ---- ---- ---- énergies marines
 11 ---- ---- ---- ---- ---- énergies nouvelles
  8 ---- ---- ---- ---- ---- énergies propres
 278 ---- ---- ---- ---- ---- énergies renouvelables
       4 ---- ---- ---- énergies renouvelables en 2020
           2 ---- énergies renouvelables sont effectivement intermittentes ou
variables
       2 ---- ---- ---- énergies solaire et éolienne
  4 ---- ---- ---- ---- ---- énergies vertes
```

L'inventaire distributionnel, tableau 5.13, réalisé et trié après la forme *énergies* permet de hiérarchiser les entités adjacentes de cette même forme et de visualiser leur nombre de répétitions. L'extrait montre la large place réservée généralement aux différents types d'énergie, pour la plupart leurs termes génériques, tels que renouvelables (278 fois), fossiles (60 fois), alternatives (12 fois), nouvelles (11 fois), propres (8 fois), vertes (4 fois), marines (3 fois), solaire et éolienne (2 fois), décarbonées (2 fois), alternatives renouvelables (2 fois), faiblement carbonées (2 fois). En ce qui concerne 2020 et 2050, ce sont des objectifs énergétiques à atteindre dans le but de limiter les GES [165] et de créer des nouvelles énergies, concertés lors de conférences par des instances nationales, européennes et internationales.

Dans ce contexte complexe, le réacteur pressurisé européen, appelé EPR, l'un des symboles de la troisième génération de réacteur nucléaire, apparaissait comme un nouveau souffle. Le tableau 5.14 ci-dessous retrace les grands jalons relatifs à la construction des EPR dans le monde. Aujourd'hui que pouvons-nous en restituer, anticiper et prévoir ?

Quatre EPR sont en cours de construction, mais aucun n'est encore opérationnel à ce jour ; un premier à Olkiluoto en Finlande, un deuxième à Flamanville en France, et les deux derniers à Taishan en Chine. Un cinquième est prévu à Penly en France, et un sixième à *Hinkley Point*[166] (Royaume-Uni), mais pour le moment, tous les travaux sur ces deux projets sont suspendus.

[165] GES : gaz à effet de serre

[166] Après dix ans de préparation, le projet de centrale nucléaire EPR à Hinkley Point a été signé. http://www.lemonde.fr/economie/article/2016/09/29/le-gouvernement-britannique-signe-l-accord-nucleaire-de-hinkley-point-avec-edf_5005674_3234.html (consulté le 01/10/2016)

Tableau 5.14 Quelques jalons chronologiques concernant les réacteurs EPR dans le monde

EPR Olkiluoto (Finlande)[167]	
2003-12	Signature du contrat entre Areva et TVO pour la construction du 1er EPR à Olkiluoto.
2005	Début du chantier.
2009	Initialement date d'entrée en service du réacteur d'Olkiluoto 3.
2013 (6-12)	Suite à de nombreux retards, notamment au décalage du système destiné au pilotage et au contrôle du réacteur, une nouvelle date de mise en service est annoncée.
2013 - 2015	Areva est en litige avec son client TVO pour partager l'ensemble de cet énorme surcoût. Une décision arbitrale est attendue début 2015 par la Chambre de commerce internationale.
2016-6	Fin de construction de la centrale et début des essais.
2018	Mise en service de l'EPR.
EPR Flamanville[168] **(France)**	
2006-5	EDF décide de lancer le projet d'EPR à Flamanville et dépose une demande d'autorisation de création auprès des pouvoirs publics.
2007-12	Début des travaux de l'EPR Flamanville.
2012-12-3	Le coût de construction[169] du réacteur pressurisé européen (EPR) de Flamanville est dépassé : EDF annonce « *avoir relevé de 2 milliards d'euros son estimation du coût, portée à 8,5 milliards, inflation comprise* ». Le devis initial, en 2005 prévoyait un montant de 3,3 milliards d'euros.
2016	Objectif de première production commercialisable pour l'EPR Flamanville (4 ans de retard dus à des études d'ingénierie supplémentaires, à l'intégration de nouvelles exigences réglementaires et de sécurité suite à la catastrophe nucléaire de Fukushima), (prévisions de 2012).
2017-12	Date prévisionnelle de mise en service de l'EPR Flamanville.
EPR Taishan 1 et 2 (Chine)[170]	
2006-10	Partenariat industriel signé entre EDF et l'électricien chinois China Guangdong Nuclear Power Holding Company (CGNPC).
2008-8-10	Concrétisation à Pékin, des accords finaux fixant la construction et l'exploitation de deux centrales nucléaires de technologie EPR à Taishan, dans la province du Guangdong, sur le modèle du réacteur EPR en cours de construction à Flamanville. Le groupe français est partenaire à 30% de ce projet aux côtés de la compagnie chinoise China General Nuclear (CGN).
2009	Début des travaux.
2011-10	Pose du dôme du bâtiment réacteur de la première tranche réalisée avec succès, puis pose du dôme du bâtiment réacteur de la deuxième tranche intervenue le 12 septembre 2012.
2011 - 2015	Ces 2 EPR en cours de construction par Areva dans le sud-est de la Chine, Taishan 1 et 2, ne connaissent pas les mêmes problèmes que les 2 autres EPR, les retards ne sont pas significatifs, et ils bénéficient des retours d'expérience des deux chantiers finlandais et français.
2015-1[171]	EDF annonce la livraison, pour la fin de l'année 2015 de Taishan 1 et quelques mois plus tard de Taishan 2.
2016-6	La livraison de Taishan est repoussée en attendant les résultats de tests complémentaires demandés par l'ASN[172] sur l'EPR à Flamanville.

[167] http://www.lemonde.fr/economie/article/2011/10/12/l-epr-finlandais-prend-du-retard_1586073_3234.html (consulté le 20/08/2015).
[168] http://energie.edf.com/nucleaire/carte-des-centrales-nucleaires/presentation-48324.html (consulté le 18/10/2014).
[169] « Le coût de l'EPR de Flamanville revu encore à la hausse » http://www.lemonde.fr/planete/article/2012/12/03/le-cout-de-l-epr-de-flamanville-encore-revu-a-la hausse_1799417_3244.html (consulté le 18/10/2014).
[170] http://asie.edf.com/activites/nucleaire-46732.html (consulté le 20/08/2015).
[171] Annonce par Hervé Machenaud directeur de la branche Asie-Pacifique d'EDF.
[172] http://france3-regions.francetvinfo.fr/basse-normandie/manche/nord-cotentin/cherbourg-en-cotentin/edf-suspend-les-controles-techniques-sur-l-epr-de-flamanville-1032255.html (consulté le 25/06/2016)

Projet construction EPR Penly (France)[173]	
2009-1-30	Le Président de la République annonce la création d'un deuxième réacteur électronucléaire de type EPR sur le site de Penly.
2009-7	Ouverture d'un débat public.
2010 - 2015	Statu quo ?
Projet construction EPR Hinkley Point (Royaume-Uni)[174]	
2015-4-2	La filiale anglaise d'EDF a annoncé ce jour la suspension des travaux préparatoires en vue de la construction de la centrale nucléaire de Hinkley Point. EDF Energy accumule les déboires sur Hinkley Point depuis le démarrage de la phase d'étude il y a quelques années. Le budget estimé a été revu plusieurs fois à la hausse, tandis que la date de livraison n'a cessé de glisser pour se fixer, selon les dernières indications, à 2023 au plus tôt.

5.4.10 Veille active et veille ciblée par poly-cooccurrences évolutives de l'EPR

Pourquoi seulement les poly-cooccurrences ?

Nous venons de démontrer l'efficacité de la restitution d'informations par le réseau cooccurrentiel des formes *énergie* et *énergies*. En effet, le calcul des cooccurrences est très efficace pour restituer le contexte d'un événement, toutefois, dans une recherche de veille active et ciblée, nous devons recourir, dans la plupart des cas, aux calculs de poly-cooccurrences afin de récupérer un maximum de fragments informationnels, à partir d'une forme spécifique, dans une ou plusieurs branches du réseau cooccurrentiel, susceptibles d'être utiles pour la restitution d'informations. Il est à noter que tous les calculs des poly-cooccurrences sont obtenus en se basant sur ceux des cooccurrences.

Guide de lecture des figures de listes des poly-cooccurrents

Afin de faciliter la lecture des représentations des poly-cooccurrences, nous avons opté pour la liste des poly-cooccurrents plutôt que les diagrammes (les figures sont disponibles dans l'annexe N), parfois difficiles à interpréter, notamment pour la forme *EPR* et l'année 2011. Les paramètres par défaut, co-fréquence 2, seuil 10, et contexte phrase (. ?!) ont été maintenus dans tous les calculs de poly-cooccurrences.

Dans la figure 5.14 ci-dessous, la lecture se fait de la manière suivante :

- **Pole : EPR (fq : 36)** : la forme-pôle *EPR* est présente 36 fois dans le sous-corpus choisi.

Les paramètres de calculs sont :

- **Co-Freq (Co-fréquence) : 2** : il faut avoir une co-apparition minimum 2 fois, de la forme pôle, dite P1 et une autre dite P2 (qui sera trouvée dans le sous-corpus).
- **Seuil 10** : paramètre pour la spécificité de cette forme-pôle *EPR*, il faut satisfaire le seuil de probabilité à 10 (paramètre fixé par le logiciel Trameur).
- **P1-cofreq (specif)(contextes)->P2** : prenons la 7e ligne de la figure comme exemple. La forme-pôle de départ P1, ici *EPR*, co-fréquence = 8, spécificité = 18.2, contexte = 7, P2 *sûreté*. C'est-à-dire, qu'il y a 7 phrases dans le sous-corpus qui contiennent les deux formes *EPR* et *sûreté* avec une coprésence de 8. Rappelons que dans notre cas le paramètre contexte est défini par les signes de ponctuation comme «.!?». Plus concrètement, un contexte est égal à une phrase.

[173] http://www.asn.fr/L-ASN/ASN-en-region/Division-de-Caen/Centrales-nucleaires/Projet-de-creation-d-un-reacteur-EPR-sur-le-site-de-Penly (consulté le 20/08/2015).
[174] http://www.lefigaro.fr/flash-eco/2015/04/02/97002-20150402FILWWW00212-edf-suspend-les-travaux-a-hinkley-point-c-uk.php (consulté le 20/08/2015).

Pour comprendre le cas où plusieurs formes-pôles se manifestent, prenons la ligne 5 de la figure 5.14. *EPR* = P1, *réacteur* = P2, *Finlande* = P3, P1 est 11 fois co-présent avec P2 dans 10 phrases avec une spécificité de 23,7, mais P2 est 4 fois co-présent avec P3 dans 4 des 10 phrases de P1 et P2 avec une spécificité de 10.3. Les calculs de toutes ces spécificités sont automatiquement attribués par le logiciel, Le Trameur.

Poly-Cooccurrents (Forme (annotation 1)) - Pole : EPR (fq:36)

- *Co-Freq : 2*
- *Seuil : 10*
- *P1-cofreq(specif)(contextes)->P2*

Liste des Poly-cooccurrents

EPR -10(18.5)(10)-> Areva -7(19)(7)-> chantier -3(10.3)(3)-> arbitrage -3(11)(3)-> trancher -3(11.3)(3)-> différend -3(10.3)(3)-> TVO

EPR -10(18.5)(10)-> Areva -4(10.4)(4)-> Finlande -3(10.6)(3)-> Olkiluoto -3(12.7)(1)-> 1600 -3(10.3)(1)-> MW

EPR -7(14.1)(7)-> 2013 -6(11.6)(6)-> service -4(12.3)(4)-> finlandais -4(12.2)(4)-> TVO

EPR -11(23.7)(10)-> réacteur -5(10.8)(3)-> retard -3(10.5)(3)-> démarrage

EPR -11(23.7)(10)-> réacteur -4(10.3)(4)-> Finlande

EPR -8(10.6)(6)-> nucléaire

EPR -8(18.2)(7)-> sûreté

Figure 5.14 ENRG_FR 2010 : réseau poly-cooccurrentiel *EPR*

La liste des poly-cooccurrents en 7 lignes dans la figure 5.14 ci-dessus, nous livre ligne par ligne les attractions lexicales par des suites de mots fédérées indiquant des unités d'informations.

Les formes obtenues des 5 premières lignes restituent un même événement : un retard sur le démarrage et la mise en service en 2013 du chantier EPR à Olkiluoto en Finlande, réacteur nucléaire de 1600 MW et un différend à trancher par arbitrage entre Areva et TVO[175].

Les deux autres lignes traitent du nucléaire et de la sûreté.

Nous restituons les informations par un retour au texte des formes du réseau, informations marquées en rouge.

> *<day=20101129><article=3082> title : EPR : Areva estime respecter le dernier calendrier établi (……) Les retards accumulés sur le chantier ont amené Areva à enregistrer huit provisions successives. Ce réacteur de troisième génération aura coûté au total 5,6 milliards d'euros, alors que le prix d'origine était de 3 milliards d'euros. Areva et TVO se rejettent régulièrement la responsabilité du retard du chantier de l ' EPR et ont engagé une procédure d'arbitrage pour trancher leur différend. L'un et l'autre réclament des milliards d'euros de dédommagements.*
> <day=20100311> <article=2213> *title:La sûreté du réacteur EPR de nouveau en cause | Eco(lo)(……)# Combien d'EPR sont en construction ? # Areva construit actuellement quatre EPR : une unité de 1600 MW à Olkiluoto (Finlande), une de 1600 MW à Flamanville (Manche), depuis 2007, et deux de 1600 MW chacune à Taishan (Chine). Un débat public sur la construction d'un autre EPR sur le site de la centrale nucléaire de Penly, près de Dieppe (Seine-Maritime), doit s'ouvrir le 24 mars (2010). D'autres projets sont à l'étude, notamment aux Etats-Unis et au Royaume-Uni.*

[175] Teollisuuden Voima Oyj (TVO) est une compagnie privée finlandaise de production d'électricité avec participation de plusieurs compagnies finlandaises.

> *<day=20100607><article=2515> title:Le lancement de l'EPR finlandais à nouveau repoussé*
> *# Photo prise le 15 mars 2010 du chantier de l'EPR construit par Areva et Siemens en Finlande. Le démarrage du réacteur EPR finlandais a encore été reporté de six mois lundi 7 juin et accuse désormais près de 4 ans de retard. Ce nouveau report constitue un nouveau coup dur pour son concepteur Areva, qui pourrait voir ses comptes à nouveau plombés par les surcoûts de ce chantier-phare. # "Compte tenu des progrès réalisés actuellement sur le chantier", l'exploitation nucléaire "interviendra fin 2012", a annoncé le groupe nucléaire français dans un communiqué. Entamée en septembre 2005, la construction du 3e réacteur de la centrale finlandaise d'Olkiluoto devait initialement se terminer en avril 2009. Mais la fin des travaux a été reportée à au moins quatre reprises. (……) # Le réacteur de 3e génération (EPR) d'Olkiluoto est le premier réacteur de ce type construit dans le monde, avec celui de Flamanville (Manche), dont l'exploitation commerciale est prévue en 2013. EDF avait reconnu implicitement fin 2009 que ce chantier avait aussi pris du retard. Deux EPR doivent également être mis en service à Taishan, en Chine, en 2013 et 2014.*

Le retour aux textes du sous-corpus, citations ci-dessus, atteste formellement notre déduction. Jusqu'à nos jours, les différents retards accumulés sur le chantier EPR en Finlande repousseraient sa mise en service aux environs de 2018, soit 9 ans de retard, avec un doublement de son coût initial, soit de 3,7 milliards à 7,4 milliards d'euros[176].

Poly-Cooccurrents (Forme (annotation 1)) - Pole : EPR (fq:150)

- *Co-Freq : 2*
- *Seuil : 10*
- *P1-cofreq(specif)(contextes)->P2*

Liste des Poly-cooccurrents

EPR -47(18.3)(42)-> nucléaire -11(16.5)(11)-> troisième -12(17.8)(12)-> réacteur -11(24.5)(15)-> génération -7(11.6)(7)-> construction -9(20.6)(14)-> Flamanville
EPR -47(18.3)(42)-> nucléaire -11(16.5)(11)-> troisième -12(17.8)(12)-> réacteur -11(24.5)(15)-> génération -5(10.9)(5)-> retard -5(11.4)(5)-> Flamanville
EPR -47(18.3)(42)-> nucléaire -9(13.1)(9)-> chantier -8(11.8)(10)-> réacteur -6(10.8)(13)-> construction -5(11.6)(14)-> Flamanville
EPR -47(18.3)(42)-> nucléaire -9(13.1)(9)-> chantier -8(11.8)(10)-> réacteur -5(11)(15)-> génération -5(11.5)(10)-> Flamanville
EPR -29(41.8)(28)-> chantier -8(10.7)(7)-> construction -8(18)(14)-> Flamanville -5(11.1)(14)-> génération
EPR -20(25.3)(20)-> troisième -5(10.3)(5)-> Manche -5(11.2)(11)-> Flamanville -5(11.6)(14)-> génération
EPR -21(17.5)(21)-> EDF -12(14.8)(12)-> réacteur -8(13.3)(13)-> construction -5(11)(14)-> Flamanville
EPR -21(17.5)(21)-> EDF -12(14.8)(12)-> réacteur -5(10.1)(15)-> génération -6(13.9)(10)-> Flamanville
EPR -47(18.3)(42)-> nucléaire -9(13.1)(9)-> chantier -8(11.8)(10)-> réacteur -5(12.3)(5)-> Manche
EPR -12(10.4)(12)-> service -5(11.3)(5)-> Manche -5(13.8)(5)-> 2016 -6(14)(5)-> Flamanville
EPR -29(41.8)(28)-> chantier -8(16.4)(8)-> Manche -8(17.5)(11)-> Flamanville
EPR -21(17.5)(21)-> EDF -5(10.8)(5)-> 2016 -4(10.4)(4)-> Manche -5(11.9)(11)-> Flamanville
EPR -12(10.4)(12)-> service -7(15.5)(7)-> retard -7(11.3)(7)-> mise -5(12.7)(5)-> 2016
EPR -8(13.6)(8)-> Finlande -5(10.4)(5)-> Flamanville
EPR -6(10.4)(5)-> Bouygues -5(11.8)(4)-> Flamanville
EPR -9(13.2)(8)-> Hollande -8(17.1)(7)-> François -5(10.2)(5)-> PS
EPR -8(13.6)(8)-> Finlande -4(11.6)(4)-> Olkiluoto
EPR -47(18.3)(42)-> nucléaire -5(10.7)(4)-> Bouygues
EPR -14(11)(12)-> Areva -5(14.2)(5)-> Olkiluoto
EPR -19(11.4)(16)-> réacteurs
EPR -5(11.2)(5)-> Penly

Figure 5.15 ENRG_FR 2011 : réseau poly-cooccurrentiel *EPR*

Les formes des 15 premières lignes ainsi que les lignes 18 et 20 permettent de restituer une série d'informations sur l'EPR en France : EDF construit une centrale avec un réacteur de troisième génération à Flamanville dans la Manche dont la construction des bâtiments est assurée par Bouygues. La mise en service est prévue en 2016, mais du retard est à prévoir. Le même type de réacteur EPR se trouve en Finlande et à Flamanville.

La 16e ligne montrerait la prise de position de François Hollande (PS) envers l'EPR lors de la pré-campagne présidentielle. La ligne 17 et 19 nous indique un EPR à Olkiluoto en Finlande avec Areva comme partenaire. La ligne 21 nous informe d'un EPR à Penly (Pays de Caux, Seine-Maritime).

[176] http://www.lefigaro.fr/conjoncture/2014/09/01/20002-20140901ARTFIG00374-l-epr-finlandais-d-areva-demarrera-avec-neuf-ans-de-retard.php (consulté le 20/08/2015)

Nous allons nous reporter aux textes du sous-corpus afin de vérifier ces informations.

<day=20111004> <article=3975> title: EPR Penly : l'enquête publique est reportée à 2012
 # L'association écologiste Robin des Bois avait déjà réclamé début septembre son report en attendant les résultats des tests de résistance en cours dans le parc nucléaire français. L'enquête publique sur le projet de réacteur nucléaire EPR de Penly (Seine-Maritime), qui devait débuter en octobre, est reportée à 2012 à la demande d'EDF. C'est ce qu'a annoncé, mardi 4 octobre, le ministère de l'énergie, assurant toutefois que le projet n'était "pas suspendu". (…….) # "L'enquête publique ne sera lancée qu'une fois que le dossier complet pourra être soumis à l'Autorité de sûreté nucléaire. Dans l'attente des compléments de dossiers en provenance d'EDF, l'enquête publique ne sera pas lancée avant 2012", # Quelle qu'en soit la cause, ce report repousse d'autant la mise en service finale du second EPR français, s'il était décidé de le construire. Une mauvaise nouvelle de plus pour le constructeur Areva et l'opérateur EDF, alors que le premier EPR, celui de Flamanville, a déjà vu sa mise en service retardée à 2016, quatre ans après le calendrier initial. # Pour aller plus loin : "L'EPR, chronique d'un chantier qui s'enlise"
<day=20111110> <article=4102> title: Sur le chantier de l'EPR à Flamanville, EDF est "à la moitié du chemin"
 # L'opérateur de l'EPR de Flamanville (Manche), maintient "la mise en service en 2016". Le chantier du réacteur nucléaire de troisième génération EPR de Flamanville (Manche), pomme de discorde entre le PS et les écologistes, est arrivé à la moitié de sa construction, a indiqué, jeudi 10 novembre, le maître d'œuvre EDF, qui prévoit toujours une mise en service en 2016, avec quatre ans de retard. (……) # De trois à quatre milliards d'euros ont été dépensés sur ce chantier qui doit en coûter six, a ajouté Hervé Machenaud, directeur exécutif en charge de la production et de l'ingénierie d'EDF. En 2005, le coût total avait été évalué à 3,3 milliards. # Mais, fort des leçons tirées des premières constructions d'EPR (un à Flamanville, un à Olkiluoto en Finlande et deux en Chine), Areva, le concepteur, espère gagner un milliard sur le coût de construction des EPR suivants, a indiqué Claude Jaouen, directeur de la branche réacteurs d'Areva. # Ce qui a été fait à Flamanville depuis fin 2007, "on le ferait aujourd'hui avec 20 mois de moins. On ne referait pas les mêmes erreurs", a ajouté Dominique Lagarde, directeur du secteur nouveau nucléaire d'EDF. Ces 20 mois, c'est plusieurs "centaines de millions" d'économie, selon M. Jaouen. # En début de semaine, Areva a indiqué que l'EPR en construction à Olkiluoto était achevé au deux tiers. La mise en service est prévue pour 2013, soit avec cinq ans de retard, mais l'ombre d'un nouveau retard plane sur ce chantier.
<day=20111112><article=4108> title: Le nucléaire paralyse les négociations PS-EELV
 # Pierre Moscovici et François Hollande, le 28 septembre, à Paris. Le député socialiste Pierre Moscovici, proche de François Hollande, a estimé dimanche 13 novembre sur Radio J que son parti avait fait des avancées "considérables" sur le nucléaire et qu'il revenait désormais à Europe Ecologie-Les Verts (EELV) de faire des concessions dans le cadre des négociations en cours pour 2012. Il a notamment cité des propositions du Parti socialiste "d'une ambition extraordinaire", comme le passage de 75 % d'énergie électrique d'origine nucléaire aujourd'hui à 50 % d'ici 2025, de fermer certaines centrales ou de ne pas faire de deuxième EPR. # Pierre Moscovici est en revanche resté ferme sur la question du premier EPR, actuellement en chantier à Flamanville (Manche), que les écologistes veulent arrêter. "François Hollande a pris un engagement que je redis ici : s'il est élu président de la République, nous ferons l'EPR de Flamanville", a-t-il affirmé. # Lire "Les vicissitudes de l'EPR, réacteur nucléaire de '3e génération'" # Il s'est toutefois montré optimiste sur l'issue des négociations entre le PS et EELV, déclarant qu'elles "ne sont pas dans l'impasse" mais sont "compliquées". "François Hollande et le PS souhaitent un accord avec les Verts : nous voulons diriger le pays avec eux. C'est important", a déclaré Pierre Moscovici. (……)

La consultation des articles concernés, cités ci-dessus, confirme notre hypothèse énoncée ci-dessus.

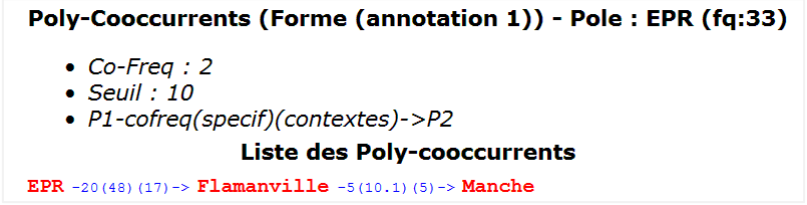


Figure 5.16 ENRG_FR : réseau poly-cooccurrentiel *EPR* pour l'année 2012 (jusqu'au mois d'avril)

Les échantillons textuels de l'année sont peu significatifs par rapport aux deux autres années, puisque seulement 4 mois de textes ont été récupérés.

> *<day=20120316> <article=4680> title: EPR de Flamanville : interruption pour plusieurs mois du bétonnage # L'opérateur de l'EPR de Flamanville (Manche), maintient "la mise en service en 2016". Il faudra "plusieurs mois"* ***pour remplacer les éléments défectueux du bâtiment du réacteur pressurisé européen (EPR) de Flamanville, a annoncé EDF*** *(le communiqué en PDF). Cela repousse d'autant la reprise du bétonnage du bâtiment, interrompue fin février. Cependant, l'opérateur maintient "la mise en service en 2016", a-t-il précisé. #* ***Pendant ces mois, il faudra pour le constructeur refabriquer quarante-six "consoles", des boîtes métalliques sur lesquelles doit prendre appui le futur pont de manutention du réacteur****. EDF a en effet décidé de les "remplacer en totalité" en raison de "défauts". Le 1er mars,* ***EDF avait déjà annoncé avoir interrompu le bétonnage en raison de défauts, dont l'ampleur et la gravité étaient en cours d'examen.*** *#* ***RETARD DE QUATRE ANS*** *# L'électricien va essayer de réorganiser le chantier pour avancer certains travaux qui devaient être effectués après la pose du dôme du réacteur, qui n'aura pas lieu à l'été, comme annoncé encore il y a quelques mois, a-t-il ajouté. # EDF a annoncé à deux reprises un report de la mise en service de l'EPR de Flamanville, qui a pris un retard de quatre ans. Le bâtiment réacteur a presque atteint sa hauteur finale. Le génie civil du chantier est terminé à 90 %, selon EDF. C'est maintenant l'électro-mécanique qui prend le relais. # Le coût de ce réacteur, lancé pour être une vitrine à l'exportation, a quasiment doublé à 6 milliard d'euros, contre 3,3 milliards en 2005. Il a été en 2011 au cœur de désaccords entre le PS et Europe Ecologie-Les Verts, ces derniers exigeant une suspension du chantier, que* ***François Hollande n'estime pas opportune****.*

Bien que les poly-cooccurrences de l'année 2012 relèvent peu d'information par sa quantité de formes, le retour au contexte à l'aide de l'article du 16 mars 2012 cité ci-dessus a permis d'en savoir un peu plus sur quelques causes susceptibles d'entrainer des retards successifs sur le chantier EPR à Flamanville.

EPR étant une forme peu présente, celle-ci constitue un véritable signal faible (Lesca, 2001) des sous-corpus. Afin d'accéder au cœur de la recherche sur *EPR* par la proximité segmentale, les calculs des segments répétés ont été appliqués sur ENRG_FR avec le seuil de sélection des segments à 10 (Salem, 1987, 1994, 2006).

Nous rappelons qu'un « signal faible » *est un « outil » d'aide à la décision. Il se présente comme une « donnée » d'apparence anodine mais dont l'interprétation que l'on en fait peut déclencher une alerte. Cette alerte indique que pourrait survenir un événement susceptible d'avoir des conséquences considérables (en terme d'opportunité, de menace ou de risque). Après interprétation le signal n'est plus qualifié de faible, mais de signal d'alerte précoce* (Lesca et Lesca, 2011). Nicolas Lesca a rappelé que *le concept de signal faible* [peut être considéré] *comme étant une donnée d'apparence insignifiante, noyée dans une multitude d'informations ayant une valeur informative pertinente afin d'alerter qu'un événement significatif aura lieu. De plus, un signal faible est anticipatif, fragmentaire, capté isolément ou en ordre dispersé, inondé par un volume significatif de données, incertain et peu fiable à priori*.[177]

[177] http://www.ressi.ch/num11/article_068 (consulté le 27/10/2015)

Les Segments Répétés (SR)[178] sont un module de Lexico 3 permettant de calculer les segments de formes qui se répètent dans toutes les séquences et toutes les différentes parties du corpus avec tous les choix de délimiteurs, demi-phrases, phrases, paragraphes, articles, jours, mois, années, etc. (Salem, 1987, 2006).

Les segments anaphoriques de la forme *EPR* : *réacteur EPR*, suite de 2 mots répétée 17 fois et *réacteur nucléaire de troisième génération,* suite de 5 mots répétée 13 fois, ont été retenus dans le cadre de notre recherche. Dans l'optique de visionner la localisation des articles concernés par la fouille textométrique, une projection des deux segments répétés a été effectuée sur la carte des sections et celle de la ventilation de formes.

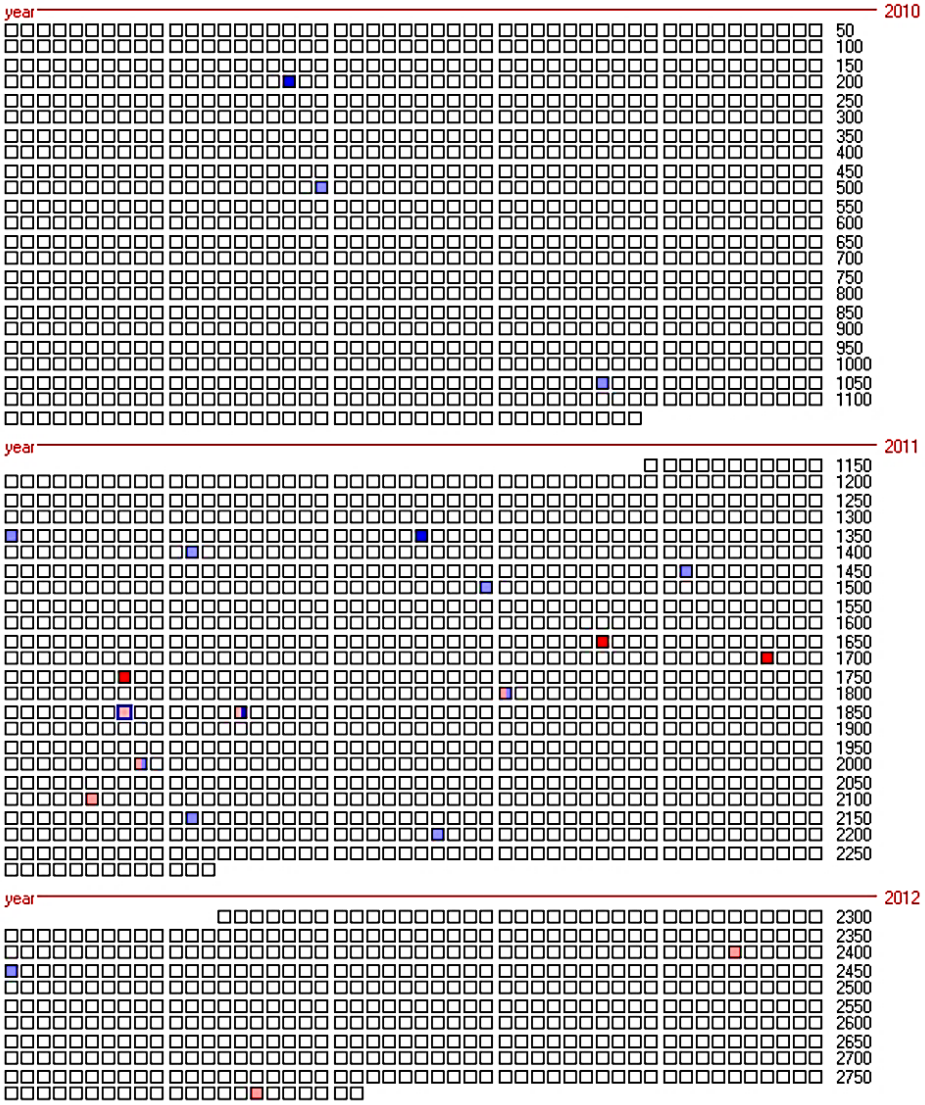


Figure 5.17 ENRG_FR de 2010 à 2012 : ventilation par année des segments répétés *réacteur EPR* en bleu et *réacteur nucléaire de troisième génération* en rouge

[178] Segments Répétés : Suite de formes dont la fréquence est supérieure ou égale à 2 dans le corpus.

Une vision globale de la répétition des deux segments répétés est ainsi réalisée. Nous pouvons constater que les deux segments ont fait couler beaucoup d'encre dans le Monde en 2011. La figure 5.17 montre également le nombre d'articles contenant les deux segments répétés :

- 2010 : 3 articles contenant le segment répété ***réacteur EPR***,
- 2011 : 7 articles contenant le segment répété ***réacteur EPR***, 5 articles contenant le segment répété ***réacteur nucléaire de troisième génération*** et 3 articles contenant simultanément les deux segments,
- 2012 : 1 article contenant le segment répété ***réacteur EPR*** et 2 articles contenant le segment répété ***réacteur nucléaire de troisième génération***.

Un recours à la ventilation des spécificités des deux segments répétés, ***réacteur EPR*** et ***réacteur nucléaire de troisième génération***, par mois, permet d'apporter des précisions supplémentaires sur la localisation et sur des informations émanant des segments concernés pour la période retenue par rapport à l'ensemble du reste des articles.

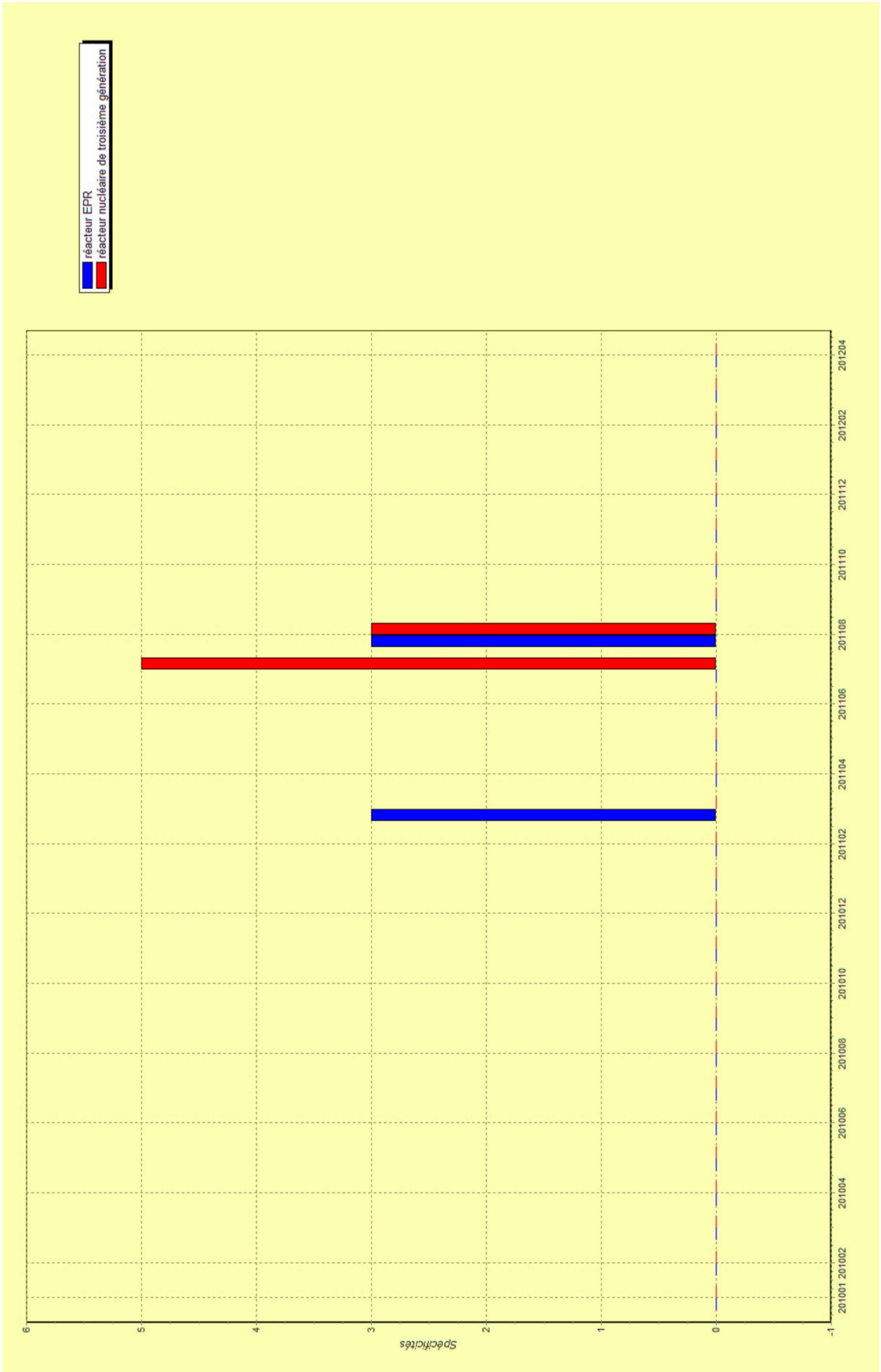


Figure 5.18 ENRG_FR de 2010 à 2012 : ventilation par mois des segments répétés *réacteur EPR* et *réacteur de troisième génération*

La ventilation, figure 5.18 ci-dessus, atteste une forte présence des deux segments répétés entre mars, et octobre 2011, en particulier le mois de mars, très touché par l'événement de Fukushima, les mois de juillet et août de cette même année, marqués par des états des lieux, des audits de contrôles accrus, etc. du parc nucléaire dans le pays le plus nucléarisé au monde.

Une projection par mois des deux segments répétés sur la carte des sections, permet d'accéder directement aux articles qui relatent les événements et les informations, voire à des renseignements vraisemblablement primordiaux pour la veille technologique et stratégique.

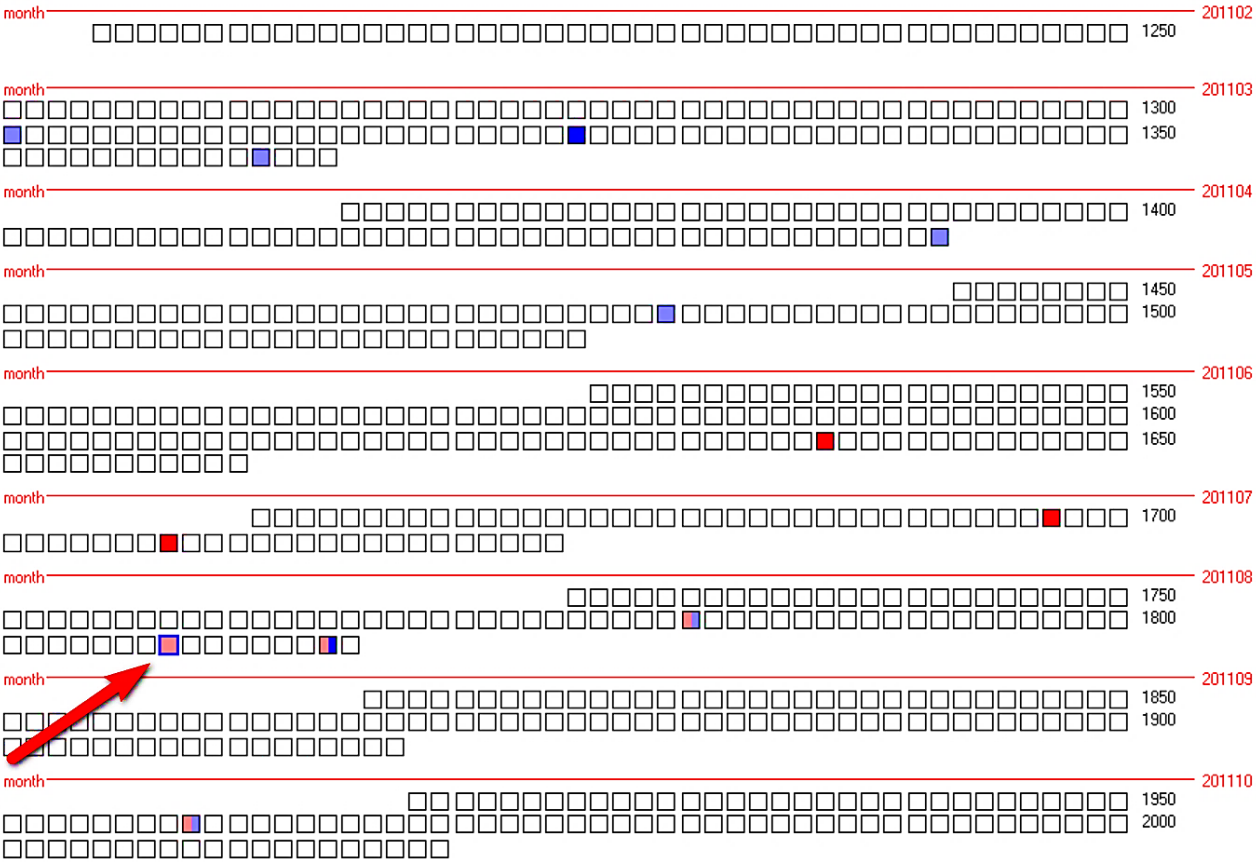


Figure 5.19 ENRG_FR de février à octobre 2011 : carte des sections pour les formes *réacteur EPR* en bleu et *réacteur nucléaire de troisième génération* en rouge, un carré = un article

Dans la carte des sections, lorsque les couleurs affichées sont atténuées, cela signifie que la répétition des formes sélectionnées est moindre par rapport aux couleurs plus foncées. La flèche rouge indique l'article contenant seulement le segment ***réacteur nucléaire de troisième génération***, article cité en entier ci-dessous, extrêmement révélateur puisqu'il relate la situation de l'ensemble des problèmes concernant le projet EPR en France et dans le Monde. En effet, c'est grâce à la consultation attentive de chacun des articles en couleurs que nous avons pu le repérer.

Nous avons marqué en rouge les mots liés aux multiples problèmes relevant de la construction de ces réacteurs EPR et à leur sécurité dans l'article ci-après.

<day=20110831><article=3854> title: L'EPR, chronique d'un chantier qui s'enlise

 # L'EPR de Flamanville en novembre 2009. Le temps est à l'orage au-dessus de l'EPR de Flamanville (Manche). Une fois de plus, le chantier du ***réacteur nucléaire de troisième génération*** *a été épinglé pour des défaillances. Cette fois, ce sont des malfaçons dans le gros œuvre qui ont fait l'objet d'une lettre au vitriol adressée par l'Autorité de sûreté nucléaire (ASN) à l'opérateur EDF, le 18 juillet, parmi d'autres courriers de réprimande révélés par Le Canard Enchaîné mercredi 31 août.*

 # La semaine dernière, treize autres faiblesses avaient déjà été constatées par le gendarme du nucléaire, tandis que samedi, on apprenait que le site n'était pas totalement aux normes sismiques. Et l'on ne compte plus les lettres, rapports ou documents internes égrénant les lacunes de la future centrale, tant vantée par le gouvernement, et présentée comme "la plus sûre au monde" par le fabricant Areva. En Finlande et en Chine aussi, où sont construits trois autres réacteurs du même type, les chantiers accumulent d'importants retards et sont la cible de nombreuses critiques. L'EPR, d'un fleuron du nucléaire français, est ainsi en passe de devenir l'une des technologies les plus décriées. # Sur le papier, le réacteur pressurisé européen (European pressurized reactor), conçu par Areva et l'allemand Siemens dans les années 1990, est censé représenter, en termes de sûreté, un modèle dans le monde. Le réacteur, d'une puissance de 1 650 mégawatts, aurait été conçu pour résister à la chute d'un avion gros porteur, et ses multiples systèmes de sécurité doivent le mettre à l'abri d'un accident détruisant le cœur du réacteur. Les piscines de refroidissement des combustibles usés seront même protégées par une enceinte de confinement. Au final, le risque de prolifération des matières radioactives serait quasiment nul.

 # MALFAÇONS DANS LA CONSTRUCTION

 # En réalité, depuis le début de la mise en chantier, en décembre 2007, du réacteur de Flamanville, les ingénieurs de l'ASN, qui contrôlent le site deux fois par mois, ont relevé plusieurs centaines de failles, consignées dans des compte-rendus d'inspection. # Ce sont tout d'abord des faiblesses de construction. Alors que le chantier progressait dans sa phase de génie civil, l'ASN a régulièrement mis en cause, ces derniers mois, des problèmes dans les opérations de bétonnage, féraillage et soudage. Dans des lettres adressées à EDF entre octobre 2010 et août 2011, et relevées par Le Canard Enchaîné mercredi, le gendarme égrène ainsi d'importantes malfaçons pouvant "porter préjudice à la qualité finale des structures" : "des piliers de béton percés comme du gruyère ou grêlés de nombreux 'nids de cailloux'", des "erreurs de ferraillage" ou encore "l'absence de nettoyage des fonds de coffrage, encombrés d'un mas de ligatures et autres objets non identifiés".

 # "Ce qui pose problème, c'est que ces défaillances sont récurrentes et portent sur nombre de structures du site et en particulier des éléments centraux de la sûreté de la future centrale", déplore Sophia Majnoni, chargée de campagne nucléaire chez Greenpeace France. Des trous ou des fissures ont ainsi été observés sur la cuve des réacteurs, le dôme qui protège le réacteur, ou encore le radier, c'est-à-dire la dalle de béton de sécurité située sous le réacteur. Des défaillances qui avaient poussé l'ASN à suspendre les travaux de bétonnage pendant trois mois, en mai 2008, une première dans l'histoire des centrales nucléaires françaises.

 # CONCEPTION DÉFAILLANTE

 # Bien plus embêtant, les insuffisances affectent aussi l'EPR jusque dans sa conception. Ainsi, les autorités de sûreté française, britannique et finlandaise demandaient-elles, en novembre 2009, d'"améliorer la conception initiale de l'EPR". Motif : une défaillance du système de contrôle-commande, cerveau du réacteur. Depuis, l'ASN n'a pas obtenu de réponse totalement satisfaisante de la part d'EDF et Areva. # L'EPR est, par ailleurs, régulièrement pointé du doigt pour son exposition aux risques. En 2003, le réseau Sortir du nucléaire a ainsi rendu public un document confidentiel défense, interne à EDF, qui montrait que le réacteur ne résisterait pas à la chute d'un avion de ligne. "Le risque terroriste n'a donc pas été pris en compte dans la conception, de même que le risque de séismes élevés, contre lesquels la sûreté des bâtiments est insuffisante, et le risque de sécheresse, les futurs réacteurs devant consommer 67 mètres cubes par seconde, soit bien davantage que les centrales actuelles qui puisent entre 3 et 10 mètres cubes par seconde", déplore Marc Saint-Aroman, chargé de mission pour l'ONG. # Du côté des autres centrales en construction dans le monde, on fait le même constat de multiples défaillances. "En Finlande, le chantier d'Olkiluoto, débuté en 2005 donc plus avancé qu'en France, a révélé des écarts dans les activités de montage du matériel électrique et mécanique", livre Thomas Houdré, directeur des centrales nucléaires au sein de l'ASN. En Chine, à Taishan, les travaux des deux réacteurs nucléaires, débutés en 2009, rencontrent aussi des problèmes de construction, selon Greenpeace.

 # SÛRETÉ BRADÉE

 # "Qu'il y ait des écarts sur des chantiers de cette ampleur est inévitable, tente de relativiser Thomas Houdré. Ce qui est importe, c'est que tout soit mis en conformité par les opérateurs avant la fin du chantier." # Mais pour les associations, ces futurs réacteurs font craindre une sûreté bradée, malgré les nombreux contrôles, dans la mesure où les opérateurs subissent d'énormes pressions économiques et politiques pour achever les chantiers au plus vite. "EDF a dû arbitrer entre la sûreté et les coûts. Pour réagir aux pénalités, il accélère les cadences, ce qui multiplie les problèmes", s'inquiète Charlotte Mijeon, porte-parole du réseau Sortir du nucléaire. # Les problèmes à répétition ont en effet considérablement ralenti l'avancement des travaux. La mise en service de Flamanville 3,

> *initialement prévue l'an prochain, a ainsi été reportée à 2016, tandis que la facture du chantier a été revue à la hausse, à 6 milliards d'euros, soit près du double des estimations initiales. De la même façon, en Finlande, Olkiluoto 3 ne démarrera pas avant 2013, soit avec quatre ans de retard et un budget aussi multiplié par deux.*
> *# PERTE DE SAVOIR-FAIRE*
> *# Fallait-il, alors, se lancer dans de tels chantiers, les plus gros jamais construits dans le monde ? A l'origine, l'EPR de Flamanville a été retenu en France pour renouveler des compétences qui disparaissent peu à peu, alors que les ingénieurs du programme électronucléaire français, qui ont construit les 58 réacteurs de l'Hexagone entre 1970 et 1990, partent à la retraite. # Mais EDF a sous-estimé la lourdeur et la complexité du projet. Et l'ingénierie nucléaire française a perdu de son savoir-faire et de sa performance, dans la mesure où vingt ans séparent la dernière centrale construite, à Civaux, dans la Vienne, de Flamanville. Enfin, coordonner des milliers d'ouvriers et d'ingénieurs de toutes les nationalités employés par plusieurs niveaux de sous-traitants complique encore la tâche. A tel point que la mise en chantier du cinquième EPR, prévue l'an prochain à Penly (Seine-Maritime), a été gelée. # "Le programme EPR est un désastre industriel et financier, en plus d'être dangereux. L'ASN devrait décider d'arrêter le chantier", estime Sophia Majnoni. A la fin du mois de mars, après l'accident dans la centrale japonaise de Fukushima, le gendarme du nucléaire avait bien évoqué un moratoire sur le chantier de Flamanville. Mais l'idée a finalement été écartée, en raison des pressions politiques et économiques. §*

La consultation de l'intégralité de cet article cité ci-dessus expose l'historique, les péripéties et les véritables causes du retard de la mise en service de l'EPR en France.

Cet article dresse un état du chantier concernant la construction de l'EPR français à Flamanville au mois d'août 2011, sachant que sa construction a démarré quatre ans plus tôt (en 2007) et qu'elle devait se terminer en 2014. Or, depuis la parution de cet article, la date de mise en service a encore reculé à 2017 voire plus tard suite à divers contretemps. Fukushima a déclenché une prise de conscience de toutes les autorités de l'ampleur des conséquences d'un tel événement naturel. Les contrôles de sécurité et de sûreté des installations en cours et existantes sont renforcés de façon draconienne, les pires scénarios catastrophes sont désormais envisagés. La vitrine du nucléaire français tant vantée par ses concepteurs rencontre de sérieuses difficultés dans la conception de sa structure et ce sur les trois chantiers dans le monde (en Finlande, en France et en Chine), toutes ces informations sont annoncées par les ONG, notamment Greenpeace. Les contrôles effectués par l'organisme ASN sont permanents et les non-conformités dénoncées sans qu'EDF et Areva n'apportent de véritables réponses à tous ces manquements. Les chantiers trainent en longueur, ce qui a pour effet d'augmenter les coûts financiers et les délais, qui sont sans cesse revus à la hausse. S'ajoutent également des problèmes de management des équipes dans l'un des chantiers, le plus important d'Europe, voire du monde.

Nous venons de mettre en évidence par une expérience empirique la recherche de l'information sur l'EPR dans le monde par la veille textométrique, méthode s'avérant très efficace pour la restitution de l'information et de la détection de signaux faibles.

En dépit de la petite quantité de production de la forme *EPR*, les informations associées à cette forme peuvent être considérées comme de type signaux faibles. L'approche de la textométrie permet toutefois de les détecter, *« informations parfois difficilement perceptibles ou identifiables »* (Ansoff 1975 ; Hermel, 2010 : 94 ; MacMurray, 2012).

Rappelons que les signaux faibles sont l'un des types des informations recherchées (Ansoff, 1975, 1989). Ils sont par nature essentiellement vagues, qualitatifs, anticipatifs, incertains et équivoques. Ils peuvent se préciser par une veille anticipative. Ces informations anticipatives difficiles à prendre en compte devraient permettre d'éclairer des changements futurs dans l'environnement de l'entreprise (Lesca, 2001 ; Caron-Fasan, 2001).

Au-delà de l'exactitude de ces informations, que pouvons-vous prévoir et anticiper ?

5.4.11 Prévision et anticipation

Les difficultés et problèmes révélés au cours de la construction de l'EPR dans la Manche permettront de bénéficier des retours d'expériences et d'anticiper la mise en place de bonnes pratiques pour la naissance de ses frères jumeaux dans le monde (en Finlande et en Chine).

L'EPR, projet phare du nucléaire défendant le grand savoir-faire des tricolores, est soumis à des épreuves d'exigences sans précédent. Les enjeux de sûreté et de sécurité nucléaire ont plus que jamais été révisés et défendus malgré les énormes pressions sociales, politiques et économiques. La construction en Finlande se poursuivrait en tirant profit des leçons françaises et probablement devancerait celle de Flamanville. Mais à la vue des déboires successifs rencontrés par les deux centrales européennes, la première véritable centrale EPR susceptible de voir le jour serait en Chine dans la ville de Taishan. Cette expérience démontre encore une fois la nécessité de l'efficacité de la coordination sociale, politique et économique dans le développement des technologies de pointe de grande envergure, la clé de toutes les «*Success Stories*».

Mais que se passe-t-il aux États-Unis ?

5.4.12 Cooccurrences évolutives autour de la forme *energy* sur le sous-corpus américain

En anglais, la forme *energy* est un nom généralement non-comptable, c'est-à-dire, elle reste dans la plupart des cas au singulier. Sa forme plurielle *energies* est réservée lorsqu'il s'agit d'efforts physiques et mentaux d'une personne pour faire quelque chose, « *energies : A person's physical and mental powers, typically as applied to a particular task or activity*»[179]. Nous proposons la traduction suivante : Pouvoirs physiques et mentaux d'une personne, généralement appliqués à une tâche ou une activité particulière (Traduction de l'auteur). Dans notre cas, la forme *energy* est toujours au singulier[180].

Les calculs de cooccurrences évolutives ont été effectués avec les paramètres par défaut (Co-fréquence 2, seuil 10) à partir de la forme spécifique *energy* sur ENRG_US afin d'obtenir un réseau de mots cooccurrents évolutifs (figures disponibles dans l'annexe I). Des restitutions d'informations seront consacrées à la forme-pôle *energy* afin de mieux cerner le contexte.

Guide de lecture pour les tableaux de cooccurrences

Dans l'entête du tableau

- **Pôle** : forme-pôle choisie afin de trouver les formes qui vont être co-présentes avec celle-ci
- **fq** : fréquence totale d'apparition de cette forme dans le corpus
- **Co-Freq** : nombre minimum de la co-apparition des deux formes dans le corpus
- **Seuil** : seuil minimum pour le calcul de probabilité de la spécificité du cooccurrent

Dans le corps du tableau

- **Cooccurrents** : forme co-présente avec la forme-pôle
- **Fq (Cooc)** : fréquence absolue du cooccurrent dans le corpus
- **co-Fq** : nombre de la co-apparition des deux formes dans le corpus
- **specif** : coefficient de la spécificité du cooccurrent dans le corpus
- **contextes** : nombre de séquences où la forme-pôle et son cooccurrent sont co-présents

[179] http://www.oxforddictionaries.com/definition/american_english/energy

[180] Energy: [uncountable] a source of power, such as fuel, used for driving machines, providing heat, etc.
Ex: solar/nuclear energy. It is important to conserve energy. An energy crisis (= for example when fuel is not freely available). http://www.oxfordlearnersdictionaries.com/definition/english/energy?q=energy

Tableau 5.15 ENRG_US de 2010 à 2012 : synthèse des résultats des cooccurrences évolutives autour de la forme *energy*

2010							2011						2012
Pôle : *energy* ; fq : 663. Co-Freq : 2 | Seuil : 10							Pôle : *energy* ; fq : 663. Co-Freq : 2 | Seuil : 10						Pôle : *energy* ; fq : 89.
Pôle	**cooccurrents**	**équivalent en français**	**Fq (Cooc)**	**co-Fq**	**specif**	**contextes**	**cooccurrents**	**équivalent en français**	**Fq (Cooc)**	**co-Fq**	**specif**	**contextes**	
energy	climate	climat	276	44	10.2	41	climate	climat	377	57	10.2	52	
energy	solar	solaire	154	53	27.8	47	nuclear	nucléaire	133	42	19.2	37	
energy	fossil	fossile	53	20	12.4	20	solar	solaire	101	28	11.9	26	
energy	efficient	efficace	46	22	16.1	21	policy	politique	121	36	15.8	34	
energy	renewable	renouvelable	77	58	**	57	power	puissance	289	60	16.8	55	
energy	power	puissance	286	47	11.3	46	renewable	renouvelable	89	68	**	63	aucune donnée
energy	efficiency	efficacité	42	22	17.2	22	efficiency	efficacité	48	25	18.3	25	
energy	fuels	carburants	56	22	13.9	20	sources	sources	80	38	25.0	38	
energy	sources	sources	73	23	12.2	22	certificates	certificats	11	9	10.2	9	
energy	wind	vent	108	33	16.2	29	clean	propre	139	58	33.3	53	
energy	clean	propre	145	58	34.3	57							

Pour des raisons de visibilité, les formes communes obtenues par les calculs des cooccurrences entre 2010 et 2012 (jusqu'au mois d'avril) ont été colorées en orange dans le tableau 5.15 ci-dessus.

5.4.13 Synthèse d'informations d'*energy dans* ENRG_US

Formes communes des cooccurrences : ***climate, solar, renewable, power, efficiency, sources, clean***

Par analyse longitudinale du tableau ci-dessus, la forme *energy* est cooccurrente avec les formes (*climate, solar, renewable, power, efficiency, sources, clean*) dans les deux années du sous-corpus. L'analyse de ces formes communes en anglais américain (tableau 5.15) démontre que le champ sémantique de la forme *energy* présente une forte corrélation de fréquence avec *climate* (276 et 377) et *power* (286 et 289), mais respectivement en coprésence, 44 fois et 57 fois pour *climate*, 47 fois et 60 fois pour *power*. Les formes *renewable* et *clean* sont les formes les plus cooccurrentes en 2010 (co-Fq 58), et en 2011 ce sont *renewable* et *power* les plus cooccurrentes. La spécificité de la forme renouvelable (renewable) tend vers l'infini dans les deux années et présente une forte répétition dans les phrases, ceci est dû à l'expression énergies renouvelables (*renewable energy*). Quant à l'expression énergie propre (clean energy), celle-ci reste également récurrente.

Les formes communes aux deux années nous indiquent une préoccupation constante du changement climatique, ainsi que la recherche et le développement des énergies propres, renouvelables et efficaces, en particulier, une volonté d'exploitation de l'énergie solaire.

En 2010, nous constatons que les formes associées à *energy* sont principalement les types d'énergies et les sources d'énergies tels que l'énergie solaire, éolien et fossile, alors qu'en 2011, les énergies fossiles et l'énergie du vent disparaissent, mais le nucléaire (*nuclear*) et la politique énergétique (*policy et certificates*) sont apparus.

Nous faisons deux hypothèses concurrentes : l'apparition de la forme nucléaire dans les cooccurrences est peut-être liée soit à l'événement de Fukushima, soit à une politique énergétique favorisant le nucléaire.

A l'issue des recherches infructueuses par nos outils et méthodes textométriques sur les informations relayant des impacts politico-économiques clairement affichées dans les articles du sous-corpus, nous recourons à des sources externes dans l'optique de récupérer des informations complémentaires.

Un autre article du même journal paru le 17 mars 2011, mais classé sous la rubrique *US/politics* nous dévoile la face cachée de ces informations par une prise de position du président Obama.

Un extrait de l'article[181]

POLITICS >>>Obama's Speech on Japan (Text) >>>MARCH 17, 2011[182]

(……)Here at home, nuclear power is also an important part of our own energy future, along with renewable sources like wind, solar, natural gas and clean coal. Our nuclear power plants have undergone exhaustive study, and have been declared safe for any number of extreme contingencies. But when we see a crisis like the one in Japan, we have a responsibility to learn from this event, and to draw from those lessons to ensure the safety and security of our people.

That's why I've asked the Nuclear Regulatory Commission to do a comprehensive review of the safety of our domestic nuclear plants in light of the natural disaster that unfolded in Japan. (……)

Traduction française

Titre : Discours d'Obama sur le Japon (Texte), le 17 mars 2011

(……)Ici, chez nous, l'énergie nucléaire occupe également une place importante de notre propre avenir de l'énergie, ainsi que les sources renouvelables comme le vent, l'énergie solaire, le gaz naturel et le charbon propre. Des études exhaustives ont été menées sur nos centrales nucléaires, celles-ci ont été déclarées sûres pour toutes sortes de contingences extrêmes. Mais quand nous voyons une crise comme celle du Japon, nous avons la responsabilité d'appréhender la situation de cet événement, et d'en tirer des leçons afin d'assurer la sûreté et la sécurité de notre peuple.

C'est pourquoi j'ai demandé à la Commission de réglementation nucléaire de faire un examen complet de la sécurité de nos centrales nucléaires nationales à la lumière de la catastrophe naturelle qui s'est déroulée au Japon. (……)

L'extrait démontre que la catastrophe de Fukushima ne semble pas exercer d'influence sur la stratégie politique nucléaire, mais engendrer simplement un regain du renforcement de sa sûreté et de sa sécurité.

Après avoir appréhendé ces révélations, nous tentons de comprendre comment la forme *Fukushima* se répertorie dans la rubrique *environment* du NYT.

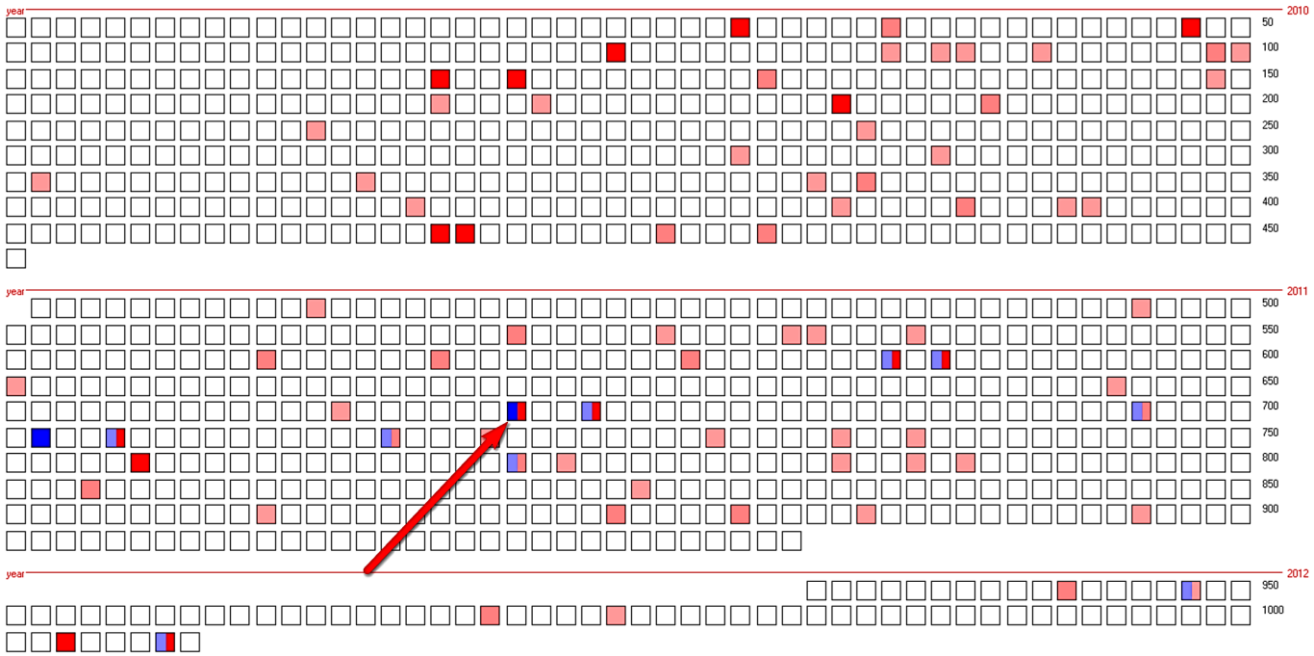


Figure 5.20 ENRG_US de 2010 à 2012 : carte des sections pour les formes *Fukushima* en bleu et *nucléaire* en rouge, un carré = un article

[181] L'intégralité de cet article est disponible : https://www.whitehouse.gov/blog/2011/03/17/president-obama-we-will-stand-people-japan (consulté le 25/05/2015)

[182] http://www.nytimes.com/2011/03/17/us/politics/18obama-japan-text.html (consulté le 25/05/2015)

L'extrait de l'article fléché sur la carte des sections rapporte la position catégorique post-Fukushima de l'Allemagne (Rüdinger, 2013).

> *<day=20110512><<article=3684> title : Panel Urges Germany to Close nuclear Plants by 2021*
> *# The recommendations, which have not been made public, will go to a panel of specialists meeting in a closed session in Berlin this weekend. Mrs. Merkel said this week that Germany would certainly end its reliance on nuclear energy and that the only question was how long nuclear would be needed as a "bridge technology" until other forms of energy could meet the country's needs. # Nuclear energy provides 22.6 percent of Germany's electricity, according to the energy Ministry. (……)*
> *She quickly changed her mind in March, as the damage to the Fukushima Daiichi plant became apparent. She ordered seven of Germany's power plants to be temporarily closed, instituted a moratorium on construction of new reactors, ordered an intensive review of security and safety measures, and appointed the Ethics Commission. # She announced the decision days before regional elections in southwestern Germany, where the Greens soundly defeated the governing conservatives. §*

Traduction française

> <Jour=20110512><<article=3684> Titre: Le comité invite instamment l'Allemagne à fermer ses usines nucléaires d'ici 2021
> # Les recommandations, qui n'ont pas été rendues publiques, iront à un comité de spécialistes réunis à huis clos à Berlin ce week-end. Mme Merkel a déclaré cette semaine que l'Allemagne mettrait certainement fin à sa dépendance à l'énergie nucléaire et que la seule question était combien de temps serait nécessaire pour que le nucléaire se maintienne comme une technologie de transition jusqu'à ce que d'autres formes d'énergie soient susceptibles de répondre aux besoins du pays. # L'énergie nucléaire fournit 22,6 % de l'électricité de l'Allemagne, selon le ministère de l'énergie. (……)
> Elle a rapidement changé d'avis en Mars suite à l'évidence des dommages causés à la centrale de Fukushima Daiichi. Elle a demandé la fermeture temporaire de sept des centrales allemandes, institué un moratoire sur la construction des nouveaux réacteurs, ordonné un examen approfondi des mesures de sécurité et de sûreté, et nommé une commission d'éthique. # Elle a annoncé ses décisions les jours précédant les élections régionales dans le sud-ouest de l'Allemagne, où les Verts ont battu les conservateurs au pouvoir.

La carte des sections montre la disparition des formes *nuclear* (en rouge) et *Fukushima* (en bleu) dans les articles d'ENRG_US. Il est évident que la catastrophe de Fukushima n'a que très peu de poids dans cette rubrique *environnent* du NYT. Bien que les articles d'actualités, de politiques nationales et internationales soient classés dans d'autres rubriques, cette faible apparition de la forme *Fukushima* conforte notre hypothèse de départ.

L'article sur la dénucléarisation radicale de l'Allemagne explique que le NYT veille de près à la politique énergétique européenne, probablement pour en extraire les bonnes pratiques en matière de développement d'énergies nouvelles. Il est à noter que l'organisation politique est sensiblement semblable en Allemagne et aux États-Unis, une partie de la surveillance du nucléaire s'effectuant aux niveaux respectivement des *länder* et des Etats[183].

Notre expérience montre que peu d'informations intéressantes se dégagent du sous-corpus ENRG_US contrairement à celles d'ENRG_FR, toutefois elle nous livre certaines pistes de réflexion : d'une part, le nucléaire est un secteur sensible, par conséquent, les informations sont difficilement accessibles, d'autre part les sociétés nucléaires aux États-Unis ne sont pas totalement étatiques, ainsi elles ont le droit de protéger leurs données privées.

Les articles du NYT, classés probablement en fonction du lieu où l'événement s'est produit, apporteraient des complexités supplémentaires, à moins que les outils exploitant des mégadonnées soient disponibles, pour les études transversales et multidisciplinaires, c'est-à-dire, les traitements automatisés multimodaux et croisés, traitements permettant de rechercher simultanément les informations relatives aux identités, lieux, événements, vidéos, images, etc.

[183] Extrait d'un entretien sur ARTE Journal du 30 août 2013 : http://www.arte.tv/fr/nucleaire-les-dinosaures-n-ont-pas-leur-place-dans-un-champ-de-fleurs/7633092,CmC=7633414.html

5.4.14 Synthèse d'informations d'*EPR* dans ENRG_US

A la différence d'ENRG_FR, la forme *EPR* dans ENRG_US pourrait prendre d'autres formes :

- E.P.R., for European pressurized reactor,
- EPR, third generation, pressurized water reactor (PWR),
- US EPR ou US-EPR.

Cependant, toutes les formes citées ci-dessus sont absentes dans ENRG_US, d'autres formes sont susceptibles d'être liées ou associées à EPR, telles que

- Areva ou AREVA,
- Flamanville,
- Manche,
- Olkiluoto,
- Taishan (ville en Chine où l'EPR est en cours de construction).

Parmi toutes les formes mentionnées ci-dessus, seule la forme *Areva* n'est apparue qu'une seule fois dans un article relatant les traitements de déchets nucléaires (figure 5.21, ci-dessous).

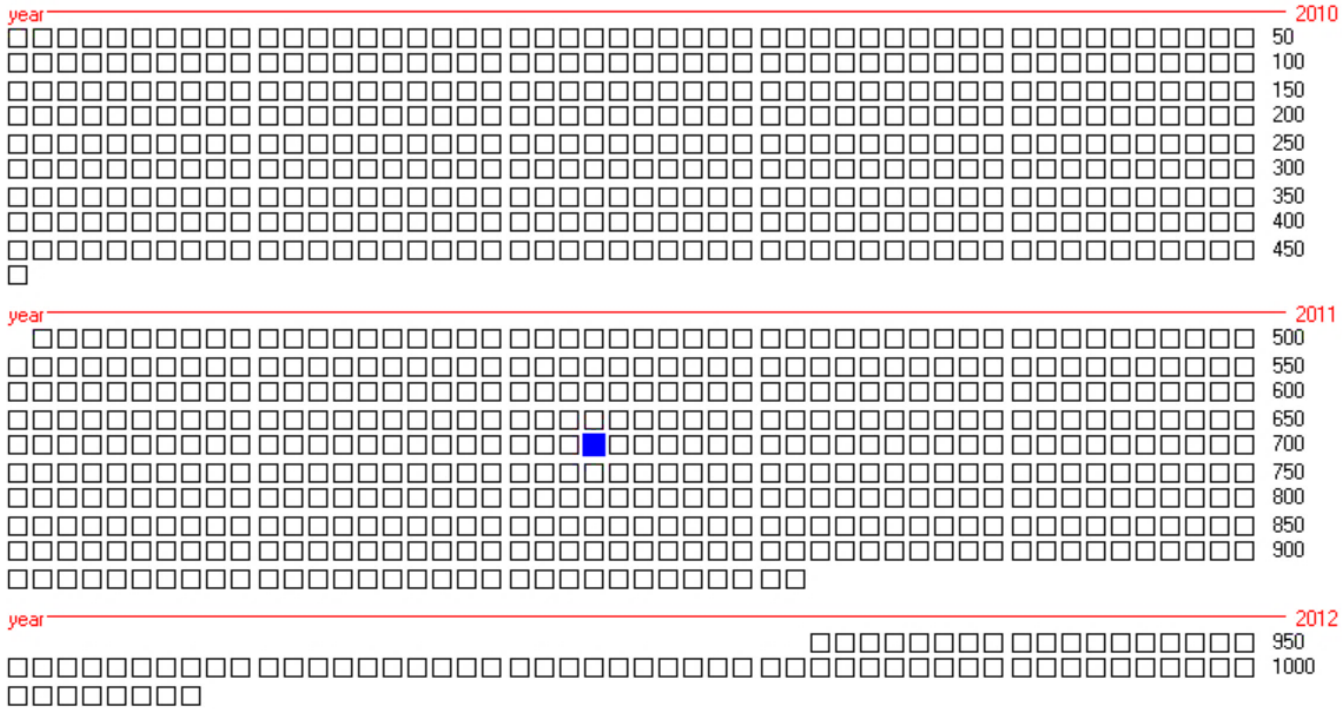


Figure 5.21 ENRG_US de 2010 à 2012 : carte des sections pour la forme *Areva* en bleu, un carré = un article

Le passage concerné

<day=20110513><article=3687> title : Panel on nuclear Waste Disposal to Propose Above - Ground Storage
*(…….)But recent studies argue that there is plenty of virgin uranium and thus no reason to recycle. And American utilities have been reluctant to use the plutonium fuel, even when asked by the energy Department in an effort to help dispose of surplus plutonium from the weapons program. # **Areva**, the French nuclear company, has been arguing that recycling cuts the cost of disposal and eliminates the need to mine uranium, which itself is environmentally damaging. (…….)*

Traduction française

<day=20110513><article=3687> title : Un groupe d'études propose le stockage des déchets nucléaires en surface
(……)Mais des études récentes font valoir qu'il y a beaucoup d'uranium pur et donc aucune raison de le recycler. Les services publics américains ont été réticents à utiliser le combustible au plutonium, même lorsque le Département de l'énergie lui a demandé de faire un offort pour disposer des surplus de plutonium issu du programme d'armement. # Areva, société nucléaire française, a fait valoir que le recyclage réduit le coût de stockage et évite l'exploitation des mines d'uranium, qui est dommageable pour l'environnement. (.....)

Un seul article évoque le savoir-faire d'Areva dans les traitements et retraitements des déchets nucléaires aux États-Unis, malgré la place importante occupée par les technologies nucléaires.

Afin de vérifier l'exactitude du classement des articles contenant la forme *EPR* par le NYT, une requête de recherches a été exécutée dans leur moteur de recherche interne. Le résultat obtenu pour la période du 1er janvier 2010 au 31 décembre 2012, figure 5.22 ci-dessous, conforte notre hypothèse sur le classement des articles.

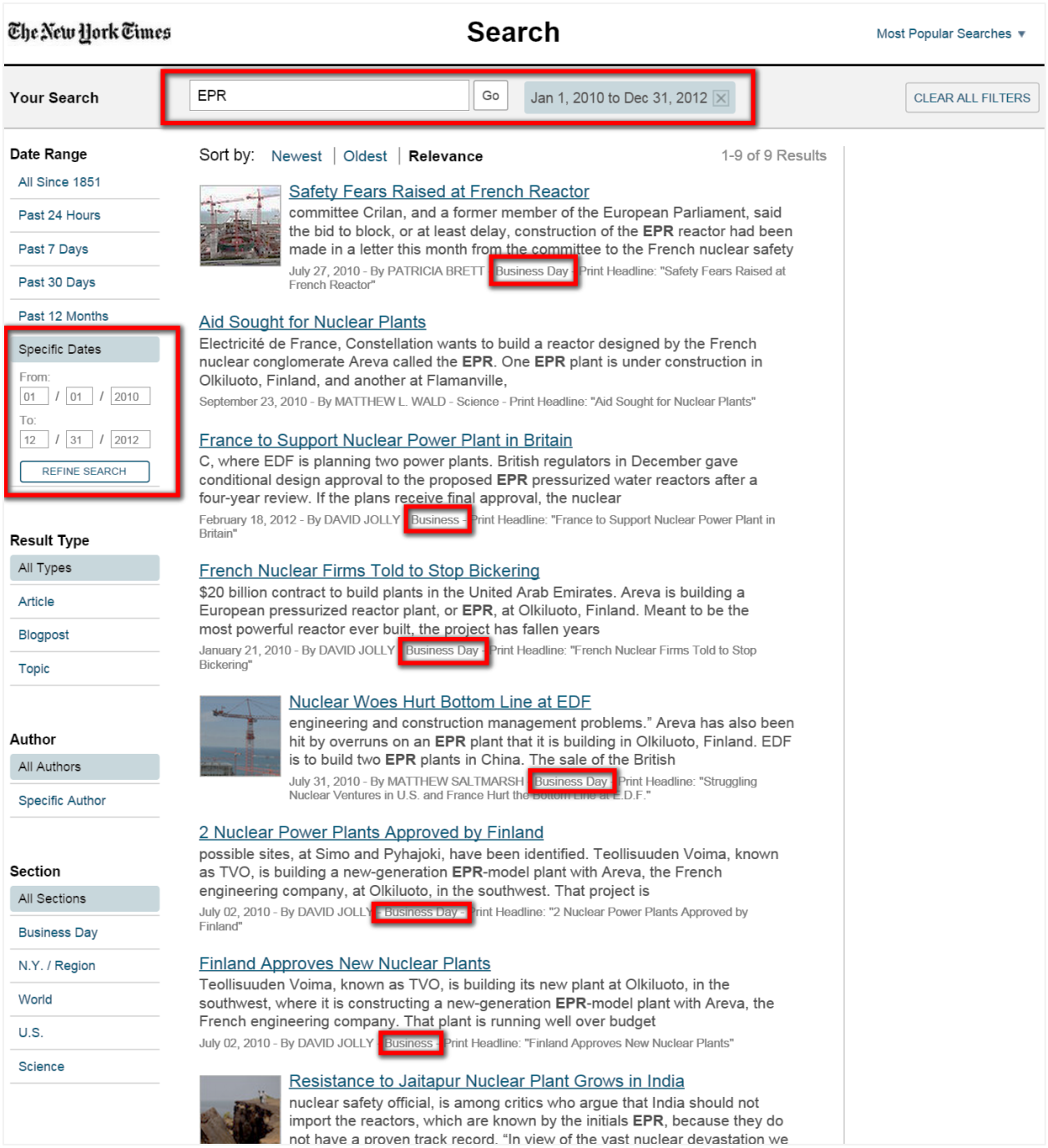


Figure 5.22 Extrait de la page des résultats de la recherche de la forme *EPR* dans le moteur de recherche interne de nyt.com, consulté le 24/08/2015

La figure 5.22 montre clairement que les articles concernant la forme *EPR* ont été classés dans la rubrique *Business ou Business Day*. Ceci explique la raison pour laquelle nous n'avons pas d'articles évoquant le sujet EPR.

Pour des raisons pratiques, les informations américaines sur l'EPR étant absentes, les restitutions, les prévisions et les anticipations générales de ce dernier s'effectueront à la fin de l'exploration du sous-corpus chinois, chapitre 6.

Conclusion du chapitre

La comparaison entre les sous-corpus français et américain a permis d'établir les constats suivants : d'une part, une variation du nombre de formes entre janvier 2010 et avril 2012 pour le sous-corpus ENRG_FR, d'autre part, concernant le sous-corpus ENRG_US, les résultats ont montré des variations relativement faibles de formes et d'hapax entre janvier 2010 et avril 2012. Par ailleurs des irrégularités relatives à la répartition du nombre d'occurrences et du nombre d'articles ont été constatées dans les deux sous-corpus entre janvier 2010 et avril 2012. Ces variations d'emploi de formes peuvent s'expliquer en partie par le nombre de catastrophes naturelles sur cette période.

Nos différentes analyses quantitatives des textes des articles des deux journaux utilisés à savoir Le Monde et le *New York Times*, pour les sous-corpus français et américain sur le thème de l'énergie, nous montrent que les informations événementielles ne sont pas traitées de la même façon dans ces deux quotidiens. Nous avons constaté que les journaux relatent communément les événements importants. L'absence de groupement commun dans les AFC a été relevée entre ENRG_FR et ENRG_US. Les analyses ont aussi permis de démontrer que l'importance accordée à certains événements n'était pas la même selon le pays. Ainsi, le *New York Times* donne plus d'importance à la pollution du Danube en Hongrie que le journal Le Monde. En France, les différentes analyses textométriques menées ont permis de voir que les sujets se rapportant au nucléaire et au changement climatique demeurent des préoccupations majeures des Français.

Le chapitre 6 sera consacré au sous-corpus ENRG_CN.

6. *Environnement*, *énergies* et *EPR* dans le sous-corpus chinois ENRG_CN

Nous commencerons par présenter ce sous-corpus issu d'articles de la rubrique vert de *QQ.com*, et de la rubrique Protection de l'environnement du site *sina.com.cn*. Ces sites mettent à la disposition des articles provenant de 1 123 médias. Nous expliquerons par la suite les raisons pour lesquelles la période de 2010 à 2012 a été retenue. Les principales caractéristiques textométriques de ce sous-corpus seront aussi décrites, puis nous extrairons notamment les données suivantes :

- nombre de paragraphes,
- nombre d'occurrences,
- nombre de formes,
- évolution du nombre d'articles,
- accroissement du vocabulaire,
- segments répétés.

La proximité textuelle, les spécificités positives et les cooccurrences seront aussi étudiées. Le chapitre se terminera par une comparaison des trois sous-corpus d'ENRG restreinte aux seuls thèmes des EPR et des énergies.

6.1 Présentation du sous-corpus issu de supports divers

Rappelons que le sous-corpus chinois, ENRG_CN, provient d'un regroupement d'articles de deux sites majeurs d'informations qui proposent eux-mêmes les articles de près de 1 100 médias, agences de nouvelles, presse en ligne etc. : les *news* de la rubrique *Vert* du site *QQ.com* et les *news* de la rubrique *protection de l'environnement* du site sina.com.cn, couvrant la période du 23-03-2008 au 23-04-2013. La taille du fichier en .txt est de 67,1 Mo (déjà segmenté en mots).

6.1.1 Une période de rupture

La date de début correspond au premier article disponible dans les deux rubriques et la date de fin est celle du dernier lancement de la requête informatique.

Principales caractéristiques textométriques d'ENRG_CN

Tableau 6.1 ENRG_CN : principales caractéristiques textométriques

Nombre d'occurrences: 10447521 Nombre de formes: 98699
Nombre d'hapax: 33669 Fréquence maximale: 551602

	Num	Partie	Occurenc	Formes	Hapax	Fmax	Forme
✓	1	2008	3984	1231	684	323	的
✓	2	2009	930	529	405	60	的
✓	3	2010	3460916	59155	20138	188717	的
✓	4	2011	4919088	70072	23239	256106	的
✓	5	2012	2062188	45583	15252	106369	的
✓	6	2013	415	256	185	27	的

Le tableau 6.1 nous montre l'ensemble du sous-corpus qui compte 10 447 521 occurrences et 98 699 formes. Il se divise en 221 569 paragraphes. La forme la plus fréquente est la préposition du chinois « 的, de».

La particularité d'ENRG_CN réside dans la diversité de l'origine des articles. En effet, il ne s'agit pas d'un seul journal, mais d'un groupement de 1 123 médias, dont la quasi-totalité de la presse écrite chinoise en ligne.

L'évolution du nombre d'articles du sous-corpus chinois

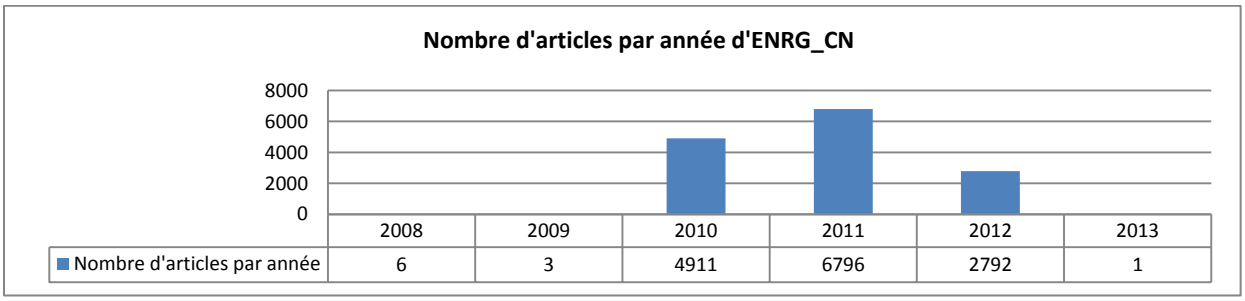


Figure 6.1 ENRG_CN : évolution du nombre d'articles par année

Le tableau 6.1 et la figure 6.1 ci-dessus montrent que les années 2008 et 2009 sont particulièrement peu productives, seulement 9 articles au total. A partir de 2010, une recrudescence de production d'articles est constatée avec 4 911 articles et 3 460 916 occurrences. Elle atteint son apogée en 2011 avec 6 796 articles et 4 919 088 occurrences. Un seul article a été récupéré au printemps 2013 (mars) dont le nombre d'occurrences est 415. Le nombre d'articles de l'année 2010 est presque le double de celui de l'année 2012 (2 792 articles), tandis que celui de l'année 2011 est presque le triple de l'année 2012.

Une première remarque est à noter à propos de la quantité textuelle totale d'ENRG_CN, cette quantité dépasse largement ENRG_FR et ENRG_US sur le nombre total d'articles de la période comparable (2010, 2011 et 2012). Cela est dû aux nombreuses sources textuelles lors des récupérations des données.

Une deuxième remarque porte sur le nombre d'articles de la période 2008 à 2013 (années 2008, 2009, 2013 incomplètes) témoignant d'une grande disparité au cours de la période. L'année 2011 occupe la première place tant par son nombre d'articles que par son nombre d'occurrences ou son nombre de formes. Ce résultat confirme que plus le nombre d'articles est grand, plus il y a de formes.

Cette exubérance de production d'articles est identique dans les deux sites (Sina et QQ). Ce changement important s'explique par une volonté politique de s'adapter aux exigences énergétiques internationales et par la prise de conscience étatique et publique face aux pollutions grandissantes (se reporter à l'annexe G).

Afin de visualiser une évolution plus fine d'ENRG_CN, la figure 6.2 ci-dessous trace la répartition du nombre d'articles par mois.

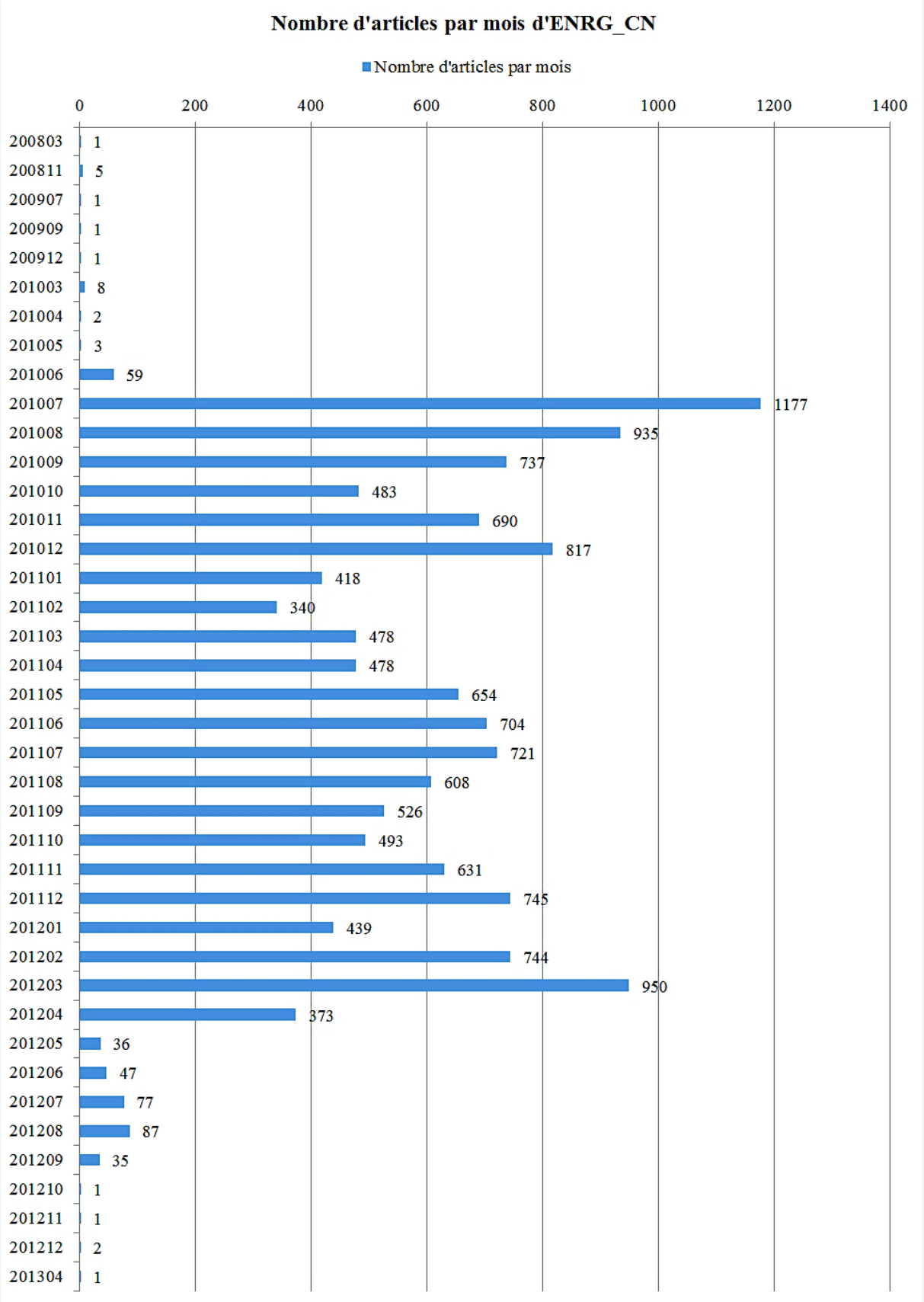


Figure 6.2 ENRG_CN de mars 2008 à avril 2013 : évolution du nombre d'articles pour les rubriques *Vert* et *Protection de l'environnement*

L'extraction automatique du sous-corpus ENRG_CN a en effet été réalisée entre les mois de mars 2008 et avril 2013 ; cependant l'analyse de la figure 6.2 met à jour la discontinuité de la répartition des articles dans ENRG_CN. Pour certains mois d'une année, par exemple du mois d'avril au mois d'octobre 2008 ou les mois d'août, d'octobre et de novembre 2009, nous constatons l'absence totale d'articles.

6.1.2 Sélection de périodes d'ENRG_CN

Après l'analyse des constats, nous écartons les années 2008, 2009 et 2013, au vu du nombre très faible d'occurrences par rapport au reste de la période retenue. Ces données, statistiquement insuffisantes et peu significatives, ne permettent pas d'analyser ces périodes et d'établir des liens entres les formes et l'ensemble des faits.
Désormais, nous nous intéressons uniquement aux années 2010, 2011 et 2012 pour ENRG_CN constituant le sous-corpus sélectionné.

6.2 ENRG_CN : données restreintes aux années 2010, 2011 et 2012

Pour la période retenue, nous allons successivement analyser la répartition mensuelle du nombre d'occurrences et du nombre de formes.

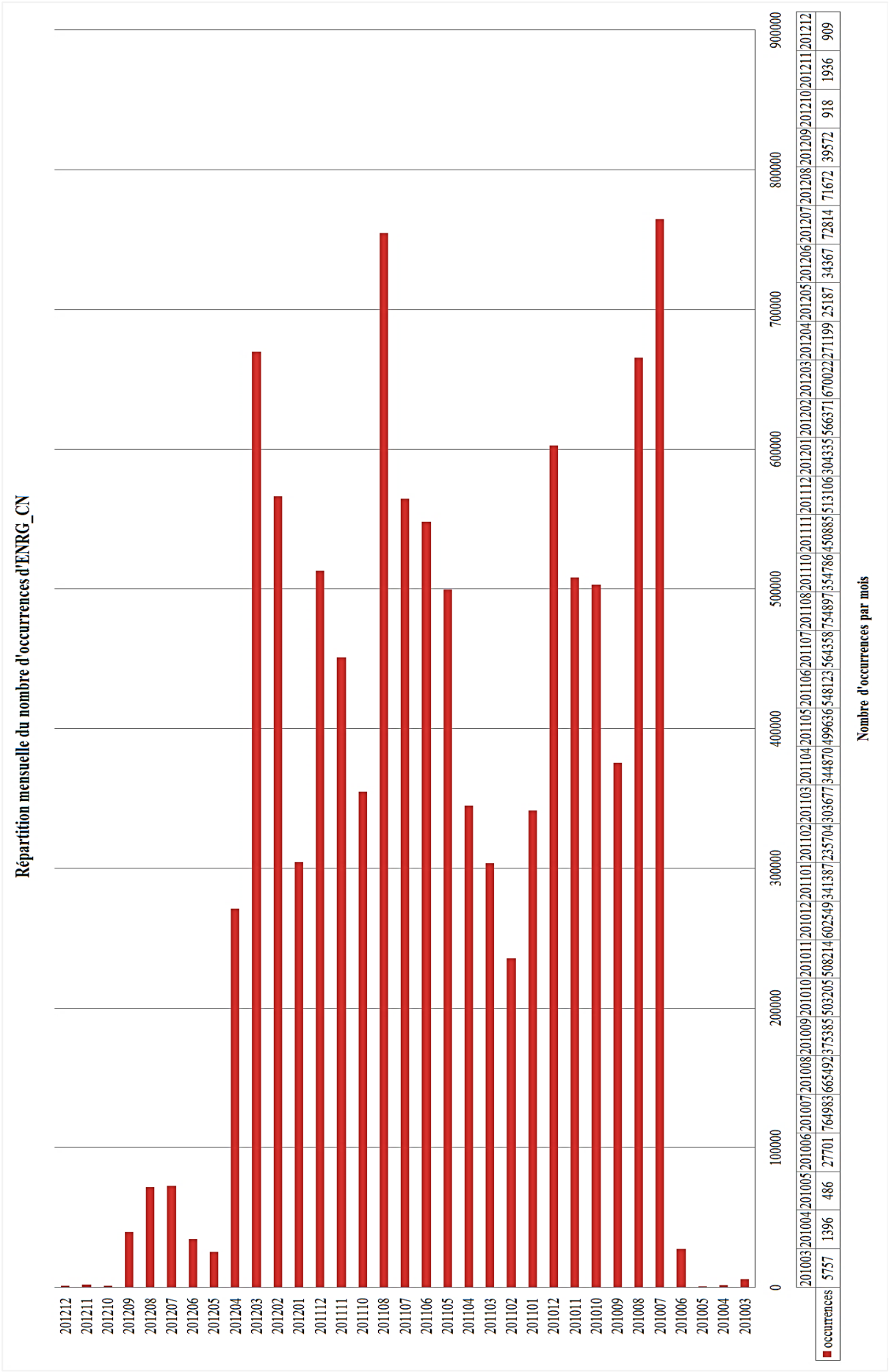


Figure 6.3 ENRG_CN de mars 2010 à décembre 2012 : répartition mensuelle du nombre d'occurrences

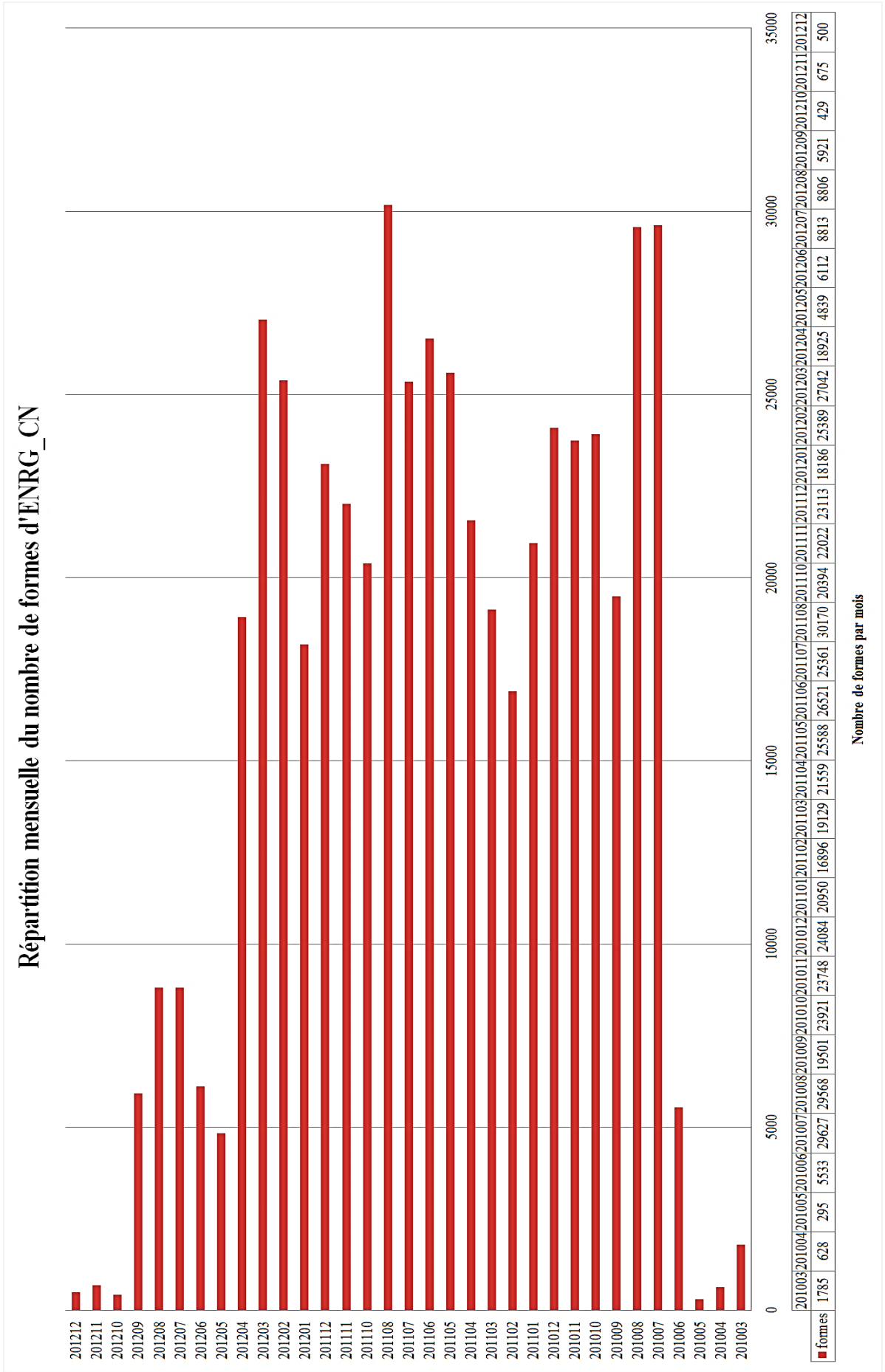


Figure 6.4 ENRG_CN de mars 2010 à décembre 2012 : répartition mensuelle du nombre de formes

D'après la figure 6.4 ci-dessus, le mois d'août 2011 atteint un nombre de formes maximal, un mois riche en événements. Mais selon les figures 6.1, 6.2 et 6.3, le record absolu de production d'articles a eu lieu au mois de juillet 2010 avec 1 177 articles. Ce constat nous interpelle. Nous faisons l'hypothèse que ce nombre record d'articles est dû à l'Exposition Universelle 2010 de Shanghai qui s'est déroulée du 1er mai 2010 au 31 octobre 2010. Le calcul de spécificités (seuil de probabilité à 5 et fréquence minimale à 10) du mois nous donnera éventuellement une réponse. Nous ferons une sélection de spécificités du mois de juillet 2010 dans le tableau qui suit.

Tableau 6.2 ENRG_CN : sélection des spécificités positives du mois de juillet 2010 et leurs traductions

Forme	Equivalent français	Frq. Tot.	Fréquence	Coeff.
馆	Pavillon (Exposition Universelle)	860	325	***
纳凉	Profiter de l'air frais (se protéger de la chaleur)	76	72	***
上杭	Shanghang (nom du siège de comté)	201	147	***
大连	Dalian	895	246	***
铜	Cuivre	728	268	***
世博会	Exposition Universelle	731	324	***
海洋	Océan	3764	543	***
矿业	Industries minières	2189	968	***
鱼	Poisson(s)	2596	456	***
事故	Accident(s)	4759	900	***
污水	Eaux polluées	5759	840	***
世博园	Parc de l'expo	217	121	***
上海	Shanghai	4622	672	***
泄漏	Fuite	1393	304	***
清	Clair, éclairer, nettoyer.	1720	340	***
紫金	Zijin (nom d'une société)	1814	996	***
爆炸	Explosion	535	161	***
汀江	Tingjiang (ville)	540	403	***
上杭县	Shanghang (siège du comté dans le Fujian)	423	299	***
紫金山	Zijinshang (nom d'un lieu ou d'une société)	420	286	***
铜矿	Mine de cuivre	490	288	***
福建省	Province du Fujian	476	159	***
世	Exposition Universelle	595	215	***
渗漏	Fuite (par pénétration)	473	235	***
油污	Huile lourde (pollution)	624	193	***
福建	Province du Fujian	753	187	*50*

Le tableau 6.2 des spécificités du mois de juillet 2010 nous livre des fragments d'informations concernant des événements locaux en Chine, événements non seulement liés à l'Exposition Universelle de Shanghai mais également à autres faits tels que :

1. Divers sujets écologiques et climatiques (formes en vert) liés à l'Exposition Universelle de Shanghai dont le thème principal est l'écologie.
2. Explosion de deux oléoducs (formes en rouge) servant au déchargement d'un tanker dans le port de Dalian, au nord-est du pays, a provoqué une marée noire, révélant ainsi la vulnérabilité des infrastructures chinoises.
3. Fuite (formes en gris) d'acide de cuivre (cuivre et acide nitrique), a pollué les eaux fluviales à Tingjiang dans la province du Fujian.

Nous poursuivons l'étude textométrique en analysant la courbe d'accroissement de vocabulaire et le diagramme de Pareto sur la période 2010, 2011 et 2012.

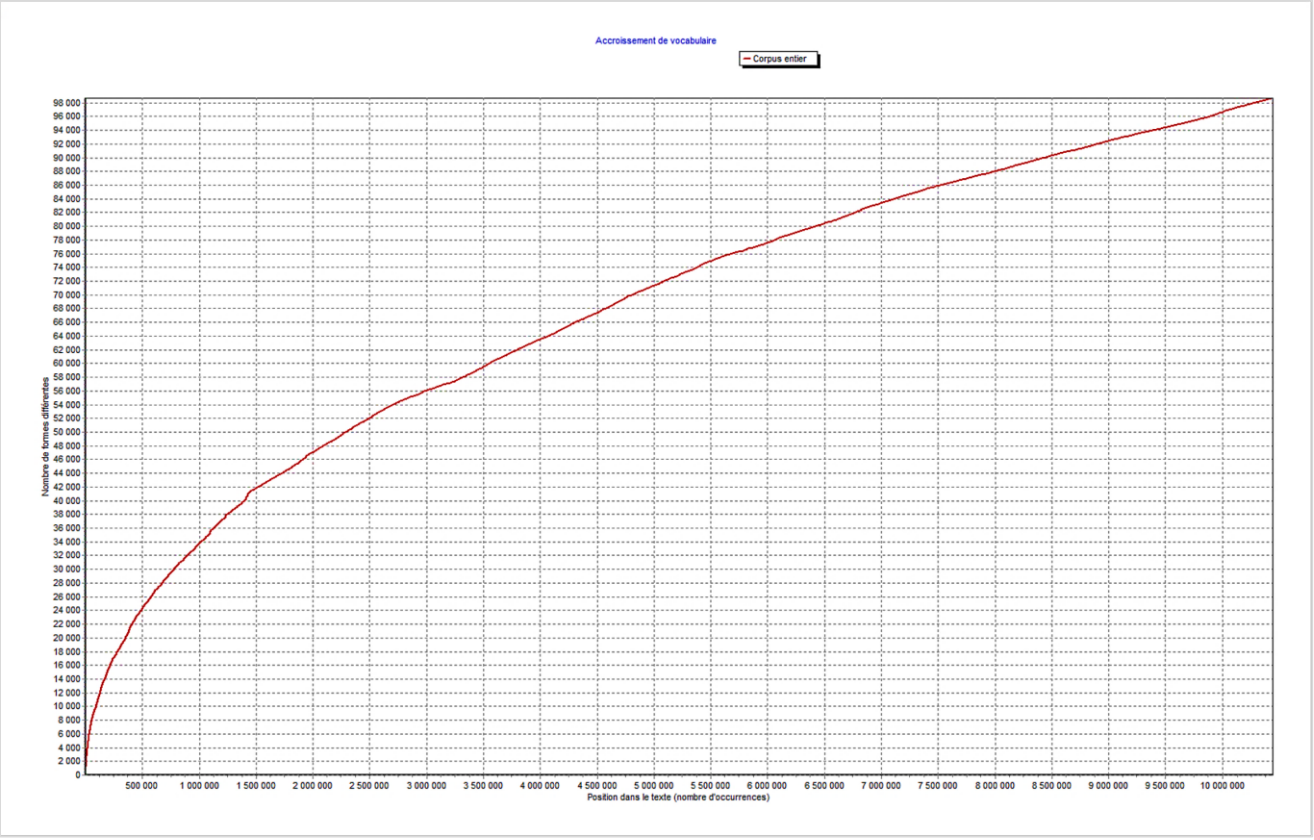


Figure 6.5 ENRG_CN de 2010 à 2012 : accroissement de vocabulaire

Tableau 6.3 ENRG_CN de 2010 à 2012 : principales caractéristiques textométriques

Nombre d'occurrences: 10442524 Nombre de formes: 98680
Nombre d'hapax: 33663 Fréquence maximale: 551192

	Num	Partie	Occurenc	Formes	Hapax	Fmax	Forme
✓	1	2010	3460989	59157	20139	188717	的
✓	2	2011	4919298	70074	23239	256106	的
✓	3	2012	2062237	45584	15251	106369	的

Le diagramme d'accroissement de vocabulaire (figure 6.5) et le tableau 6.3 montrent, comme pour les sous-corpus ENRG_FR et ENRG_US, l'apparition de nouvelles formes que nous détaillerons ci-après. L'ensemble compte 10 442 534 occurrences et 98 680 formes, le renouvellement de formes se caractérise par une croissance progressive en deux phases, phénomène normal[184] dans les calculs d'accroissement de vocabulaire :

1. après 2 000 000 d'occurrences, pour chaque tranche de 500 000 occurrences, une croissance de 4 000 formes,
2. par la suite, à partir de 7 000 000 d'occurrences, le nombre de formes augmente de 2 000 environ pour chaque tranche de 500 000 occurrences supplémentaires.

[184] Selon le même tutoriel de Lexico 3.

Tableau 6.4 ENRG_CN de 2010 à 2012 : principales caractéristiques textométriques par mois

Nombre d'occurrences: 10442524 Nombre de formes: 98680
Nombre d'hapax: 33663 Fréquence maximale: 551192

	Num	Partie	Occurenc	Formes	Hapax	Fmax	Forme
✓	1	201003	5765	1786	986	318	的
✓	2	201004	1398	629	401	76	的
✓	3	201005	489	296	215	15	的
✓	4	201006	27762	5535	2673	1412	的
✓	5	201007	766191	29629	10481	41063	的
✓	6	201008	666468	29570	10676	36216	的
✓	7	201009	539898	24188	8797	29311	的
✓	8	201010	340513	20225	7672	18546	的
✓	9	201011	509082	23750	8820	28224	的
✓	10	201012	603423	24086	8745	33536	的
✓	11	201101	341827	20952	7992	17746	的
✓	12	201102	236094	16898	6401	12573	的
✓	13	201103	304167	19131	7229	15897	的
✓	14	201104	345446	21561	8102	18165	的
✓	15	201105	500295	25590	9464	25437	的
✓	16	201106	548821	26525	9568	28475	的
✓	17	201107	565213	25363	8751	30213	的
✓	18	201108	756212	30173	10645	37966	的
✓	19	201110	355525	20398	7401	18308	的
✓	20	201111	451543	22024	7703	23765	的
✓	21	201112	514155	23117	8199	27561	的
✓	22	201201	304825	18188	6594	15423	的
✓	23	201202	567126	25391	9082	28832	的
✓	24	201203	671033	27046	9552	34827	的
✓	25	201204	271588	18927	7252	14665	的
✓	26	201205	25223	4840	2345	1300	的
✓	27	201206	34414	6113	2982	1732	的
✓	28	201207	72893	8815	3916	3591	的
✓	29	201208	71761	8808	3906	3598	的
✓	30	201209	39607	5922	2678	2156	的
✓	31	201210	919	430	298	55	的
✓	32	201211	1937	676	418	136	的
✓	33	201212	911	501	369	54	的

Le tableau 6.4, sur les 33 mois de la période sélectionnée, nous livre les caractéristiques principales du sous-corpus ENRG_CN, en particulier la production de formes, d'occurrences et d'hapax en fonction du temps. Par exemple, les mois de juillet, août 2010 et août 2011 sont les mois où les nombres d'hapax sont les plus importants.

6.2.1 Typologies sur 2010, 2011 et 2012

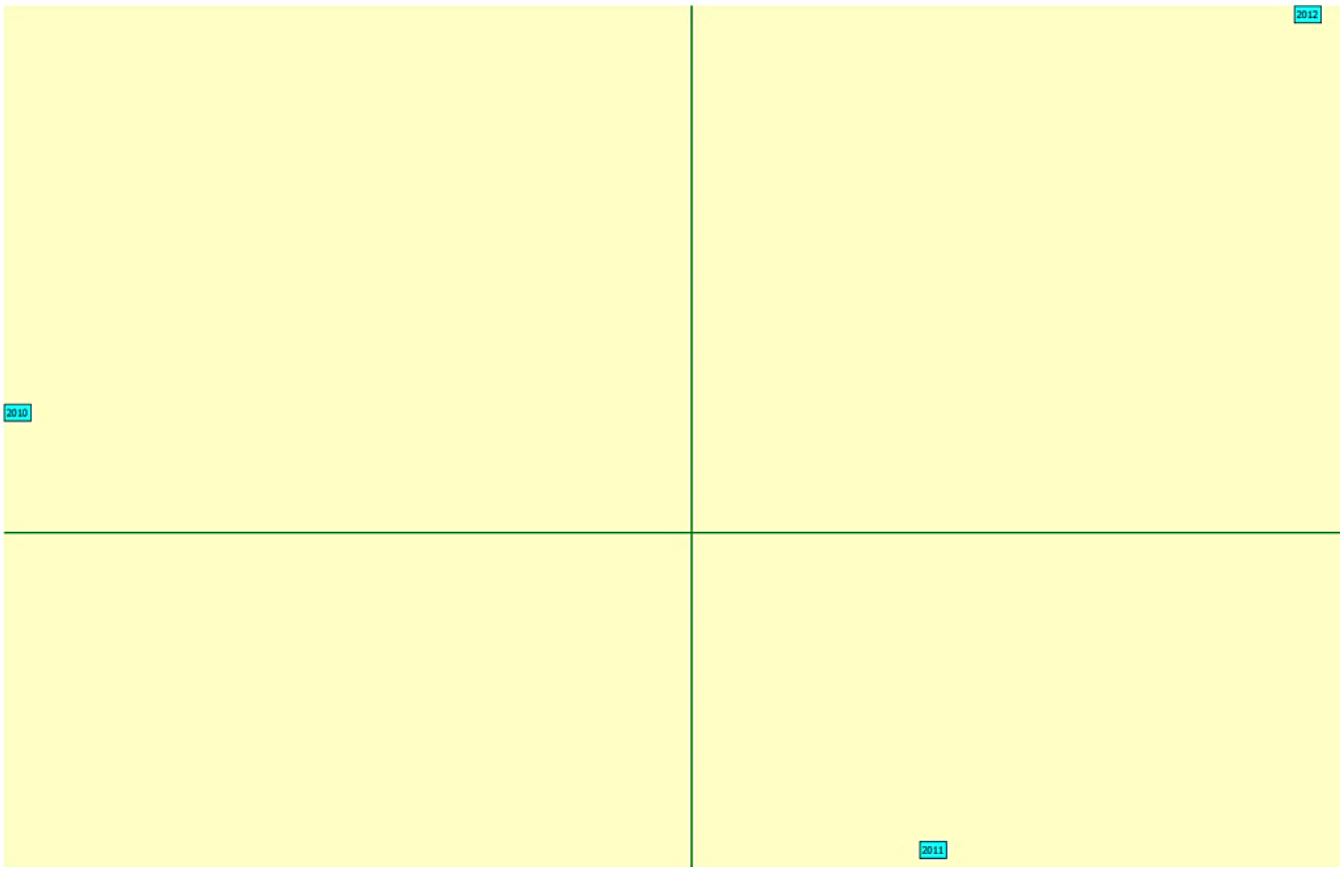


Figure 6.6 ENRG_CN de 2010 à 2012: analyse factorielle des correspondances

L'AFC sur les trois années montre qu'il n'y a pas de proximités annuelles textuelles. Mais ceci ne veut nullement dire qu'il n'y en a pas sur les mois. Afin d'étudier la proximité de ces textes en chinois, une AFC sur tous les mois a été effectuée avec la fréquence minimale à 80.

Sur la figure de l'AFC (figure 6.7, ci-dessous), les 33 mois sont répartis par leurs proximités textuelles dans quatre zones formées par les deux axes. Nous remarquons une nette séparation entre les mois se trouvant à gauche et à droite de l'axe vertical. Nous tentons de grouper les séries chronologiques et d'en restituer la répartition dans le tableau 6.5 ci-dessous.

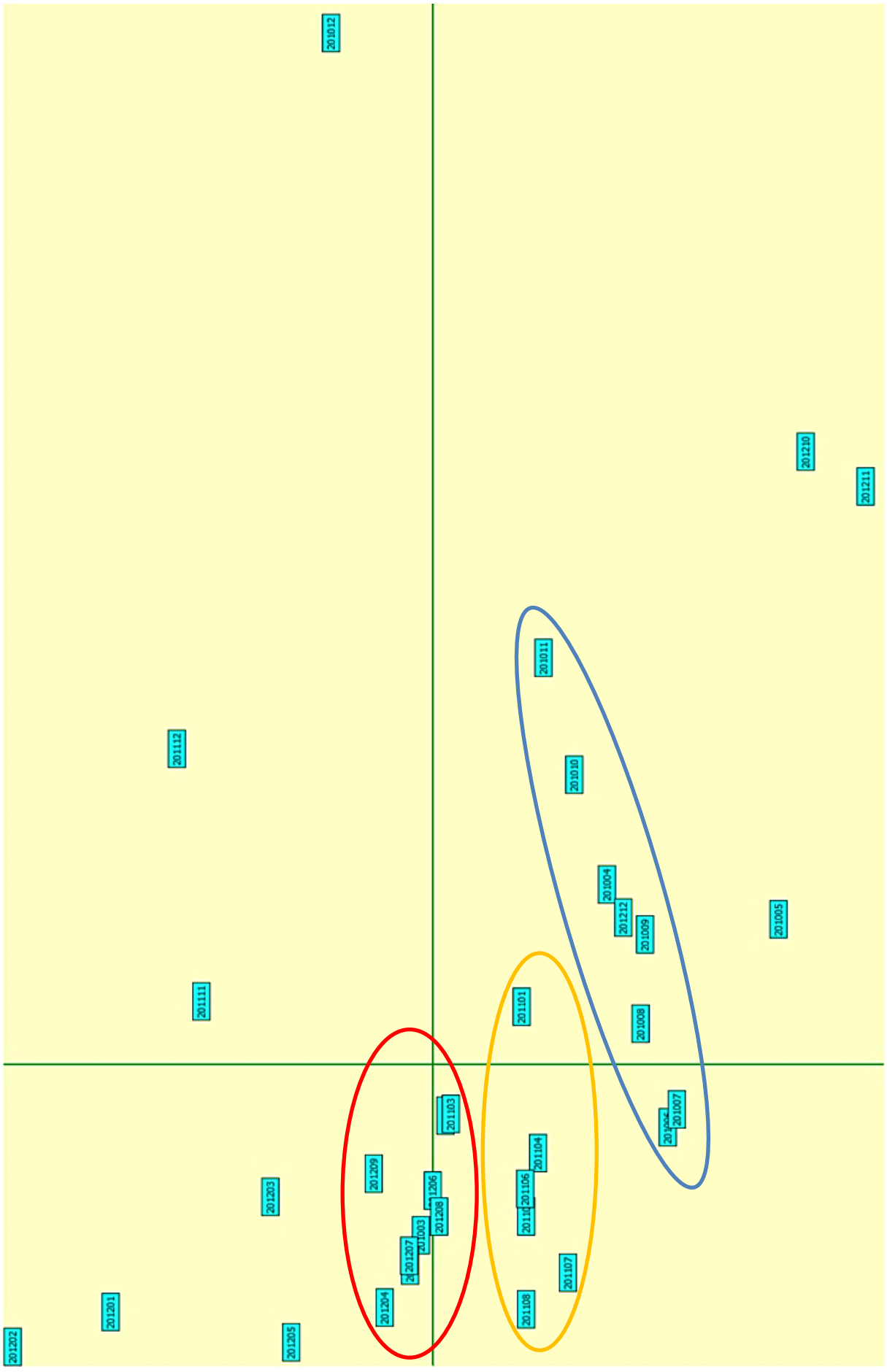


Figure 6.7 ENRG_CN de 2010 à 2012 : analyse factorielle des correspondances sur les mois des trois années

Tableau 6.5 ENRG_CN de 2010 à 2012 : restitution de la répartition des mois de l'AFC

en haut à gauche : 9 mois	**en haut à droite : 3 mois**
en 2010 2010-03 **en 2011** 2011-10 (caché par 2012-07) **en 2012** 2012-01, 2012-02, 2012-03, 2012-04, 2012-05, 2012-07, 2012-09	**en 2010** 2010-12 **en 2011** 2011-11, 2011-12
sur l'axe horizontal : 2 mois 2012-06 et 2012-08	
en bas à gauche : 9 mois	**en bas à droite : 10 mois**
en 2010 2010-06, 2010-07 **en 2011** 2011-02 (caché par 2011-03), 2011-03, 2011-04, 2011-05 (caché par 2011-06), 2011-06, 2011-07, 2011-08	**en 2010** 2010-04, 2010-05, 2010-08, 2010-09, 2010-10, 2010-11 **en 2011** 2011-01 **en 2012** 2012-10, 2012-11, 2012-12

Des fragments de séries chronologiques (surlignés en vert) dans le tableau, ont été mis en évidence dans les différents groupements de mois par l'AFC.

Nous pouvons regrouper les mois de ces trois années sans faire appel aux deux axes de l'AFC et nous obtenons les groupements et les séries chronologiques suivants par proximité :

Cercle rouge : 2010-03, 2011-02, 2011-03, 2011-10, 2012-04, 2012-06, 2012-07, 2012-08, 2012-09,

Cercle orange : 2011-01, 2011-04, 2011-05, 2011-06, 2011-07, 2011-08,

Cercle bleu : 2010-04, 2010-06, 2010-07, 2010-08, 2010-09, 2010-10, 2010-11, 2012-12.

Les séries chronologiques (colorées en vert ci-dessus) repérées dans les trois cercles nous livrent un autre clivage textuel à l'intérieur de ces trois années étudiées. Des sous-thèmes événementiels récurrents se manifestent à travers cette sous-structure chronologique et révélatrice, sous-structure formée par les trois cercles sur le plan de l'AFC.

Comme les trois années d'ENRG_CN sont complètement distinctes sur le plan de l'AFC obtenu par année (figure 6.6), un calcul de spécificités par année semble pertinent pour comprendre leurs différences en restituant les événements essentiels.

6.2.1.1 En 2010

Au vu du nombre d'articles en 2010, nous limitons la sélection des formes obtenues par le calcul de spécificités dont le coefficient est supérieur à 50 (marqué par trois étoiles *** dans les tableaux de sélection des spécificités positives). Une sélection plus large est nécessaire pour mettre en lumière quelques grands événements nationaux et internationaux qui ont marqué cette année.

Tableau 6.6 ENRG_CN 2010 : sélection des spécificités positives et leur traduction

Forme	Equivalent français	Frq. Tot.	Fréquence	Coeff.
减*	diminuer* ou moins*	13614	6878	***
墨西哥	Mexique	1016	895	***
会议	conférence	7234	4122	***
节能	économie d'énergie	13829	6583	***
世博园	parc de l'Exposition Universelle	217	198	***
生态	écologie	13295	5592	***
馆	pavillon	860	601	***
高温	température élevée	824	503	***
气候	Climat	14816	8620	***
排*	Emission*	15684	7627	***
空调	climatisation	1151	646	***
绿色	vert	14094	6148	***
上杭	Shanghang (district de la province du Fujian)	201	190	***
哥本哈根	Copenhague	1446	1163	***
公约	convention	1428	791	***
松花江	la rivière Songhua	373	265	***
汀江	Dingjiang (ville)	540	513	***
渗漏	fuite	473	318	***
上杭县	le district de Shanghang	423	387	***
铜矿	cuivre	490	372	***
变化	changement	10427	5606	***
紫金山	Zijinshan (nom de société)	420	365	***
世博会	Exposition Universelle	731	594	***
议定书	protocoles	3267	1854	***
限	limiter	3237	1609	***
环保	protection environnementale	28774	10823	***
税	taxes, impôt	3316	1626	***
产能	capacité de production	2798	1440	***
昆*	Cancun (kun)*	3601	3245	***
坎*	Cancun (kan)*	3561	3226	***
安平县	le comté Anping	134	132	***
矿业	exploitation minière	2189	1699	***
紫金	Zijin (nom de société)	1814	1634	***
低*	faible	16181	7805	***
十一五	onzième quinquennat	1942	1012	***
二氧化碳	dioxyde de carbone	2628	1294	***
碳*	carbone	22186	10901	***
京都	Kyoto	2609	1393	***
能耗	consommation d'énergie	2645	1345	***

Le tableau de spécificités de l'année 2010 (tableau 6.6) expose une sélection d'informations concernant des événements locaux en Chine et internationaux :

1. Participation de la Chine à la Conférence de Cancun sur le climat au Mexique décembre 2010 (formes en vert) et divers sujets écologiques et climatiques (formes en vert).
2. Divers sujets écologiques et climatiques (formes rouges) liés à l'Exposition Universelle de Shanghai (thème écologie) mai 2010.
3. Fuite (formes en gris) d'acide de cuivre (cuivre et acide nitrique) polluant les eaux fluviales à Tingjiang de la province du Fujian.
4. Explosion de deux oléoducs (formes roses) servant au déchargement d'un tanker dans le port de Dalian, au nord-est du pays, a provoqué une marée noire, révélant ainsi la vulnérabilité des infrastructures chinoises.
5. Programme politique mettant l'accent sur des mesures concrètes[185]liées à l'environnement et au développement durable demandées par le Parti Communiste chinois (la forme turquoise).

En élargissant le calcul de spécificités sur l'année 2010, nous retrouvons les événements locaux déjà évoqués sur le mois de juillet 2010 (points 2, 3, 4 de la liste ci-dessus). D'autres sujets apparaissent, d'ordre international (point 1) et d'ordre politico-économique (point 5).

[185] Se reporter à « De meilleures normes en Chine ? » - *HEC Eurasia Insitute*, *TOPIC*, mars 2008, http://www.hec.fr/var/fre/storage/original/application/0f91fd2d83663f0f5cd0d1fdf0931bec.pdf (consulté le 10/04/2015).

6.2.1.2 En 2011

Nous poursuivons l'analyse de notre sélection des spécificités positives sur l'année 2011.

Tableau 6.7 ENRG_CN 2011 : sélection des spécificités positives et leur traduction

Forme	Equivalent français	Frq. Tot.	Fréquence	Coeff.
辐射	radioactivité ou radiation	971	705	***
香精	arômes	324	291	***
焚烧	incinération	2937	1911	***
台湾	Taïwan	995	707	***
乳	lait	987	769	***
苹果	Apple ou pomme ou nom d'un journal à Taiwan	917	724	***
安全	sécurité	10718	5968	***
电场	champ électrique	1007	724	***
电	électrique	15939	8557	***
奶粉	lait en poudre	1147	832	***
防腐剂	conservateur	330	297	***
添加	additifs ou ajouter à	1070	758	***
食品	aliments	10433	7235	***
勾兑	frelater ou mélange	283	256	***
瘦肉	viande maigre	642	537	***
油	huile	8060	5305	***
馒头	brioches ou gâteaux cuits à la vapeur	288	261	***
豆浆	lait de soja	498	454	***
铅酸	plomb-acide	499	445	***
风	vent	10246	6735	***
猪	porc	708	598	***
康菲	ConocoPhillips[186] (nom d'une société américaine)	683	544	***
中海	CNOOC	749	563	***
猪肉	porc (viande de porc)	725	603	***
有机	organique	2211	1415	***
餐	repas	1376	968	***
牛肉	bœuf	381	349	***
地沟	caniveaux ou égouts	2037	1647	***
蓬莱	Penglai (ville au Shangdong)	362	307	***
渣	scories (résidus)	1418	977	***
厨	cuisine	2009	1416	***
酯	ester (composé produit par la réaction entre un acide et un alcool)	365	311	***
奶	lait	1351	1062	***
添加剂	additif	1578	1277	***
牛奶	lait de vache	692	518	*50*
苹果公司	Apple (Groupe)	346	295	*50*

[186] ConocoPhillips (NYSE : COP) est une entreprise américaine spécialisée dans l'extraction, le transport et la transformation du pétrole. Elle exploite aussi des réseaux de stations-service dans différents pays.

Le tableau de spécificités de l'année 2011 (tableau 6.7) présente des informations saillantes relatives aux événements locaux et internationaux :

1. Mobilisation générale de la détection de radioactivité en Chine suite à la catastrophe de Fukushima (en rose).
2. Scandales alimentaires en Chine[187] (en jaune) :
    - récupération d'huile distillée dans les caniveaux à la sortie des restaurants,
    - transformation du porc en bœuf afin de vendre la viande plus chère (le porc est passé au borax, un minerai qui lui donne l'aspect et le goût du bœuf, mais ce minerai est strictement interdit dans l'usage alimentaire, il entre dans la composition des détergents et des pesticides et est cancérigène[188]),
    - deux entreprises taiwanaises ont utilisé des plastifiants et des agents d'opacification, qui ont contaminé un grand nombre de boissons et d'aliments en Chine et à Taiwan.
3. Apple est accusé par des ONG d'utiliser des sous-traitants chinois qui polluent (en turquoise).
4. En juin, deux déversements d'hydrocarbures dans la mer de Bohai ont duré un certain nombre de jours. CNOOC[189] n'a pas réagi rapidement. L'affaire est devenue un scandale. L'Administration maritime de Chine a ordonné à l'exploitant américain (*ConocoPhillips*) de cesser leurs activités (en rouge).
5. Différentes pollutions aux métaux lourds dans les eaux fluviales contaminent l'agriculture et l'alimentation en Chine ; de plus, de grandes quantités de scories ont été générées dans les eaux lors des traitements.

Ces séries d'événements expliquent la raison pour laquelle la richesse du nombre de formes apparues est importante au mois d'août 2011.

6.2.1.3 En 2012

Nous terminons l'analyse des spécificités positives par l'année 2012, dernière année complète de notre sous-corpus.

[187] Se reporter à « De meilleures normes en Chine ? » - *HEC Eurasia Insitute*, *TOPIC*, mars 2008, http://www.hec.fr/var/fre/storage/original/application/0f91fd2d83663f0f5cd0d1fdf0931bec.pdf (consulté le 10/04/2015).
[188] Il est à noter que pour désigner une substance qui favorise l'apparition d'un cancer, on emploie le terme cancérogène, alors que pour parler d'une substance qui favorise le développement d'un cancer, on utilise plutôt le terme cancérigène.
[189] *China National Offshore Oil* Corporation (CNOOC) est la troisième compagnie pétrolière chinoise derrière *Sinopec* et *PétroChina*, son rôle est plus orienté vers l'exploitation de ressources pétrolières et gazières extérieures à la Chine, en coopération avec des entreprises étrangères.

Tableau 6.8 ENRG_CN 2012 : sélection des spécificités positives et leur traduction

Forme	Equivalent français	Frq. Tot.	Fréquence	Coeff.
中药	médecine chinoise traditionnelle	503	401	***
河池市	Hechi (ville dans le Guangxi)	247	238	***
鸡蛋	Œufs	513	364	***
望江县	Wangjiang (siège d'un comte)	94	94	***
质量	qualité	7916	2590	***
鄱阳湖	le lac Poyang	977	478	***
污染源	source de la pollution	951	421	***
药用	médicinal	251	193	***
胆	vésicule biliaire ou la bile (l'organe)	2308	1820	***
归*	rentrer ou retour (nom de marque)*	926	556	***
空气	air	6987	2167	***
安全	sécurité	10718	2771	***
湖泊	lac	1103	450	***
委员	membre du comité	1064	491	***
苯酚	phénol	101	92	***
药品	médicaments	1114	644	***
核电厂	centrale nucléaire	206	144	***
彭泽	Pengze (nom centrale)	205	193	***
药业	pharmaceutique	311	263	***
龙江	Longjiang (nom d'une rivière)	493	355	***
真*	zhen (marque pharmaceutique)*	1894	763	***
广西	Guangxi (province)	1036	513	***
核电	puissance nucléaire ou électricité nucléaire	3302	1052	***
水价	prix de l'eau	691	323	***
明胶	gélatine	606	603	***
归真堂*	Temple Guizhen (marque pharmaceutique)*	597	504	***
郴州	Chenzhou (ville)	163	127	***
超标	dépasser normes autorisées	3854	1354	***
堂*	Temple (marque pharmaceutique)*	663	454	***
水质	qualité de l'eau	3442	1133	***
PM2	PM 2.5 (Les particules en suspension)	4111	2073	***
熊	ours	3413	2538	***
江河	rivière(s)	787	454	***
航空	aviation	2487	831	***
引流	détournement d'une rivière (rediriger)	254	212	***
柳州	Liuzhou (ville dans le Guangxi)	372	341	***
胶囊	capsule	553	529	***
铅	plomb	2750	963	***
重金属	métaux lourds	2694	1061	***
黑熊	ours noir	754	482	***
大气	atmosphère	2635	859	***
保健食品	compléments alimentaires	172	146	***
饮用水	eau potable	1511	701	***
镉	Cadmium	1514	1048	***
沉降	précipitation (terme chimique) [190]	461	253	***
胆汁	bile (un liquide organique)	346	264	***
自来水	eau courante	1368	640	***
活	Vivant	1391	676	***
净化器	Filtre	85	80	*50*
保健品	compléments alimentaires	161	119	*50*
镇江市	Zhenjiang (ville)	114	96	*50*
价格	prix	6373	1740	*49*
欧盟	Union Européen	4382	1265	*48*

[190] La subsidence en géologie est un lent affaissement de la lithosphère entraînant un dépôt progressif de sédiments sous une profondeur d'eau constante.

Le tableau de spécificités de l'année 2012 (tableau 6.8) nous livre encore une fois les événements saillants de l'année, principalement des faits nationaux, mais nous voyons apparaître un fait international important aux yeux des chinois, *la taxe carbone* :

1. Pollution des eaux fluviales : une fuite de cadmium contamine des cours d'eau à Liuzhou ainsi que dans d'autres villes au Guangxi en Chine (formes en jaune).
2. Pollution de l'air avec la particule fine PM2.5 (formes en turquoise).
3. Pollution des eaux fluviales à Zhenjiang (Jiangsu), surdosage de phénol par l'usine de traitement des eaux (formes en gris foncé).
4. Scandale sur le prélèvement de la bile d'ours noir (pratique courante de la médecine traditionnelle) dans la société chinoise Guizhentang ou Gui Zhen Tang (formes en gris clair).
5. Capsules toxiques de médicaments : normes non respectées et produits frelatés (formes en vert clair) lors de leur production.
6. Scandale des faux-œufs fabriqués avec de la gélatine et de l'alun de potassium[191] (formes sans couleur).
7. Taxe carbone sur les émissions polluantes pour les compagnies aériennes, imposée par l'Union Européenne (formes en rose), un événement international saillant pour les chinois.
8. Arrêt de la construction d'une centrale nucléaire « Pengze » faisant suite à des études diligentées par les autorités compétentes et à des enquêtes d'utilité publique et de sécurité nucléaire, études et enquêtes peu approfondies, une première dans l'Histoire chinoise (formes en rouge). Nous pouvons en déduire que cet arrêt forme un signal faible.

Mise à part la sémantique informationnelle de ces formes chinoises, la lecture de ces tableaux 6.6 et 6.8 de spécificités chinoises nous montre immédiatement des erreurs de segmentation, phénomène très récurrent dans le traitement automatique des langues, problème impactant souvent le fondement du traitement automatique du chinois.

Par exemple, dans les tableaux 6.6 et 6.8, les formes indiquées par une étoile :

- 减/jiǎn 排 /pái : réduction des émissions (de CO2 ou de gaz à effet de serre),
- 低/dī 碳 /tàn : à densité carbonique faible,
- 坎/kǎn 恩/ēn : Cancun,
- 归/guī 真/zhēn 堂/táng : nom d'un groupe pharmaceutique, dans le tableau 6.8.

En effet, ces formes étoilées dans les deux tableaux ci-dessus ont été séparées par le segmenteur automatique (ICTCLAS). Le segmenteur les a considérées chacune comme un mot individuel, alors qu'elles auraient dû être reconnues comme un seul mot, autrement dit, une seule forme. Ainsi, le comptage des formes peut être biaisé par ces erreurs de segmentation. Toutefois, l'outil « segments répétés », permettant de repérer les séquences textuelles via leurs fréquences, pourrait partiellement ou parfois intégralement récupérer les informations véhiculées dans les formes non lexicalisées dans les dictionnaires. Par conséquent, les problèmes de la segmentation sont intrinsèquement liés à la veille informationnelle et à la textométrie du chinois.

La restitution d'informations à travers les trois tableaux précédents de spécificités annuelles d'ENRG_CN nous mène à un contexte chinois très particulier, scandales à répétition, dénonciations récurrentes des fraudes par les médias. Ceci prouve justement le contraire de l'absence de la liberté

[191] L'alun de potassium encore nommé sulfate double d'aluminium et de potassium, est un sel double de formule chimique KAl(SO4)2•12 H2O.

d'expressions en Chine, une idée reçue véhiculée chez les occidentaux. Le pouvoir central laisse publier certains types d'informations.

Comme en témoignent les études des ENRG_FR et ENRG_US dans les chapitres précédents, les informations sur les thèmes *énergies* et *environnement* émanant de la presse française sont consacrées pour l'essentiel au nucléaire, celles des Américains sont beaucoup plus disparates avec une absence de la forme *EPR*. Quant à celles des médias chinois, elles sont très dénonciatrices et révélatrices d'une volonté croissante et déterminée d'améliorer les conditions environnementales. Ces constats nous conduisent à focaliser nos veilles sur l'énergie nucléaire, plus particulièrement l'électricité nucléaire dans ENRG_CN, thème commun et comparable dans les trois sous-corpus.

A l'issue des toutes premières recherches sur le thème *énergie nucléaire*, nous nous sommes heurtés encore une fois aux problèmes de segmentation du chinois. En effet, la notion d'énergie se traduit par 能源 /néng yuán, et la production de l'électricité par 发电 /fā diàn, mais les deux formes ont été parfois séparées en quatre formes individuelles à savoir, énergie pour 能/néng, sources pour 源/yuán, produire pour 发/fā, et électricité pour 电/diàn. Par exemple, la forme 能 néng désigne à la fois le verbe pouvoir et tous les types d'énergies. Ces difficultés de segmentation ont pour conséquence de complexifier l'accès aux informations et à l'interprétation des résultats textométriques. Dans nos recherches chinoises, seules les formes correctement segmentées ont été retenues.

Quatre formes associées aux thèmes énergies 能源/néng yuán/énergie, 核能/hé néng/énergie nucléaire, 核电/hé diàn/électricité nucléaire et 核/hé/nucléaire ont été projetées sur la carte de ventilation.

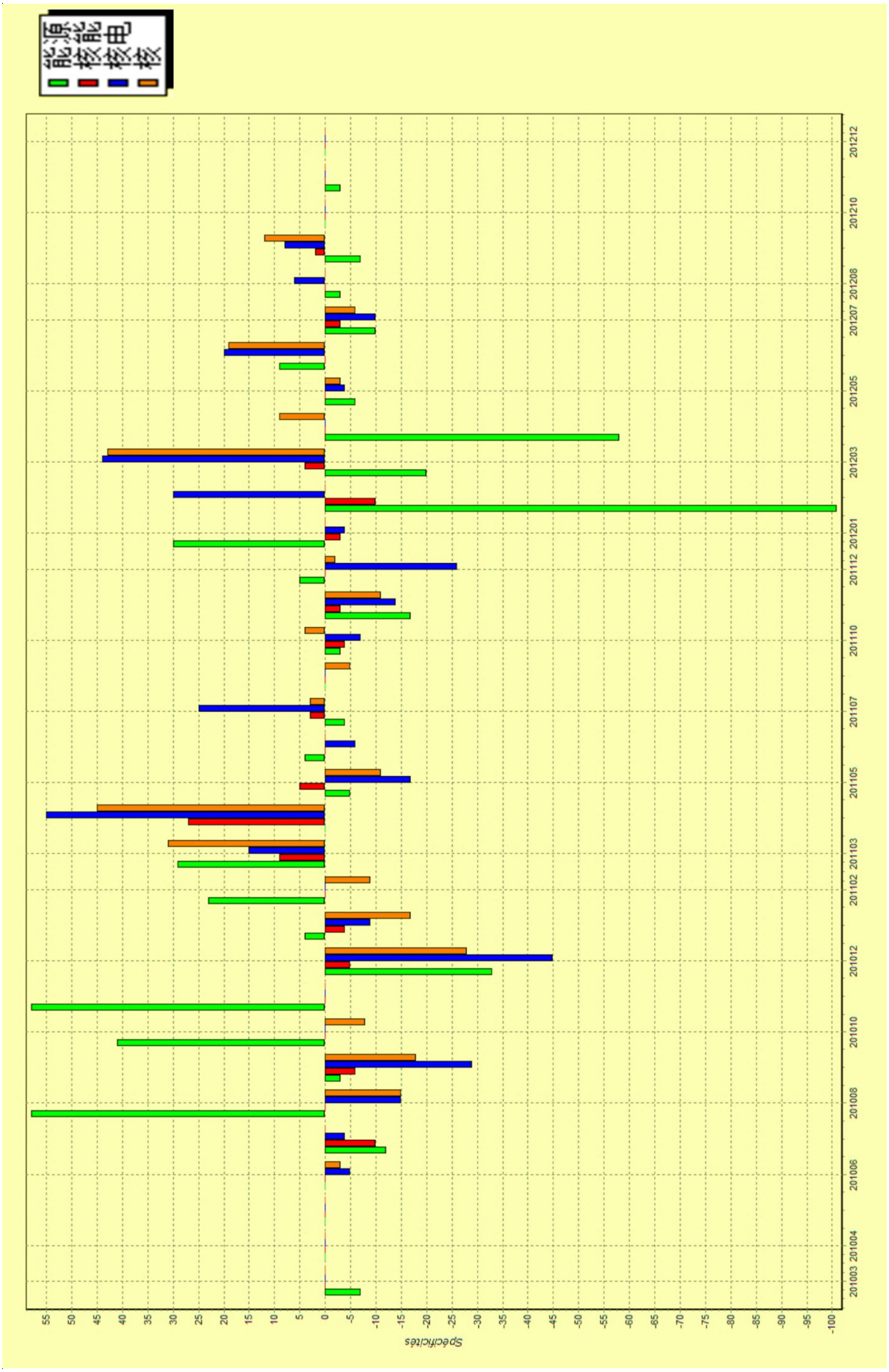


Figure 6.8 ENRG_CN de 2010 à 2012 : ventilation par mois des formes 能源/néng yuán /énergies en vert, 核能/hé néng/énergie nucléaire en rouge, 核电/hé diàn/électricité nucléaire en bleu et 核/hé/nucléaire en orange

La ventilation de ces formes dans ENRG_CN montre leur répartition et évolution au fil du temps. L'énergie a été le sujet saillant aux mois de juillet, septembre et novembre 2010, aux mois de janvier et mars 2011. L'Exposition de Shanghai dont le thème principal est : «Une ville meilleure, une vie meilleure», a été un des événements déclencheurs des discours sur l'énergie. L'énergie revient au cœur de l'actualité à la fin de décembre 2011, alors que l'énergie nucléaire n'est qu'un thème récurrent aux alentours de l'événement de Fukushima (mars et avril 2011), mais c'est bel et bien cette catastrophe qui a déclenché la focalisation des médias sur l'énergie nucléaire ainsi que sur l'électricité nucléaire. Quant au 核/hé/nucléaire, un terme générique, celui-ci est accompagné par sa forme dérivée, 核能/hé néng/l'énergie nucléaire, tout au long de l'évolution dans la période retenue avec probablement des discussions sur la sécurité et la sûreté du nucléaire en novembre 2011 et en avril 2012.

Figure 6.9 ENRG_CN de mars 2010 à septembre 2010 : extrait de la carte des sections contenant les formes de la figure 6.8

L'extrait de la carte des sections (figure 6.9) montre que les discours sur l'énergie ont été déclenchés au mois de juillet 2010 et se poursuivent de manière très dense. Nous avons retenu un article caractéristique dont l'extrait est présenté ci-dessous.

<day=20100727><media=东方早报><article=1066># 国 经 中心 : " 智能 能源 网 " 被 列为 重大 课题 # 早报 讯 国家 能源 局 总工程师 吴贵辉 26 日 在 上海 举办 的 上海 世博会 国家电网 馆 特别 活动日 主题 论坛 上 表示, 目前 正 在 编制 的 能源 行业 " 十二五 " 规划 , 已 将 发展 智能 电网 作为 重要 内容 纳入 其中 . 国家 能源 局 正 联合 有关 部委 组建 工作组 , 大力 推进 智能 电网 标准化 工作 . (……)

Traduction française

Le 27 juillet 2010, l'article du Quotidien du matin de l'Est (de la Chine) Le Centre national de l'économie annonce le projet Réseaux intelligents de l'énergie comme un enjeu majeur. Le 26 juillet , selon le Quotidien du matin, Monsieur Wu Guihui, ingénieur général en chef représentant le Conseil national de l'énergie a déclaré lors des événements spéciaux de la journée thématique du forum à l'Exposition Universelle de Shanghai, qui s'est tenue au Pavillon des Réseaux nationaux d'électricité de Chine : Le douzième plan quinquennal du secteur de l'énergie, qui est en cours de préparation actuellement, a classé le développement du réseau d'électricité intelligent comme un objectif primordial. Le Conseil national de l'énergie est en coordination avec les instances et ministères concernés afin de former des groupes de travail dans l'objectif de promouvoir et déployer la normalisation du réseau d'électricité intelligent. (……)

6.2.2 Cooccurrences évolutives autour des formes énergie(s) sur le sous-corpus chinois

Les calculs de cooccurrences évolutives ont été effectués à partir des formes 能源/néng yuán /énergie sur ENRG_CN afin d'obtenir un réseau de mots cooccurrents évolutifs (figures disponibles dans l'annexe N). Par la suite, nous sélectionnons les cooccurrents associés à *énergie, électricité et politique énergétique*, puis les synthétisons dans le tableau 6.9 ci-dessous. Des restitutions d'informations sélectionnées et centrées sur nos thèmes de recherche seront consacrées aux formes 能源/néng yuán/énergie, mais des retours aux textes des articles seront appliqués uniquement sur la forme *EPR*.

Tableau 6.9 ENRG_CN de 2010 à 2012 : extrait et synthèse des résultats des cooccurrences évolutives pour la forme 能源/néng yuán/énergie

2010			2011		2012	
Pôle	**Cooccurrents**	**Equivalent français**	**Cooccurrents**	**Equivalent français**	**Cooccurrents**	**Equivalent français**
能源 (énergie)	2020年	année 2020	2020	année 2020	2020年	année 2020
能源 (énergie)	生物	biologique	2020年	année 2020	生物	biologique
能源 (énergie)	碳	charbon	沼气	biogaz	燃煤	charbon
能源 (énergie)	煤炭	charbon, houille	生物	biologique	煤炭	charbon, houille
能源 (énergie)	煤	charbon ou houille	燃煤	charbon (combustible)	十二五	douzième quinquennat
能源 (énergie)	十二五	douzième quinquennat	煤炭	charbon et charbon de bois	核电	électricité nucléaire
能源 (énergie)	核电	électricité nucléaire	煤	charbon ou houille	电	Électrique
能源 (énergie)	电	électrique	十二五	douzième quinquennat	风能	énergie éolienne
能源 (énergie)	能	énergie	海上	en mer (offshore ou oversea)	核能	énergie nucléaire
能源 (énergie)	风能	énergie éolienne	风能	énergie éolienne	太阳能	énergie solaire
能源 (énergie)	太阳能	énergie solaire	风力	énergie éolienne ou force de vent	化石	fossile
能源 (énergie)	风力	force du vent (l'énergie éolienne)	核能	énergie nucléaire	天然气	gaz naturel
能源 (énergie)	化石	fossile	能	énergie ou verbe pouvoir	水电	hydroélectricité
能源 (énergie)	天然气	gaz naturel	太阳能	énergie solaire	光	lumière
能源 (énergie)	水能	énergie hydraulique	热能	énergie thermique	新型	nouveaux types
能源 (énergie)	水电	hydroélectricité	燃气	gaz combustible	秸秆	paille
能源 (énergie)	氢	hydrogène	煤层气 (煤矿瓦斯)	gaz de couche, gaz de houille (coalbed Methane ou CBM)	石油	pétrole
能源 (énergie)	光	lumière	天然气	gaz naturel	油气	pétrole et gaz
能源 (énergie)	石油	pétrole	地热	géothermie	发	produire
能源 (énergie)	发	produire	水电	hydroélectricité	发电	produire ou générer de l'électricité
能源 (énergie)	发电	produire ou générer de l'électricité	光	lumière	清洁	propre
能源 (énergie)	洁净	propre, pur ou purifier	十一五	onzième quinquennat	电网	réseau éléctrique
能源 (énergie)	太阳	soleil	石油	pétrole	页岩	schiste
能源 (énergie)	风	vent	光热	photothermal	太阳	soleil
能源 (énergie)			发电	produire ou générer de l'électricité	林业	sylviculture
能源 (énergie)			洁净	propre ou purifier	风	vent
能源 (énergie)			电力	puissance électrique		
能源 (énergie)			并网	regroupement du réseau (électrique)		
能源 (énergie)			电网	réseau électrique		
能源 (énergie)			国家电网	réseau national électrique		
能源 (énergie)			页岩	schiste		
能源 (énergie)			太阳	soleil		
能源 (énergie)			风	vent		

Légende
- sources ou types d'énergies
- réseau électrique
- politiques, plans et objectifs

Les formes communes des cooccurrents concernent d'abord la politique énergétique avec le douzième plan quinquennal (*douzième plan quinquennal*), qui en fixe les objectifs (*l'année 2020*). Pour connaitre davantage d'informations, se reporter à l'annexe G.

Depuis les années 2000, la Chine participe à des conférences internationales sur le climat et l'environnement et se trouve parfois contrainte de tenir des engagements sur la diminution des émissions de gaz à effet de serre (Zhou, Fridley et al, 2011). Par conséquent, la Chine, l'un des grands émetteurs de CO2, se voit dans l'obligation de rechercher de nouvelles énergies.

Les cooccurrents évolutifs sélectionnés dans ce tableau nous indiquent que la Chine diversifie ses sources d'énergies (Wu, Huang et Deng, 2009 ; Ye, Yang et al, 2010), *biologique, éolienne, solaire, propre, géothermie, schiste, biogaz, nouveaux types, sylviculture, etc.* et cherche constamment à produire ou à mieux produire de l'électricité (*réseau électrique intelligent*), en raison de sa croissance démographique et économique, car qui dit croissance démographique et économique, dit augmentation de la consommation d'énergies.

La seule apparition de la forme *en mer (offshore)* (Tan, Zhao et al, 2013) en 2011 indique que la Chine se tourne vers l'exploitation de l'énergie éolienne. Ainsi, un inventaire distributionnel de cette forme a été réalisé, tableau 6.10 ci-dessous.

Tableau 6.10 ENRG_CN de 2010 à 2012 : extrait de l'inventaire distributionnel trié après la forme 海上/hǎi shàng/en mer (offshore)

```
  4 ---- ---- ---- ---- ---- 海上 安全 sécurité en mer
  3 ---- ---- ---- ---- ---- 海上 采油 exploitation du pétrole en mer
 21 ---- ---- ---- ---- ---- 海上 的 en mer
    2 ---- ---- ---- ---- 海上 的 风 vent en mer
        2 ---- ---- 海上 电力 配套 措施 等 équipements et mesures en mer……
 758 ---- ---- ---- ---- ---- 海上 风 vent en mer
    6 ---- ---- ---- ---- 海上 风 场 parc éolien en mer
  720 ---- ---- ---- ---- 海上 风 电 électricité éolienne en mer
        5 ---- ---- 海上 风 电 3000 万 千瓦 électricité éolienne en mer 30 M KW
        10 ---- ---- 海上 风 电 500 万 千瓦 électricité éolienne en mer 5 M KW
       2 ---- ---- ---- 海上 风 电 安装 installation du parc éolien en mer
       8 ---- ---- ---- 海上 风 电 产业 industrie éolienne en mer
           2 ---- 海上 风 电 产业 整体 竞争力 compétitivité globale de l'industrie éolienne
(d'électricité) en mer
       2 ---- ---- ---- 海上 风 电 产业链 chaine industrielle des éoliennes en mer
         4 ---- ---- 海上 风 电 的 成本 coût d'électricité des éoliennes en mer
         2 ---- ---- 海上 风 电 的 大规模 éoliennes en mer à grande échelle
         3 ---- ---- 海上 风 电 的 电价 prix d'électricité des éoliennes en mer
           2 ---- 海上 风 电 的 资源 量 多 importantes sources d'électricité éoliennes en mer
      26 ---- ---- ---- 海上 风 电 发展 développement de l'électricité éolienne en mer
         2 ---- ---- 海上 风 电 发展 规划 programme de développement de l'électricité éolienne en
mer
       8 ---- ---- ---- 海上 风 电 工作 travail de l'électricité éolienne
         6 ---- ---- 海上 风 电 工作 座谈会 séminaires de travail de l'électricité éolienne
       5 ---- ---- ---- 海上 风 电 工程 projet de l'électricité éolienne en mer
      12 ---- ---- ---- 海上 风 电 规划 planification de l'électricité éolienne en mer
       7 ---- ---- ---- 海上 风 电 和 l'électricité éolienne en mer et……
           3 ---- 海上 风 电 和 清洁 煤 l'électricité éolienne en mer et charbon propre
      19 ---- ---- ---- 海上 风 电 机组 turbines de l'électricité éolienne en mer et……
```

L'inventaire réalisé hiérarchise les 4 catégories de mini-contextes (catégories colorées en gris commençant par 4, 3, 21 et 758) se trouvant à droite de cette forme 海上/hǎi shàng/en mer (offshore) et nous indique que l'*offshore* en Chine a le vent en poupe et que ce pays accentue

l'exploitation de son espace maritime (Tan, Zhao et al, 2013 ; Node-Langlois, 2015). Ces 4 catégories se déclinent elles-mêmes en sous-catégories. Il faut noter que les mini-contextes des 2 premières catégories n'ont pas été affichés pour des raisons de taille. En effet, la forme 海上/hǎi shàng/en mer (offshore) a été répétée 786 fois au total dans ce sous-corpus et uniquement en 2011, dont 4 fois associées à la *sécurité en mer*, 3 fois à l'*exploitation du pétrole en mer*, 21 fois à *en mer*, 758 fois au *vent en mer*. Dans la catégorie des 758 mini-contextes associée à la forme *vent en mer*, nous retrouvons, entre autres, *parc éolien en mer* 6 fois, *électricité éolienne en mer* 720 fois, etc. Ainsi, nous pouvons anticiper que des opportunités se créent dans le secteur de l'industrie des éoliennes (se reporter également à l'annexe G).

Deux autres formes 页岩/yè yán/*schiste* (uniquement en 2011 et 2012) et 林业/lín yè/*sylviculture* (uniquement en 2012) nous révèlent que, d'une part, le gaz de schiste suscite l'intérêt des leaders des grandes puissances économiques ainsi que ceux des pays de grande consommation d'énergies, d'autre part, à partir de 2012, la Chine se penche à nouveau sur les problèmes de l'afforestation de son territoire et cherche parallèlement à capter le gaz carbonique (CO2) en produisant de l'énergie biologique grâce à la sylviculture (tableau 6.11 ci-dessous). On remarque que les mini-contextes colorés en gris peuvent correspondre parfois à des titres ou à des citations.

Tableau 6.11 ENRG_CN de 2010 à 2012 : extrait de l'inventaire distributionnel trié après la forme 林业 /lín yè/sylviculture

55 ---- ---- ---- ---- **林业** 生 物质 matières biologiques issues de la sylviculture
48 ---- ---- ---- **林业** 生 物质 能源 énergies biologiques issues de la sylviculture
3 ---- ---- **林业** 生 物质 能源 产业 industrie de l'énergie biologique issue de la sylviculture
3 ---- **林业** 生 物质 能源 可 替代 2025 万 吨 标准煤 的 化石 énergie biologique issue de la sylviculture pourrait remplacer 20 250 000 tonnes de fossile en équivalent charbon[192].
3 ---- **林业** 生 物质 能源 所 占 比例 更 是 微乎其微 la proportion de l'énergie biologique de la syviculture est minime.
2 ---- **林业** 生 物质 能源 替代 1070 万 吨 标准煤 的 化石 能源 l'énergie biologique issue de la sylviculture pourrait remplacer 10 700 000 tonnes de fossile en équivalent charbon.
100 ---- ---- ---- ---- ---- **林业** 碳 carbone par la sylviculture…
97 ---- ---- ---- ---- **林业** 碳 汇 les puits de carbone (puits CO2) [193]par la sylviculture
2 ---- **林业** 碳 汇 的 持续 增加 accroissement stable des puits de CO2.
16 ---- ---- ---- **林业** 碳 汇 交易 commerce de puits de CO2
10 ---- ---- **林业** 碳 汇 交易 试点 projets de pilotage du commerce de puits de CO2

Parmi les informations les plus répétées du sous-corpus chinois, l'extrait du tableau des segments répétés (tableau 6.12, ci-dessous), les plus fréquents, témoigne d'une grande préoccupation de la Chine en matière de renouvellement énergétique, de protection environnementale et de développement durable.

[192] La *tonne équivalent charbon* est une unité de mesure de l'énergie (symbole tec), unité couramment utilisée dans l'industrie du charbon ; elle correspond à 1 tec = 8,141 MWh et 1 tec = 11 061 CHh.
http://www.universalis.fr/encyclopedie/tonne-d-equivalent-charbon/ (consulté le 26/08/2015)
[193] Un puits de carbone, *carbon sink* en anglais, est un réservoir qui capte et stocke le carbone atmosphérique.
http://www.futura-sciences.com/magazines/environnement/infos/dico/d/climatologie-puits-carbone-13132/
(consulté le 26/08/2015)

Tableau 6.12 ENRG_CN de 2010 à 2012 : extrait du tableau des segments répétés, les segments les plus fréquents du sous-corpus

Segments répétés	Equivalent français	Fréquence dans le sous-corpus
减 排	réduction des émissions	12530
低 碳	carbone à densité faible	10669
新 能源	nouvelles énergies	8754
气候 变化	changement climatique	8423
风 电	électricité éolienne	7680
光 伏	photovoltaïque	6109
环境 保护	protection de l'environnement	4881
节能 减排	économies d'énergie	4436
可 持续	durable	3791
坎 昆	Cancun	3565
新 能源 汽车	véhicules équipés de nouvelles énergies	3440
可 再生 能源	énergies renouvelables	3135
可 持续 发展	développement durable	2610

6.2.3 Les termes chinois du nucléaire

Si la séquence *énergie nucléaire* en français ou *nuclear energy* en anglais parait tout aussi simple et évidente par son aspect morphosyntaxique et sémantique, dans la langue chinoise, cette notion du nucléaire ne semble pas aussi limpide que nous le désirons.

En effet, en chinois l'énergie nucléaire se traduit par le mot suivant, 核能 /hé néng/énergie nucléaire, le premier caractère 核/hé signifie le noyau, par désignation spécifique au sens propre comme au sens figuré, il devient nucléaire ; le deuxième caractère 能/néng indique l'énergie ; l'ensemble des deux caractères forme le sens : l'énergie provenant des noyaux des atomes (核子 hé zi), autrement dit, l'énergie nucléaire. Le mécanisme de la formation des mots chinois veut que la forme 核/hé/nucléaire ou noyau(x) soit l'élément de composition lexicale ou la forme polylexicale de toutes les notions en rapport avec le nucléaire, telles que 核电站/hé diànzhàn/centrale nucléaire, 核 电 /hé diàn/électricité nucléaire, 核 试 验 /hé shìyàn/essai nucléaire, 核 武 器 /hé wǔ qì/armement nucléaire, 核导弹/hé dào dàn/missile nucléaire, 核潜艇/hé qián tǐng/sous-marin nucléaire, etc. La complexité est tout justement liée à cette forme polylexicale, car lorsque cette forme 核 (hé) prend le sens de noyau(x) de quelque chose, la compréhension du mot généré par cette racine peut être erronée. Par exemple, 核心/hé xīn/central, fondamental, 冰核/bīng hé/noyau de grêlon, 桃 核 /táo hé/noyau de pêche, 榄 核 /lǎn hé/noyau d'olive, 肺 结 核 /fèi jié hé/tuberculose, 核桃/hé táo/noix, etc. De surcroît, la forme 核 (hé) peut également être un verbe, comme 核查/hé chá/(re)vérifier ou examiner avec grand soin, ou 核准/hé zhǔn/approuver, etc. Donc, dans un contexte aussi imaginaire qu'il soit, il suffit qu'il y ait une simple erreur de segmentation dans toutes les possibilités énoncées ci-dessus pour que le sens du mot soit différent. *De facto*, les segmenteurs automatiques sont soumis constamment à des aléas. Dans un cas extrême, nous pouvons facilement construire un exemple avec les mots précédents afin de montrer la complexité

sémantique de tel contexte avec tels mots. Imaginons : 揽核核电站安全核查报告显示该地区冰雹的冰核近似核桃般大小。(Lǎn hé hédiànzhàn ānquán héchá bàogào xiǎnshì gāi dìqū bīngbáo de bīng hé jìnsì hétáo bān dàxiǎo), en français, cette phrase se traduit par : selon les rapports des contrôles de sécurité de la centrale nucléaire de Lanhe, la taille des grêlons de cette région est comparable à celle des noix. Les segmentations de cette phrase par Jieba, Hylanda et ICTCLAS sont respectivement les suivantes :

- 揽_核 _核电站 _安全 _核查 _报告 _显示 _该 _地区 _冰雹 _的 _冰核 _近似 _核桃 _般 _大小 (Jieba)
- 揽_核_核电站_安全_核查_报告_显示_该_地区_冰雹_的_冰_核_近似_核桃_般_大小 (Hylanda)
- 揽/核 核电站/ 安全/ 核查/ 报告/ 显示/ 该/ 地区/ 冰雹/ 的/ 冰/ 核/ 近似/ 核桃/ 般/ 大小/ (ICTCLAS)

Visiblement, le toponyme 揽核/Lànhé n'a pas été reconnu dans le dictionnaire des trois segmenteurs. A l'exception de Jieba, les deux autres segmenteurs n'ont pas pu reconnaitre le mot 冰核/bīng hé, mot très rare dans l'usage courant du chinois. D'autres mots relevant de ce même phénomène, mots non recensés dans les dictionnaires des segmenteurs sont à noter comme par exemple, 弃核/qì hé/ abandonner le nucléaire, etc. A la différence d'un programme informatique (algorithmes et dictionnaires), un locuteur natif chinois serait capable de rectifier et de distinguer la nuance ou la différence des sens au vu de la perception du contexte.

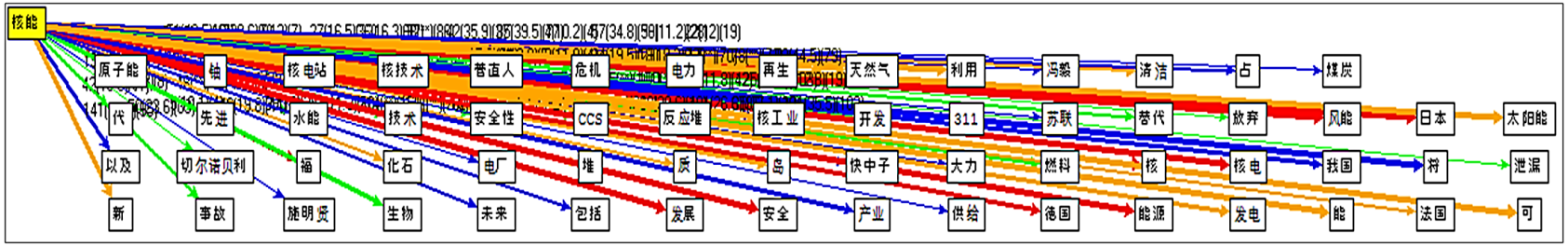
Cependant, si nous voulons brasser un maximum d'informations sur le nucléaire et sélectionner la forme nucléaire 核/hé comme forme-pôle, alors la séparation des caractères concernés tels que 冰核/bīng hé/noyau de grêlon peut parfois biaiser sensiblement les calculs de la textométrie en particulier ceux de cooccurrences et poly-cooccurrences, puisque la forme 核/hé/noyau, séparée du mot 冰核/bīng hé, constituera un point de connexion entre la forme-pôle 核/hé/nucléaire et d'autres chaînes de mots. Toutefois, dans les discours de nos corpus thématisés, ce genre de mots reste relativement peu présent, environ une dizaine de formes sont à exclure des résultats.

6.2.4 Cooccurrences des termes 核能/hé néng/énergie nucléaire et 核/hé/nucléaire

Guide de lecture des réseaux cooccurrentiels du Trameur :
Les réseaux cooccurrentiels sont constitués de traits colorés d'épaisseurs différentes correspondant aux nombres de contextes entre la forme-pôle et ses cooccurrents.
L'épaisseur du trait se décline en 4 niveaux en fonction du nombre de contexte :
1 <= niveau 1 <= 20 ; 21<= niveau 2 <=40 ; 41<= niveau 3 <= 60 ; 61 <= niveau 4.
Les couleurs dépendent de l'indice de spécificités attribué automatiquement :
10 <= bleu <= 12 ; 13<= vert <=24 ; 25<= orange <= 50 ; 51 <= rouge.

Les paramètres par défaut, co-freq : 2, seuil 10, contexte . ! ? ont été maintenus pour tous les calculs de cooccurrences. L'ordre d'affichage des cooccurrences est attribué par le logiciel Trameur.

Tableau 6.13 ENRG_CN de 2010 à 2012 : équivalent français des cooccurrents de la forme 核能 /hé néng/énergie



ENRG_CN 2010, 2011 et 2012 : cooccurrents de la forme 核能/hé néng/énergie nucléaire

原子能 énergie atomique	铀 uranium	核电站 centrale nucléaire	核技术 technologie nucléaire	普直人 Naoto Kan	危机 crise	电力 électricité	(可) 再生 renouvelable	天然气 gaz naturel	利用 utiliser ou profiter	冯毅 FENG Yi (secrétaire général assoc)	清洁 propre	占 occuper	煤炭 charbon ou houille		
代 génération	先进 avancé	水能 énergie de l'eau	技术 technologie	安全性 sécurité	CCS *Carbon Capture and Storage*	反应堆 réacteur	核工业 industrie nucléaire	开发 exploitation	311 (le 11 mars 2011)	苏联 URSS	替代 substitution	放弃 abandonner	风能 énergie éolienne	日本 Japon	太阳能 énergie solaire
以及 et	切尔诺贝利 Tchernobyl	福 (岛) Fukushima	化石 fossile	电厂 centrale ou usine électrique	(反应) 堆 réacteur	(物) 质 matière	(福) 岛 Fukushima	快中子(反应堆 *Fast breeder reactor*) surgénération ou surrégénération	大力 pleinement ou grands moyens	燃科 (燃科组成) fioul, mazout ou fuel	核 nucléaire ou noyau	核电 électricité nucléaire	我国 Chine	将 vouloir ou auxiliaire du futur	泄漏 fuite
新 nouvelle nouveaux	事故 accident	施明贤 (Michael Schaefer)	生物 biologie	未来 futur	包括 inclure	发展 développement	安全 sûreté	产业 industrie	供给 alimentation	德国 Allemagne	能源 énergie	发电 produire de l'électricité	能 énergie ou verbe pouvoir	法国 France	可 pouvoir

6.2.5 Trois réseaux cooccurrentiels

En examinant le tableau 6.13 ci-dessus sur les formes cooccurrentes du mot pôle 核能/hé néng/ énergie nucléaire, nous pouvons les regrouper en quatre grandes catégories grâce à leur appartenance sémantique :

1. différents types et sources d'énergies et développement durable : énergie atomique, uranium, énergie de l'eau, fossile, biologie, renouvelable, nouvelle, propre, substitution, gaz naturel, fuel (mazout), charbon (houille), énergie éolienne, énergie solaire, CCS,
2. production d'électricité : électricité, électricité nucléaire, produire de l'électricité, centrale d'électricité, alimentation, développement,
3. activités nucléaires : centrale nucléaire, technologie nucléaire, sécurité, futur, génération, réacteur, sûreté, industrie nucléaire, surgénération, nucléaire, Allemagne, France,
4. accidents nucléaires : accidents, Tchernobyl, URSS, Fukushima, Naoto Kan, 11 mars, Japon, crise, fuite.

Ces quatre catégories se traduisent par une convergence de sens de ces formes associées aux inquiétudes et aux enjeux majeurs de la Chine en plein essor démographique et économique, à savoir, produire de l'électricité par tous les moyens. Cet objectif est confirmé par l'extrait d'un article du Quotidien du matin de l'Est, pointé par une flèche bleue et illustré dans la figure 6.9 ci-dessus, dont la traduction se trouve en dessous de cette figure. Or, selon les résultats poly-cooccurrentiels issus d'ENRG_FR, la notion de production d'électricité n'a pas été mise en exergue, contrairement à ceux d'ENRG_US où la puissance de production d'électricité est toujours une grande préoccupation énergétique. Ces analyses des cooccurrents et poly-cooccurrents montrent que, en dépit de la catastrophe de Fukushima, il y a une similitude de besoins énergétiques de la part de la Chine et des Etats-Unis (*cf.* chapitre 2 et annexe G), deux grands producteurs d'énergie, mais aussi deux grands pollueurs de la planète, tandis que la France cherche de manière velléitaire à se débarrasser de sa dépendance nucléaire.

Mais dans le contexte de la mondialisation, la Chine cherche davantage à produire de l'électricité. Il faut que cette électricité soit abondante, propre et issue de sources d'énergie durables. Le tableau 6.12 plus haut, présentant les segments les plus répétés, montre que les énergies éolienne et photovoltaïque sont les types d'énergie les plus répétés dans ENRG_CN. Le tableau 6.14 ci-dessous illustre la hiérarchie des mini-contextes du segment répété (光) 伏 发电 /guāngfú fādiàn/produire de l'électricité photovoltaïque, segment le plus répété parmi ceux contenant la forme 发电/fādiàn/ produire de l'électricité. Il est à noter que la forme 光 伏 /guāngfú/photovoltaïque, étant un terme technique et scientifique, n'a pas été correctement segmentée par le segmenteur ICTCLAS, d'où l'ajout manuel de la parenthèse pour le caractère 光/guāng/lumière devant chaque segment. Ceci souligne l'enjeu majeur de la segmentation du chinois dans les traitements de veille.

Tableau 6.14 ENRG_CN de 2010 à 2012 : inventaire distributionnel complet trié après le segment répété (光)伏 发电 /guāngfú fādiàn/produire de l'électricité photovoltaïque

```
831 ---- ---- ---- ---- ---- (光)伏 发电 produire de l'électricité photovoltaïque(…)
   5 ---- ---- ---- ---- (光)伏 发电 安装 installation pour produire de l'électricité photovoltaïque(…)
      4 ---- ---- ---- (光)伏 发电 安装 量 quantité d'installations pour produire de l'électricité photovoltaïque(…)
   3 ---- ---- ---- ---- (光)伏 发电 才produire de l'électricité photovoltaïque afin de(…)
   13 ---- ---- ---- ---- (光)伏 发电 产业 industrie de production d'électricité photovoltaïque(…)
      2 ---- ---- ---- (光)伏 发电 产业 的 ce qui concerne l'industrie de production d'électricité photovoltaïque(…)
   2 ---- ---- ---- ---- (光)伏 发电 达到production de l'électricité photovoltaïque atteint (…)
   63 ---- ---- ---- ---- (光)伏 发电 的ce qui concerne la production de l'électricité photovoltaïque(…)
      8 ---- ---- ---- (光)伏 发电 的 补贴 subventions et aides pour la production d'électricité photovoltaïque(…)
        5 ---- ---- (光)伏 发电 的 补贴 政策 politique de subventions et aides de la production d'électricité photovoltaïque
(…)
   4 ---- ---- ---- ---- (光)伏 发电 发展 développement de la production d'électricité photovoltaïque(…)
      2 ---- ---- ---- (光)伏 发电 发展 有développement de la production d'électricité photovoltaïque peut avoir(…)
   14 ---- ---- ---- ---- (光)伏 发电 技术 technologie de la production d'électricité photovoltaïque(…)
   6 ---- ---- ---- ---- (光)伏 发电 企业 entreprises de production d'électricité photovoltaïque(…)
   27 ---- ---- ---- ---- (光)伏 发电 系统 système de production d'électricité photovoltaïque(…)
   8 ---- ---- ---- ---- (光)伏 发电 相比 en comparant avec la production d'électricité photovoltaïque(…)
   90 ---- ---- ---- ---- (光)伏 发电 项目 projet de production d'électricité photovoltaïque(…)
   8 ---- ---- ---- ---- (光)伏 发电 行业 secteur de production d'électricité photovoltaïque(…)
      2 ---- ---- ---- (光)伏 发电 行业 的 ce qui concerne le secteur de production d'électricité photovoltaïque
   3 ---- ---- ---- ---- (光)伏 发电 有de la production d'électricité photovoltaïque(…)
      2 ---- ---- ---- (光)伏 发电 有 了 la production d'électricité photovoltaïque est en train de(…)
   18 ---- ---- ---- ---- (光)伏 发电 装机容量 la puissance de production d'électricité photovoltaïque(…)
```

Cet inventaire distributionnel montre que les Chinois tout comme les Américains s'orientent vers l'énergie solaire.

Cooccurrences de la forme-pôle 核/hé/nucléaire[194]

Comme expliqué plus haut, les quelques formes colorées en jaune dans le tableau 6.15 ci-dessous sont générées par des erreurs de segmentation, formes à exclure du réseau cooccurrentiel. Le réseau cooccurrentiel de la forme 核/hé/nucléaire est beaucoup plus important que celui de la forme 核能 /hé néng/énergie nucléaire par son nombre de cooccurrents.

Le tableau 6.15 ci-dessous fournit un réseau de formes permettant d'accéder aux informations liées à la perception de la notion du nucléaire en Chine. Ainsi cinq catégories se dessinent en se focalisant sur tous les types de réacteurs et les constructeurs associés.

1. Toponymes liés aux sites nucléaires chinois : Daya Wan, Ling-Ao, Yangjiang, Taishan.
2. Types de réacteurs et constructeurs : ACPR1000, CP1000, AP1000, CNP1000, CPR1000, ACPR1000, trois, troisième génération, ITER, réacteur, Areva, Westinghouse.
3. Catastrophes et conséquences : tsunami, fuite, Tchernobyl, URSS, année 1986, Fukushima, le 11 mars, mars, explosion, radiation, radioactivités, construction, couvercle de confinement, iode, après, prolifération (propagation).
4. Activités nucléaires : industrie nucléaire, centrale, énergie atomique, mine d'uranium, combustible nucléaire, technologie nucléaire, surveillance, tritium, protection, césium, uranium, deutérium, isotope, sécurité, IAEA, fusion, bébé de sexe masculin, système de production d'électricité nucléaire.
5. Producteurs d'électricité chinois : CNGP, holding, Datang, Huaneng, Guotou (formes séparées en deux caractères, colorées en turquoise).

Nous constatons l'absence de la forme *EPR* alors que celle-ci est présente dans ENRG_FR. Cependant, son concepteur *Areva* est mentionné.

[194] La figure du réseau cooccurrentiel du tableau 6.14 est disponible dans l'annexe N, Figure N.15.

Tableau 6.15 ENRG CN de 2010 à 2012 : cooccurrents de la forme 核/hé/nucléaire

原子能 énergie atomique	ACPR1000	沙湾镇 Shawan comté	大石 (大石街) Dashi (comté du district Pānyú)	安全局 Agence nationale de la sécurité (NSA)	编制 système	释放 libérer	爆炸 explosion	十万分之一 un sur dix mille	检查 inspection	素 élément	中 Chine ou milieu ou pendant	中国 Chine	国际 international		
反应堆 réacteur	辐射 radiation ou radioactif	核燃料 combustible nucléaire	人工 artificiel	广东核电集团 CGNP	兆瓦 mégawatt(s) MW	第三 troisième	核废料 déchet nucléaire	凝结 coagulation	核电厂 centrale nucléaire	海峡 détroit	Année 1986	碘 iode	311 (11 mars, date de Fukushima)	半衰期 La demi-vie (d'une substance)	玛雅科 Complexe nucléaire Maïak (Mayak) en Russie
审视 examiner	铀矿 mine d'uranium	technologie nucléaire	防护 protection	核反应堆 réacteur nucléaire	氘 deutérium	放弃 abandonner	三 trois	核武器 armement nucléaire	法国 France	核工程 projet nucléaire	民用 civile	电力 électricité	锆 (gào) zirconium	prolifération ou propagation	影响 affecter ou Influence
海啸 tsunami	局 bureau, office	监测 surveillance	设计 conception	监管 contrôle	花生油 huile d'arachide	项目 Projet	之后 après	伊朗 Iran	地震 séisme	大唐 Datang (producteur d'électricité)	番禺区 Pānyú district	与 avec	上市 coté à la bourse	周贤生 Zhou Xiansheng (homme)	西屋 Westinghouse (US - AP1000)
技 (术) technologie	堆 réacteur	能源 énergie	国 pays	启(动) démarrer	中长期 à long terme	核政策 politique nucléaire	引发 suicider	彭泽 Pengze (comté, site candidat pour centrale nucléaire)	榄 (核) Lǎn Hé (comté)	证 certificat	暂停 suspendre	微量 Oligo- (peu nombreux)	安全性 sécurité	放射性 radioactivité	强震 choc violent
核工业 Industrie nucléaire	机组 unité de	机构 établissement	投 investissement	集团 groupe	技术 technique	大国 grand(s) pays	核辐射 radioactivité nucléaire	我国 la Chine	发展 développement	应急 urgence	水平 niveau	泄漏 fuite	受 à cause de	内陆 intérieur du continent	弃 abandonner
长 centrale ou usine	(阿海) 珐 Areva	后 après	危机 crise	国投 investissement national	建造 construire	切尔诺贝利 Tchernobyl	苏联 URSS	CNP1000	张国宝 Zhang Guobao (ministre énergies)	IAEA	代 génération	国家 pays ou national	冰粒 grésil	在建 en cours de construction	铯 césium
计量 dose	大亚湾 Dayawan Bay (ville)	院 institut	ITER	同位素 isotope	核电机组 système de production de l'électricité nucléaire	AP1000	阳江 Yangjiang (ville)	台山 Taishan (ville)	岭澳 ling-ao (ville)	运行 opération	设施 équipement	发生 avoir lieu	审批 approbation	铀 uranium	安全壳 couvercle de confinement
德国 Allemagne	男婴 bébé de sexe masculin	氚 tritium	CER (Certified Emission Reduction)	核能 énergie nucléaire	决定 décider	控股 holding	CP1000	聚变 fusion	CPR1000	华能 Huaneng groupe (producteur d'électricité)	131	废水 eaux usées	国务院 conseil d'état	部 département	常务 comité permanent
泄漏 fuite	规划 planification	刘华 Liuhua (ingénieur en Chef)	自主 autonome	委员会 comité	三月 mars	郁祖盛 Yu Zusheng (ministre env)	贺禹 Hè Yù (PDG de CGN)								

Afin de mieux cerner le milieu chinois, nous allons de manière non-exhaustive, expliquer les significations et le contexte des cooccurrents de la forme-pôle 核/hé/nucléaire en un seul caractère chinois. A la différence des analyses concernant ENRG_FR et ENRG_US, ENRG_CN présente un grand nombre de cooccurrents se rapportant à toute la branche nucléaire et à son environnement. Nous en citons quelques exemples ci-après.

Les toponymes

Nous trouvons dans le réseau cooccurrentiel de nombreux toponymes tels que *Taishan (EPR)*, *Dayawan*, *Yangjiang*, *Ling-ao*, lieux associés à des sites de centrales nucléaires chinoises, tous situés dans la province du Guangdong (se reporter à l'annexe G) et *Mayak*, un complexe nucléaire russe. Cependant, des noms de pays touchant de près le nucléaire pour diverses raisons apparaissent comme par exemple, l'*Iran* dont les relations diplomatiques avec les autres pays ont connu de graves tensions au sujet de la politique nucléaire menée par l'Iran après le 11 septembre 2001. La communauté internationale soupçonnait l'Iran de vouloir se doter de l'arme nucléaire et de mettre au point des armes de destruction massive, et s'inquiétait, craintes confirmées par l'AIEA. Les négociations ont été longues pour aboutir à un compromis à l'été 2015. Quant à la *Chine*, celle-ci développe un ambitieux programme nucléaire (se reporter à l'annexe G), contrairement à l'*Allemagne* qui, après Fukushima, a décidé de sortir du nucléaire. Nous retrouvons bien sûr la *France*, le pays du nucléaire par excellence.

Les catastrophes

Les catastrophes sont mentionnées comme *Fukushima* portant le numéro *311* qui représente la date de la catastrophe de Fukushima le 11 mars, à laquelle sont associés les cooccurrents suivants : *tsunami, sécurité, surveillance, radiation, fuite, explosion, mars,* ou encore, *Tchernobyl*, une autre catastrophe nucléaire qui a eu lieu en 1986, avec ses cooccurrents associés : *URSS, année 1986*, *radiation, radioactivité.*

L'activité nucléaire

L'association cooccurrentielle de cette activité est représentée par les formes suivantes : *Industrie nucléaire, énergie atomique, centrale, technologie nucléaire*.

Toutes les catastrophes nucléaires ont eu des conséquences touchant à de nombreux domaines d'activités tels que *Radiation, surveillance, protection, sécurité, radioactivité, IAEA, couverture de confinement.*

Sur le plan de la *sécurité,* l'accident de Fukushima a marqué les esprits et le gouvernement chinois a réagi face à cette catastrophe en prenant certaines mesures au niveau de la sûreté ou de la sécurité. Des accords ont été signés sur la base de partenariat pour des partages d'expériences et de savoir-faire comme par exemple en janvier 2015 entre EDF et CGN ou encore entre EDF et Huadian, un des premiers électriciens chinois, « *(...) afin de partager leur retour d'expérience sur l'exploitation et l'ingénierie des parcs nucléaires existants et pour maintenir les plus hauts niveaux de sûreté (...) accords signés avec nos partenaires chinois historiques viennent approfondir des coopérations existantes et poser les bases de nouveaux projets communs (...) »* a déclaré le Président-Directeur Général d'EDF Jean-Bernard Levy[195].

[195] https://www.lenergieenquestions.fr/tag/chine/ (consulté le 1/03/2015).

Le site de production de l'énergie nucléaire est souvent représenté par la forme ***Centrale***. La centrale est destinée à produire de *« l'électricité à partir d'un combustible nucléaire (…), (…) il existe plusieurs familles de réacteurs, que l'on appelle filières. Quatre constituants principaux sont nécessaires pour concevoir un cœur de réacteur : un combustible dans lequel se produit la fission ; un fluide caloporteur qui transporte la chaleur hors du réacteur ; un modérateur (sauf pour les réacteurs à neutrons rapides) qui permet de ralentir les neutrons ; des barres de commande qui contrôlent la réaction en chaîne (…) »*[196].
Cooccurrents associés à la forme *Centrale : uranium, mine d'uranium, combustible nucléaire, tritium*[197]*, isotope*[198]*, deutérium, césium*[199]

La technologie des réacteurs et leurs constructeurs (se reporter à l'annexe G, figure G.2)

Dans les récits d'ENRG_CN, nous retrouvons des descriptions techniques se rapportant aux réacteurs nucléaires utilisés en Chine :

Réacteurs[200] ***nucléaires*** : classés par type, par *puissance* exprimée en *MWe* (mégawatts électrique) et par génération : 1re génération, 2e génération, *3e génération* et 4e génération.

- Réacteur à eau pressurisée (ou REP) : eau sous pression et le combustible utilisé est de l'uranium enrichi, réacteur qui met l'accent sur la *sûreté* et la *sécurité* (résistance renforcée aux agressions externes, type chute d'avion, *choc violent*) :
  - REP réacteur à eau pressurisée : type *CPR1000, AP1000, CAP1400,*
  - PWR *pressurized water reactor* en anglais
  - EPR réacteur pressurisé européen
- Réacteur à eau bouillante (ou REB) : eau mais non pressurisée et le combustible utilisé est de l'uranium enrichi.
- Réacteur à eau lourde : combustible utilisé est de l'uranium naturel.
- Réacteur à neutrons rapides (ou RNR) : a été conçu pour utiliser la matière fissile (l'uranium et le plutonium) comme combustible nucléaire,
- Réacteur à caloporteur gaz (ou RCG) : susceptible de permettre la réalisation d'unités de petite taille (de 100 à 300 MWe), économiques et sûres.

AP1000 *:* un nouveau type de réacteur de 3e génération + développé par la compagnie américaine Westinghouse Electric Corporation, le premier de cette génération, un réacteur à eau pressurisée qui fonctionne dans un bon nombre de centrales nucléaires. Toutes les unités AP1000, soit quatre en Chine, devraient être opérationnelles d'ici 2016. La construction des unités, deux à Sanmen et deux à Haiyang dans la province du Shandong, ont été autorisées par Westinghouse et son partenaire (le groupe Shaw), en septembre 2007.
Formes associées à *AP1000 : Westinghouse, réacteurs*

[196] http://www.cea.fr/jeunes/themes/l-energie-nucleaire/le-fonctionnement-d-un-reacteur-nucleaire/les-differents-types-de-reacteurs (consulté le 27/08/2015).
[197] deutérium et tritium : atomes très légers, tous deux isotopes de l'hydrogène
[198] isotopes : atomes qui possèdent le même nombre d'électrons, mais un nombre différent de neutrons
[199] césium : un élément radioactif
[200] http://jeunes.edf.com/article/les-differents-types-de-reacteurs-nucleaires,64 (consulté le 27/08/2015).

CAP1400 : autre type de réacteur[201], développé par la Chine, sa conception est basée sur le réacteur AP1000 de Westinghouse Electric Co. La Chine possède les droits de propriété intellectuelle sur les CAP1400, ce qui permet d'exporter le réacteur. *«La technologie est en cours d'évaluation (...) pourrait être construit d'ici fin 2013 au plus tôt*», a déclaré Gu Jun, directeur général de *State Nuclear Power*.
Autres types de réacteurs, autres *technologies* : *ACPR1000, CP1000, CNP1000*

Rappelons que la forme *EPR* est absente du tableau 6.15, alors que deux EPR sont en cours de construction en Chine (*en construction, couverture de confinement*). Par contre, la forme Areva est présente, *Areva :* constructeur français développant entre autres la technologie EPR, « (…) *le groupe propose aux électriciens une offre qui couvre toutes les étapes du cycle du combustible, la conception et la construction de réacteurs nucléaires ainsi que les services pour leur exploitation. AREVA*[202] *investit également dans les énergies renouvelables afin de développer en partenariat des solutions à fort contenu technologique (...)* »[203]. Cooccurrents associés : *réacteurs, réacteurs nucléaires, trois, génération, troisième, technologie*, etc.

Autre constructeur *Westinghouse* dont la technologie AP1000 est liée à *Westinghouse Electric Company*[204], premier fournisseur au monde de la technologie nucléaire sûre et innovante, entreprise américaine créée en 1886.

Les producteurs d'électricité

Les producteurs d'électricité sont présents dans ENRG_CN, comme *Datang*, un des cinq grands producteurs d'électricité à partir du charbon, et un peu hydraulique, situé dans le nord de la Chine, société qui alimente en électricité toute la région de Pékin, Tianjin et cotée en bourse, ou *CGNP*, qui construit deux EPR à Taishan. Toutefois, un doute subsiste sur la résistance de l'acier de ses cuves où se produira la fission atomique puisqu'elles ont été forgées en France, comme celle de Flamanville[205].

D'autres cooccurrents sont présents dans le tableau 6.15, mais ceux-ci sont difficilement associables à une famille sans un retour au contexte. Mais la quintessence des formes calculées du sous-corpus fournit un précieux réseau de mots-clés pour nos restitutions de l'information. Les cooccurrents évoquent un grand nombre d'entités nommées issues du secteur de production d'électricité. Pour mieux comprendre le contexte chinois, nous tentons de réaliser de manière succincte un panorama de ce secteur.

[201] Source : le Quotidien du peuple en ligne, 4/02/2013, http://french.peopledaily.com.cn/Economie/8120966.html (consulté le 15/02/2015).
[202] Il est à noter que la branche Réacteurs et Services d'Areva a été reprise par EDF le 27 janvier 2016, journée qui a donné naissance au nouveau nucléaire français.
[203] http://www.areva.com/FR/groupe-57/leader-mondial-des-metiers-de-l-energie-nucleaire-et-energies-renouvelables.html (consulté le 19/02/2015).
[204] http://www.westinghousenuclear.com (consulté le 19/02/2015).
[205] Le Monde Economie, la Chine lance un concurrent de l'EPR français, publié le 6/05/2015, http://www.lemonde.fr/economie/article/2015/05/06/nucleaire-la-chine-lance-un-concurrent-de-l-epr-francais_4628880_3234.html (consulté le 27/08/2015).

Présentation du secteur de production d'électricité en Chine[206]

Ce domaine se décompose en trois catégories et s'organise de la façon suivante :

Cinq groupes publics de production d'électricité

1. 中国华能集团公司/Zhōngguó huánéng jítuán gōngsī/*China Huaneng Group*
2. 中国大唐集团公司/Zhōngguó dà táng jítuán gōngsī/*China Datang Corporation* (CDT)
3. 中国华电集团公司/Zhōngguó huádiàn jítuán gōngsī/*China Huadian Corporation*
4. 中国国电集团公司/Zhōngguó guódiàn jítuán gōngsī/*China Guodian Corporation*
5. 中国电力投资集团公司 简称 中电投/Zhōngdiàntóu/*China Power Investment Corporation*

Quatre sociétés publiques de production d'électricité

1. 国投电力/guó tóu diànlì/*SDIC Power Holdings* (*State Development and Investment Corporation Power*), forme mal segmentée dans le tableau 6.15, colorée en turquoise
2. 国华电力/guó huá diànlì/*Shenhua Guohua Power*
3. 华润电力/Huárùn diànlì/*China Resources Power*
4. 中广核/zhōngguǎnghé/*China General Nuclear Power Group* (CGN ou CGNP)

Deux sociétés de distribution (réseaux) d'électricité

1. 国家电网公司/Guójiā diànwǎng gōngsī/*State Grid Corporation of China* (SGCC)
2. 中国南方电网有限责任公司/Zhōngguó nánfāng diànwǎng yǒuxiàn zérèn gōngsī/*China Southern Power Grid*

Toutefois, quelques remarques sont à noter sur les sociétés ci-dessus. Celles-ci sont des sociétés mères possédant de très nombreuses filiales (Laponche, 2015). Les cinq grands groupes participent activement à la gestion et à l'exploitation des ressources fossiles, quant à CGN, elle est la seule entreprise à s'occuper du nucléaire dont l'EPR en Chine et des énergies propres sous tutelle du Conseil d'Etat. Mis à part les sociétés de distribution d'électricité, il est facile de remarquer également qu'il y a toujours une société qui gère les investissements et partenariats tant nationaux qu'internationaux dans les deux premières catégories (se reporter à l'annexe G).

6.2.6 EPR en Chine, veille active et veille ciblée par poly-cooccurrences évolutives

Malgré l'absence de l'EPR des réseaux cooccurrentiels issus des formes 核/hé/nucléaire et 核能/hé néng/énergie nucléaire, cette forme est réellement présente dans ENRG_CN. Afin de mettre en parallèle les informations d'EPR en Chine et en France, une veille active et ciblée par poly-cooccurrences évolutives a été effectuée avec les paramètres par défaut, co-freq : 2, seuil 10, contexte . ! ?.

Nous allons établir les réseaux poly-cooccurrentiels de la forme *EPR* d'abord globalement pour les trois années retenues, 2010, 2011, 2012 puis pour chaque année.

[206] http://baike.baidu.com/view/2214304.htm

Notons que, des travaux similaires ont été menés à l'aide de ce même principe (Leblanc et Martinez, 2006 ; Martinez, 2012 ; MacMurray, 2012).

Réseaux poly-cooccurrentiels de la forme-pôle *EPR* en Chine

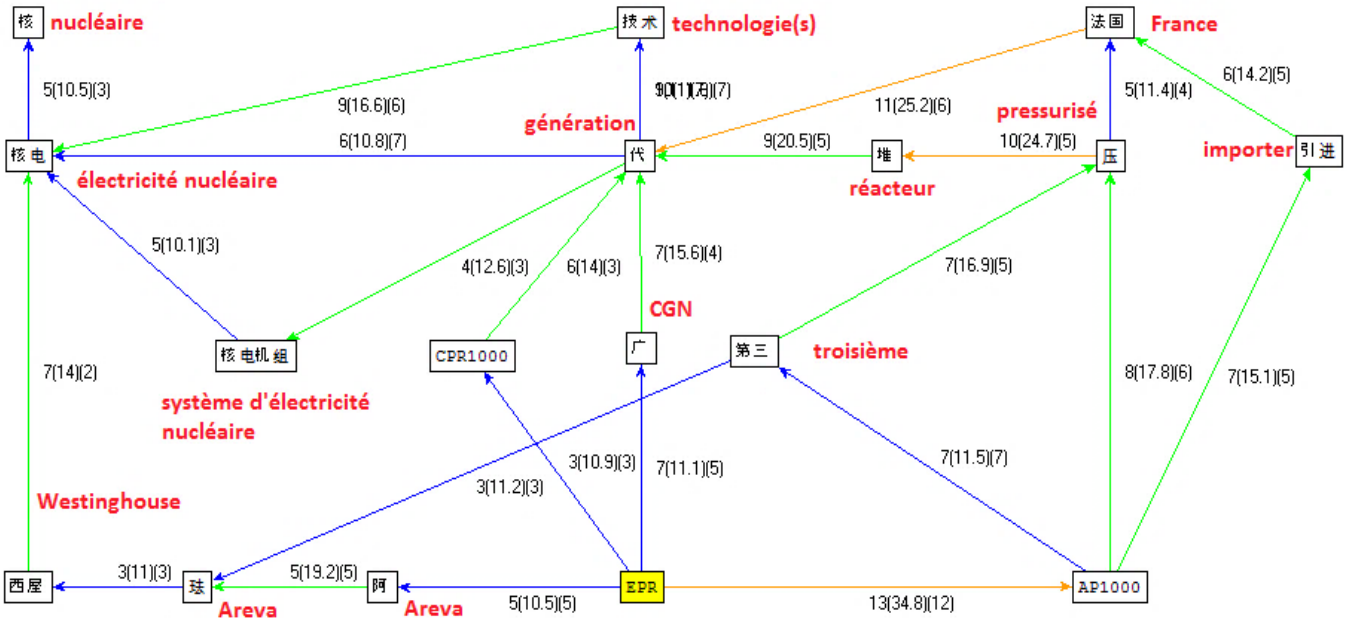


Figure 6.10 ENRG_CN de 2010 à 2012 : réseau poly-cooccurrentiel *EPR*

Dans la figure 6.10 ci-dessus, le réseau poly-cooccurrentiel de la forme-pôle *EPR* obtenu fédère *au-delà de la cooccurrence binaire*, *les unités repérées* [dans la narration du volet chinois] *suivant leur originalité fréquentielle et les* ordonne *sous une forme articulée qui se rapproche du texte naturel* (Martinez, 2012). Un réseau de mots-clés est ainsi tracé pour la période de 2010 à 2012 et formé par les unités lexicales les plus co-présentes autour de la forme-pôle *EPR*. Par conséquent, cet ensemble de formes qui vise à désigner l'emploi des mots les plus récurrents autour de la notion *EPR*, véhicule les informations les plus saillantes associées à la forme *EPR* de ce volet.

Nous pouvons reconstituer le sens partiel de ce réseau étalé sur les trois années sans retour au contexte comme suit :

EPR >>> système d'électricité >>> électricité nucléaire >>> nucléaire

EPR >>> réacteur (à eau pressurisé) de technologies de la troisième génération du nucléaire >>> Areva >>> importer >>> France

EPR >>> CGN (*China General Nuclear Power Corporation*) *>>> CPR1000 >>> AP1000 >>> Westinghouse* (*Westinghouse Electric Company*)

Au-delà de ces suites d'information reconstituées sur la période retenue ci-dessus, nous tentons de connaitre l'évolution chronologique de ces informations véhiculées par ces formes.

Réseau poly-cooccurrentiel évolutif

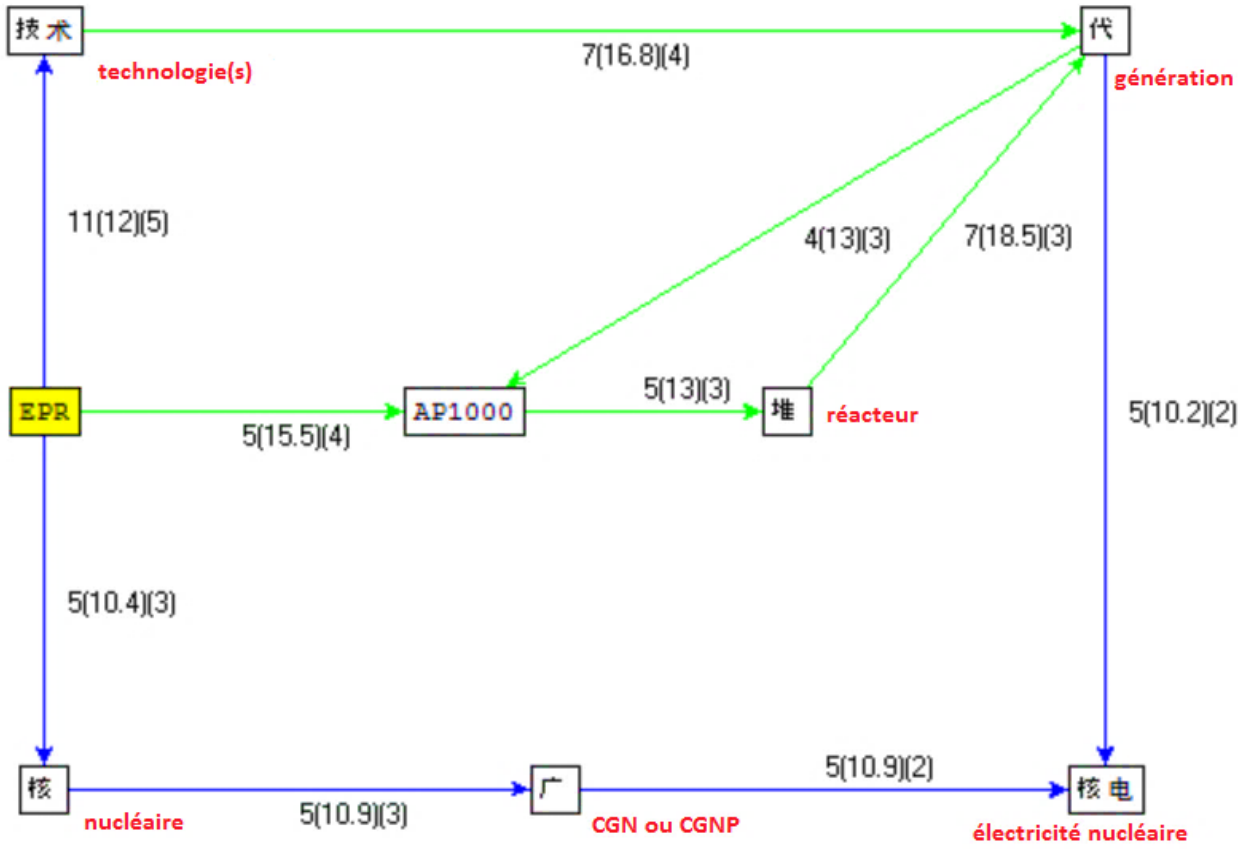


Figure 6.11 ENRG_CN 2010 : réseau poly-cooccurrentiel *EPR*

Selon les figures 6.10 (sur 3 ans) et 6.11 (sur 2010) ci-dessus, nous constatons que la forme *EPR* est associée à *AP1000*, réacteur fondé sur une technologie de troisième génération pour produire de l'électricité nucléaire par le groupe CGN. Rappelons que le réacteur AP1000 est une technologie américaine de la société *Westinghouse* marquant le début de la troisième génération du nucléaire. Celui-ci est utilisé dans un bon nombre de centrales nucléaires (Laponche, 2015).

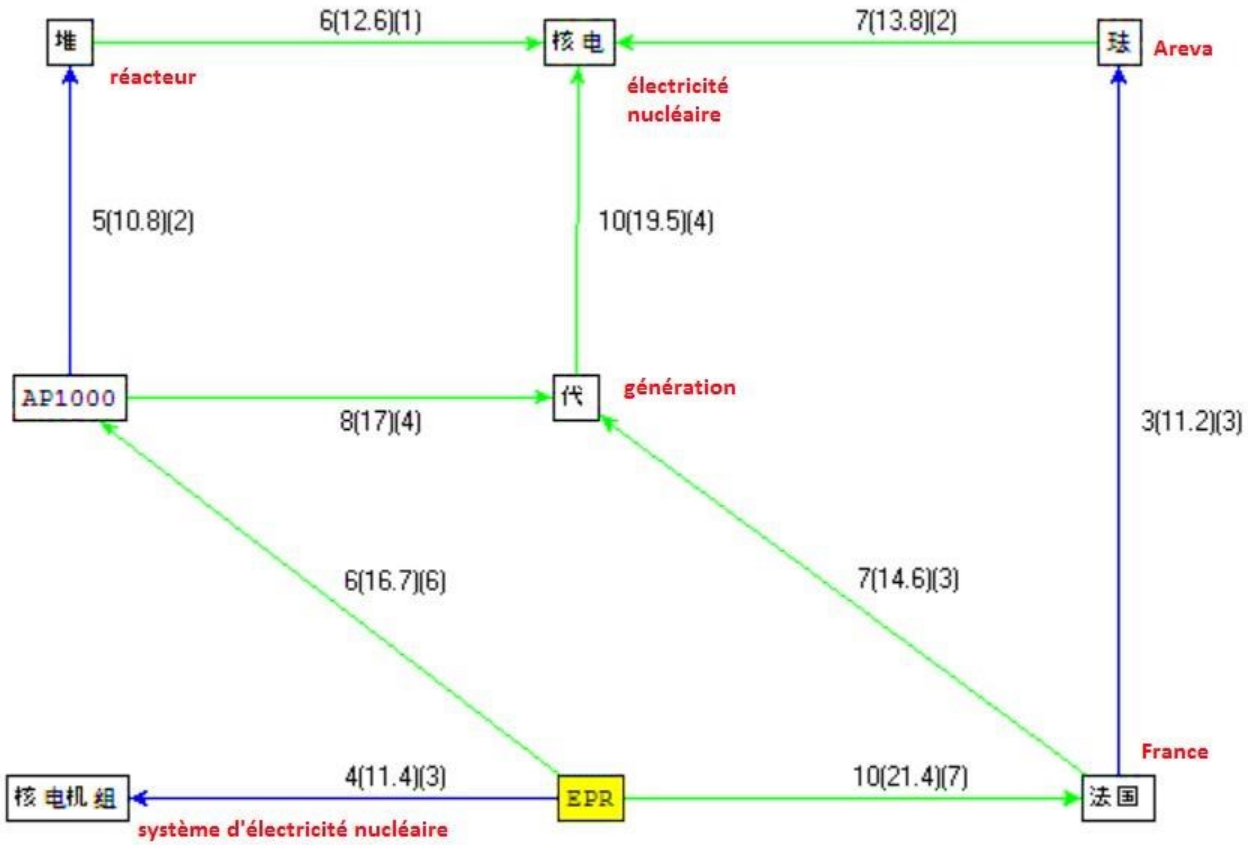


Figure 6.12 ENRG_CN 2011 : réseau poly-cooccurrentiel *EPR*

Deux notions nouvelles sont apparues dans le réseau de la figure 6.12, ci-dessus : la société Areva et la France. Nous allons expliquer la raison de leur apparition par des retours aux contextes du sous-corpus.

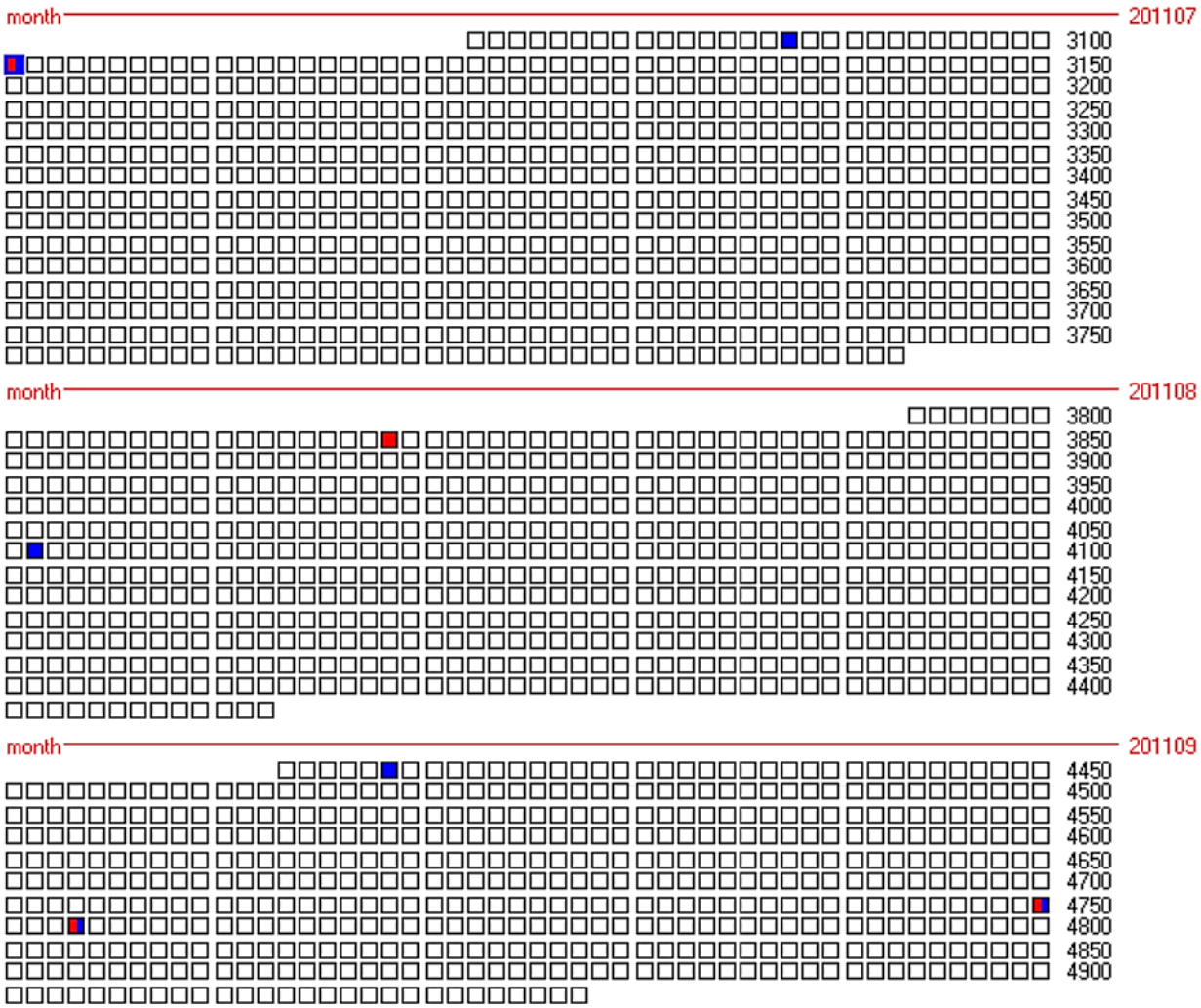


Figure 6.13 ENRG_CN 2011 : carte des sections pour les formes *EPR* en bleu et *Areva* en rouge

La projection du couple de formes *EPR* en bleu et *Areva* en rouge sur la carte des sections, figure 6.13 ci-dessus, permet de visualiser leurs répartitions simultanées dans le sous-corpus. Seuls trois articles contiennent les deux formes. Nous proposons une traduction d'extraits de ces trois articles.

Premier article

<day=20110701> <article=8020> # 日本 地震 和 核 事故 对 全球 核电 发展 产生 很 大 影响 , 最近 德国 和 瑞士 相继 宣布 逐步 退出 核电 使 法国 备受 压力 . 但 法国 领导人 重申 , 法国 的 能源 政策 不 变 , 将 继续 走 核电 发展 道路 . 法国 工业 和 能源 部 部长 贝松 近日 表示 , 德国 退出 核电 短期 会 使 法国 受到 一些 影响 , 但 不 会 改变 法国 长期 使用 核能 的 选择 . (……)

核电 安全 是 核电 发展 的 关键 , 也 是 人们 最 关心 的 问题 . 法国 原子能 委员会 主席 贝尔纳 毕戈 认为 , 日本 福 岛 核 事故 的 技术 原因 并 不 是 一个 谜 , 新 一代 核反应堆 EPR 可以 满足 安全 方面 的 需要 . 法国 阿 海 珐 集团 总裁 安妮 罗 薇 中 认为 , 新 一代 核反应堆 EPR 完全 有 安全 保障 , 即便 遇 最 坏 情形 堆 芯 熔化 , 也 可以 限制 在 核电站 内部 , 不 会 泄漏 到 空气 , 大 地 中 去 .

Traduction française

Le tremblement de terre au Japon et les accidents nucléaires ont un impact énorme sur le développement de l'énergie nucléaire dans le monde. L'Allemagne et la Suisse ont annoncé récemment leurs sorties du nucléaire, ce qui a eu pour conséquences de faire grincer des dents la France. Mais les dirigeants français ont réaffirmé que la politique énergétique en France reste inchangée, et poursuit la voie du nucléaire. Ministre chargé de l'Industrie, de l'Énergie et de l'Économie numérique, Eric Besson a récemment déclaré que l'Allemagne se retire de l'énergie nucléaire, cette décision pourrait avoir un certain impact en France à court terme, mais ne modifiera pas la prépondérance à long terme du nucléaire en France. (.....)

La sûreté nucléaire est la clé du développement de l'énergie nucléaire, elle est aussi ce qui préoccupe le plus les gens. Le président du Commissariat à l'énergie atomique (CEA), Bernard Bigot, estime que les raisons techniques de l'accident nucléaire de Fukushima ne sont pas un mystère, la nouvelle génération de réacteurs nucléaires EPR peut répondre aux exigences de sécurité. La PDG d'Areva, Anne Lauvergeon, pense que la sûreté de la nouvelle génération des réacteurs nucléaires EPR est entièrement garantie, même dans le pire des cas, si le cœur du réacteur fond, la fuite se limitera à l'intérieur de la centrale, et n'entrera ni dans l'air, ni dans la terre.

Deuxième article

<day=20110921> <article=9669> (…) # 据 了解 , 目前 国内 在建 的 第三 代 核电 示范 项目 已 达 6 个 , 这 其中 包括 4 个 从 美国 西屋 公司 引进 的 AP1000 和 2 个 从 法国 阿 海 珐 引进 的 EPR 技术 项目 . 专家 表示 , 政策 上 计划 到 2014 年 前 力 推 这些 示范 项目 的 全面 建成 . 同时 , 推动 国内 主要 核电 设备 商 全面 提升 核电 设备 设计 与 制造 的 融合 能力 , 以及 设备 关键 材料 的 自主 研发 能力 . 争取 到 2015 年 前 , 实现 稳定 年产 12 套 左右 核岛 设备 和 常规 岛 汽轮机 等 关键 设备 的 生产能力 . (…)

Traduction française

Selon certaines sources, jusqu'à présent, 6 projets pilotes de construction nationale de réacteurs de troisième génération sont en cours, dont 4 basés sur AP1000, importés de la société américaine Westinghouse aux États-Unis et 2 sur EPR, importés d'Areva en France. Les experts disent que ces 6 projets sont prévus d'être achevés d'ici 2014. Parallèlement, ces projets devraient permettre de renforcer l'amélioration concernant la conception et la fabrication d'équipements nationaux chez les grands fabricants, ainsi que des capacités d'autonomie dans la recherche et le développement des matières clés et des appareils. D'ici 2015, il faudrait atteindre une capacité de production annuelle stable d'environ 12 ensembles d'équipements de l'îlot nucléaire et de la turbine de l'îlot classique et d'autres équipements essentiels. (…)

Troisième article

<day=20110921> <article=9673> (…)# 根据 核电 安全 检查 结果 , 即将 出台 的 " 核电 安全 规划 " 提出 , 未来 新 上 核电 项目 要 按照 国际 先进 标准 设计 下一代 核电站 , 在 核电 技术 设备 上 要 全面 引进 包括 AP1000 (美国 西屋 公司 独创 先进 非 能动 压 水 堆) 和 EPR (法国 阿 海 珐 公司 研发 欧洲 压 水 堆) 在内 的 第三 代 核电 技术 , 同时 要求 尽量 新 上 大 容量 设备 , 安全 指标 和 质量 标准 均 比 " 核电 中长期 发展 规划 " 要求 更 高 .(…)

Traduction française

> Selon les résultats de l'inspection de la sûreté nucléaire, le nouveau Programme de sécurité nucléaire qui sortira prochainement, souligne que tous les nouveaux projets concernant la construction des centrales nucléaires doivent se baser sur les nouvelles normes internationales. Il faudrait introduire et déployer les technologies de réacteurs nucléaires de troisième génération, en particulier AP1000 (Westinghouse, réacteur pressurisé en évacuation passive, Passive Safety Systems[207]) et EPR (Areva R & D, France). D'ailleurs, il faudrait fabriquer les équipements dédiés à de grandes capacités de production autant que possible, et accroitre les exigences en matière des normes et des indicateurs de sécurité et de qualité, exigences plus élevées que celles du programme précédent sur le développement à long terme. (…)

Les extraits d'articles révèlent que d'une part, après l'événement de Fukushima, une véritable prise de conscience s'est opérée en matière de sécurité et sûreté nucléaire dans l'ensemble de la population chinoise, ceci explique l'apparition des formes *EPR* et *Areva* en 2011 (se reporter à l'explication du cooccurrent *sécurité* de la forme-pôle *nucléair*e dans l'annexe J), d'autre part, après avoir acquis les deux types de réacteurs de troisième génération, la Chine cherche d'abord à déployer massivement ces nouvelles technologies, en particulier AP1000. Par la suite nous faisons l'hypothèse qu'elle se familiarise aux bonnes pratiques de l'EPR afin de créer une nouvelle version locale de ce réacteur.

Grâce à la forme *CPR1000* détectée par les calculs sur les 3 ans, un article révélateur a été repéré par la projection sur la carte des sections, dont voici un extrait ci-dessous.

> <day=20100708><media= 新 华 社 Agence de Chine- 望 东 方 周 刊 Hebdomadaire Observer l'Est><article=251># 目前 , 中 广 核 集团 确立 的 战略 计划 为 : " 以 CPR1000 改进 型 (自主 品牌 中国 改进 型 压 水 堆 核电 技术) , 自主 品牌 三 代 核电 技术 CPR1700 为 主体 的 核电 技术 体 系 , 吸收 借鉴 三 代 压 水 堆 核电 技术 的 设计 理念 , 使 其 最终 与 第三 代 压 水 堆 核电站 技术 融为一体 , 形成 具有 自主 知识 产权 的 CPR 系列 核电 技术 " . # 并且 , 中 广 核 已经 开始 为 日 后 自主 品牌 的 出口 寻找 海外 市场 .

Traduction française

> # Actuellement, le plan stratégique du groupe CGN s'établit comme suit : CPR1000 version avancée et CPR1700 sont deux réacteurs dont la Chine possède les droits de propriété intellectuelle, le premier est basé sur la technologie pressurisée et le deuxième sur la troisième génération de réacteurs nucléaires. En s'inspirant de la conception technologique de la troisième génération de réacteurs pressurisés, il faut créer un système complet de centrales nucléaires avec les réacteurs pressurisés de troisième génération afin de créer un ensemble de technologies dont la Chine possèdera ses propres propriétés intellectuelles baptisées série CPR. Par ailleurs, le groupe CGN a déjà commencé la recherche de futurs marchés pour l'exportation de sa propre marque.

L'extrait ci-dessus confirme notre hypothèse.

Il est également à noter qu'une forme faux-ami d'*EPR* a été détectée lors de l'exploration, à savoir, *Extended Producer Responsibility,* EPR dans l'article N°9338*,* issu du site *Chinadialogue.*

[207] AP1000 PWR : The advanced-passive safety systems

Par ailleurs, la traduction de la forme *Areva* n'a pas été recensée dans le dictionnaire du segmenteur ICTCLAS, ce qui a conduit à une récupération partielle de son vrai nom, à savoir, 阿海珐/Ā hǎi fà /Areva. Une fois de plus, la problématique de la segmentation du chinois s'avère indéniablement cruciale dans tout traitement automatique de cette langue. La vérification de la qualité de segmentation nécessite de très bonnes connaissances de la langue et de la culture telles que : les noms propres (ex. 钱柜/qián guì/nom d'une chaîne commerciale de Karaoké), les toponymes (ex.榄核/làn hé), la terminologie (ex.竞 购 Jìng gòu/appel d'offres), les sigles (ex. 中广核/zhōng guǎng hé/CGN ou CGNP), les acronymes (ex.北约 běi yuē/OTAN), les abréviations (ex.国投/Guó tóu/*State Development & Investment Corporation* (SDIC) et 弃核/qì hé/abandonner le nucléaire), les diminutifs (néologismes) (ex. 减排/jiǎn pái/réduction des émissions de GES), les néologismes (ex. 低碳/dī tàn/densité carbonique faible), etc., où les pratiques langagières ne cessent d'évoluer.

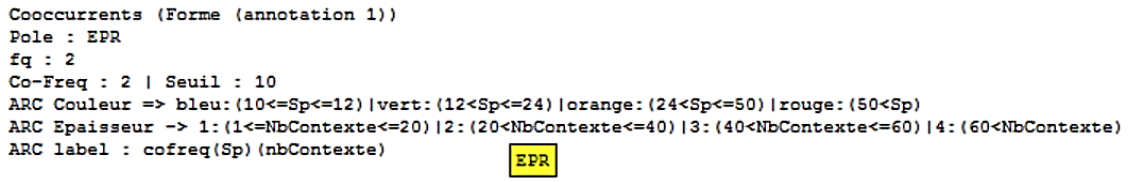


Figure 6.14 ENRG_CN 2012 : réseau poly-cooccurrentiel *EPR*

Malgré la récupération exhaustive des articles en 2012, la forme *EPR* y compris toutes ses traductions demeurent totalement absentes dans le réseau poly-cooccurrentiel d'ENRG_CN, comme le montre la figure 6.14 ci-dessus. Toutefois, cette forme a été employée deux fois dans deux articles différents du sous-corpus, dont les extraits et traductions sont ci-dessous :

<day=20120104> <media=中国经济周刊 Hebdomadaire de l'économie de Chine><article=11754> (…)
中国 在 浙江 三门 新建 的 AP1000 核电站 机组 , 第一 台 计划 在 2013 年 并网 运行 . 这 将 是 世界 上 第一 座第三 代 AP1000 核电站 , 比 美国 提前 两 年 半 . 然而 , 2011 年 3 月 11 日 , 日本 福岛 核 事故 却 给 快速 发展 的 中国 核电 事业 来 了 个 急刹车 . " 弃 核 " 之 争 甚嚣尘上 , 中国 核电 怎么办 ? # 核电 技术 发展 历程
第一 代 上 世纪 50 年代 , 前 苏联 和 美国 分别 建成 实验性 和 原型 核电机组 , 核能 发电 的 技术 可行性 被 证明 .
第二 代 上 世纪 60 年代 后期 , 陆续 建成 电功率 在 30 万 千瓦 的 压 水 堆 , 沸水 堆 , 重水 堆 , 石 墨水 冷 堆 等 核电机组 , 核电 的 经济性 也 得以 证明 . 但 第二 代 核电站 应 对 严重 事故 的 措施 比较 薄弱 . # 第三 代 上 世纪 90 年代 , 美 欧 先后 出台 了 " 先进 轻水 堆 用户 要求 " 文件 和 " 欧洲 用户 对 轻水 堆 核电站 的 要求 " .
#第三 代 核电机组 的 设计 主要 有 美国 的 AP1000 压 水 堆 和 ABWR 沸水 堆 , 以及 欧洲 的 EPR 压 水 堆 等 型号 .
第四 代 2000 年 1 月 , 美英 等 十 国 约定 共同 合作 研究 开发 第四 代 核能 技术 . 安全性 和 经济性 将 更加 优越 , 高温 气冷 堆 , 熔 盐 堆 , 钠 冷 快 堆 就 是 具有 第四 代 特点 的 反应堆 . 第四 代 目前 处在 原型 堆 技术 研发 阶段 .

Traduction française

La mise en service prévue en 2013, de la centrale située à Sanmen dans la province du Zhejiang, première centrale au monde basée sur la technologie AP1000, serait de deux ans et demi en avance par rapport à celle des Etats-Unis. Or, le 11 mars 2011, l'accident de Fukushima a donné un grand coup de frein au secteur nucléaire chinois qui était en plein essor. Le dilemme d'abandonner le nucléaire ou non sème la zizanie, alors que faire pour l'électricité nucléaire chinoise ?

L'historique du développement du nucléaire :

1. la première génération : dans les années 1950, l'ex-Union soviétique et les États-Unis ont construit indépendamment un prototype expérimental de centrale nucléaire, la faisabilité technique de l'énergie nucléaire a été prouvée.
2. la deuxième génération : à la fin des années 60, construction progressive d'un réacteur pressurisé de puissance 300.000 kilowatts, puis BWR (réacteur à eau bouillante ou REB en français, en anglais BWR pour *boiling water reactor*), puis PHWR (réacteur à eau lourde pressurisée ou *pressurised heavy water reactor*, PHWR), puis réacteur modéré au graphite ou *Graphite-moderated reactor* (Chicago Pile-1), et depuis l'économie de l'énergie nucléaire a été prouvée. Mais cette deuxième génération de centrales nucléaires a fait face à des accidents graves en raison de mesures de sécurité relativement faibles.
3. la troisième génération : dans les années 90, les États-Unis ont mis en place un réacteur à eau légère (REL) ou *light water reactor* (LWR) et l'Europe a émis une liste d'exigences en matière de centrales nucléaires LWR. La conception de la troisième génération de centrales nucléaires est principalement basée sur AP1000, réacteur à eau pressurisée et US ABWR, réacteur à eau bouillante, et EPR (*European Pressurized Reactor*).
4. la quatrième génération : en Janvier 2000, 10 pays dont les Etats-Unis et la Grande-Bretagne ont convenu de travailler ensemble à la recherche et au développement de la technologie nucléaire de quatrième génération. La sécurité et l'économie seront plus favorables, HTR, réacteur (modulaire) à lit de boulets (de l'anglais *pebble bed (modular) reactor* abrégé PBR ou PBMR), une technologie de réacteur nucléaire à très haute température, réacteur nucléaire à sels fondus (RSF) (en anglais, *molten salt reactor* : MSR), réacteur à neutrons rapides à caloporteur sodium (RNR-Na) sont les caractéristiques des réacteurs de la quatrième génération. Ces technologies de réacteurs de quatrième génération sont actuellement en phase de prototype.

<day=20120903> <media=南方报业网-南方周末><article=14477> (…)" 地平线 " 核电 项目 计划 在 英国 威尔士 安 格 尔 西 岛 Anglesey 和 格洛斯特郡 Gloucestershire 建造 反应堆 . # 在 中 广 核 的 一 位 研究员 看来 , 中国 核电 企业 与 西方 国家 的 核电 企业 联合 竞 购 , " 是 一 种 强强联合 , 各取所需 " . # 目前 我国 的 核电 企业 正 在 设计 的 第三 代 反应堆 , 包括 中 核 ACP1000 , 国 核 CAP1400 , 都 处于 设计 接近 完成 阶段 , 还 未 投入 使用 . 这次 联合 竞 购 更 多 是 两 家 国外 企 业 西屋 电气 AP1000 与 阿 海 珐 EPR 的 第三 代 主流 技术 的 角逐 . # " 英国 规定 每 一 种 反应 堆 堆 型 都 需 获得 英国 政府 许可 , 方可 在 英国 投入 使用 , 而 这 一 过程 可能 长 达 五 年 . " 道格帕尔 介绍 . # 中国 核电 技术 尚 难 打入 英国 核电 市场 , 中国 资本 已 先行 一 步 . (…)

Traduction française

Le Projet "Horizon" prévoit de construire des réacteurs nucléaires au Pays de Galles sur l'île d'Anglesey et dans le comté de Gloucestershire. (…) Selon un chercheur du groupe CGNP, le fait que les sociétés nucléaires chinoises et occidentales participent à des offres conjointes, serait une combinaison gagnant-gagnant. (…) #Les entreprises d'énergie nucléaire chinoises conçoivent des réacteurs de troisième génération, dont ACP1000 et CAP1400, de la phase de conception à la mise en service, actuellement en phase d'achèvement. Cette offre conjointe de réacteurs serait plutôt une rivalité entre Westinghouse AP1000 et la technologie EPR d'Areva, technologie phare de la troisième génération. (…)

> # « Le Royaume-Uni exige une licence d'exploitation auprès du gouvernement pour chaque réacteur avant sa mise en service, ce processus peut prendre jusqu'à cinq ans. » précise Monsieur Dong Parr. (…) La technologie de l'énergie nucléaire de la Chine a encore des difficultés à percer le marché nucléaire britannique, la capitale chinoise a fait un pas en avant.

Les deux extraits ci-dessus attestent l'ambition internationale des chinois de développer leurs propres systèmes de centrales nucléaires et d'exporter leurs savoir-faire et technologies en prenant les bonnes pratiques d'une part chez les Américains avec l'AP1000 de Westinghouse, et d'autre part en France avec l'EPR d'Areva.

6.3 Résonance trilingue globale du corpus comparable ENRG autour de la forme *EPR*

L'expérience de l'EPR, que nous venons de vivre à travers les trois sous-corpus, en trois langues et trois sphères, nous procure trois visions différentes. Cette expérience nous montre d'une certaine façon qu'il est possible d'anticiper l'avenir de ce secteur et que ces visions peuvent se traduire par une liste de mots-clés pour chacune des sphères, à savoir :

- France : déboire, coût et délais, conséquences,
- USA : business, protectionnisme, paradoxe,
- Chine : besoins, ambitions internationales et autonomie.

En décembre 2007, la France a démarré la construction de son propre EPR, qui devait être opérationnel en 2012, or il s'avère que le chantier a connu des déboires, dus à des problèmes liés au couvercle, à des éléments de la cuve d'acier, à des pièces essentielles défaillantes pour la sûreté, etc.[208], le tout entraînant des coûts et des délais supplémentaires colossaux. Les conséquences seront lourdes, la technologie française a donc subi un revers et l'avenir commercial de ce réacteur est compromis.

L'EPR n'a jamais été une préoccupation majeure du secteur nucléaire chez les Américains, puisque par tradition, ils appliquent des mesures protectionnistes [209] récurrentes envers l'extérieur. Paradoxalement, « *business is business* », les États-Unis ont vendu[210] leur technologie AP1000 à la Chine afin de décrocher le contrat de vente de leurs réacteurs. Par ailleurs, comme nous l'avons constaté dans notre travail, tous les articles EPR sont classés dans la rubrique *business* du NYT, *de facto,* l'affaire EPR n'est donc qu'une affaire de *business*.

La deuxième puissance mondiale actuelle cherche désespérément à accroître sa capacité de production d'électricité « proprement », c'est-à-dire, une énergie à la fois non polluante et totalement indépendante des technologies étrangères, en raison de ses besoins grandissants en énergie (se reporter à l'annexe G). Pour ce faire, elle a développé, entre autres, un programme nucléaire ambitieux en vue de créer ses propres technologies de réacteurs de pointe, à l'aide d'achats de réacteurs français EPR et AP1000 américains, de façon à ce qu'elle puisse garantir son autonomie nucléaire et les exporter à l'international.

[208] http://www.lemonde.fr/economie/article/2014/11/19/epr-de-flamanville-les-quatre-maledictions-d-un-chantier-controverse_4526032_3234.html
[209] http://www.lemonde.fr/ameriques/article/2002/06/10/le-protectionnisme-une-tradition-americaine-par-antoine-bouet_279820_3222.html
[210] http://www.sortirdunucleaire.org/La-Chine-prefere-Westinghouse-a

6.4 Résonance trilingue globale du corpus comparable ENRG autour des formes *énergies+*

Les résultats synthétisés des formes calculées de nos trois études du corpus ENRG sont disponibles dans le tableau L.1 de l'annexe L.

Points communs et points divergents des trois sous-corpus

Dans un contexte d'articles de presse, ENRG_FR, US et CN relatent communément les catastrophes nucléaires (Tchernobyl et surtout Fukushima), événements déclencheurs, révélant une profonde prise de conscience sur la sécurité et la sûreté du nucléaire, entraînant des choix et mesures politiques et énergétiques dans le monde.

Les toponymes associés aux sites nucléaires sont fortement représentés dans les trois sous-corpus, cependant, chacun se manifeste par une spécificité : dans ENRG_FR, c'est un éventail complet de tous les sites concernés par des activités nucléaires, dans ENRG_US, ce sont les centrales nucléaires à problèmes, et dans ENRG_CN, les centrales nucléaires de nouvelle génération en construction.
Les entités nommées relatives aux hommes politiques sont omniprésentes dans les trois sous-corpus, cependant une exception, la Chine n'évoque pas le nom de son président.

Les formes liées aux types et sources d'énergies (*fossiles* et *renouvelables*) sont très récurrentes en Chine et aux Etats-Unis, ce phénomène est dû à une consommation effrénée d'énergies fossiles, une pratique entrainant de graves pollutions. De fait, ils sont contraints de chercher de nouvelles énergies propres et durables (*photovoltaïque* et *éolienne*), alors que la France tente de promouvoir l'énergie éolienne, bien que la quasi-totalité de l'énergie soit nucléaire, d'où la raison du nombre important de formes se rapportant aux activités, actualités (*démantèlement*, *convois*, *CGT*, *campagne électorale*) et politiques nucléaires, mais également, aux opérateurs-fournisseurs d'électricité, aux divers organismes de surveillance nucléaire, aux mouvements écologistes, etc.

Une exubérance d'emploi de formes s'attachant à un très large panel de réacteurs (Laponche, 2015) et de constructeurs (se reporter à l'annexe G) témoigne une singularité de certaines informations chinoises.

Ainsi, dans les années à venir, nous pouvons supposer que les Etats-Unis et la Chine continueront à développer leur stratégie nucléaire en mettant l'accent sur l'accroissement des énergies solaires et éoliennes, tout en favorisant d'autres énergies propres. Quant à la France, elle se heurte à un dilemme : sortir ou non du nucléaire ?

Cette thèse montre que les formes anaphoriques, extrêmement fréquentes en français, en anglais voire dans la plupart des langues occidentales, permettant entre autres d'embellir le discours par une richesse lexicale et culturelle, sont très usitées à l'écrit dans le monde occidental. Or, dans la langue chinoise, cette pratique reste encore peu fréquente. C'est une des raisons pour lesquelles une veille textuelle en chinois par les cooccurrences et poly-cooccurrences récolte parfois des résultats plus fouillés mais à condition que la qualité de la segmentation du chinois soit garantie. Ces constats linguistiques vont pouvoir apporter des pistes de réflexion dans l'objectif d'améliorer la pertinence des moteurs de recherche et la fouille d'informations multilingues.

Les étiquettes sociales des locuteurs jouent un rôle important dans toute analyse du discours. Si la presse des principaux médias se définit comme l'une des références du bon usage, alors, il serait intéressant de scruter la façon dont les ONG affûtent leurs discours dans nos thèmes de recherche, en les comparant avec ceux de nos journaux occidentaux sélectionnés. Des repérages sont alors nécessaires selon les sources : il existe, en effet, une rhétorique gauchiste, ultra-libérale, etc.

Dans l'objectif d'appuyer et de compléter les enjeux du multilinguisme, l'approche de veille par corpus parallèle bilingue chinois-anglais sera déclinée dans le chapitre suivant.

Conclusion du chapitre

Les années retenues du sous-corpus ENRG_CN sont 2010, 2011, et 2012, étant donné que les années 2008, 2009 et 2013 ne contiennent pas un nombre suffisant de données pour mener des analyses textométriques pertinentes. Cependant, les différentes analyses réalisées ont permis de mettre en évidence une quantité textuelle du sous-corpus ENRG_CN nettement supérieure à celle des sous-corpus ENRG_FR et ENRG_US. Nous avons aussi relevé au mois d'août 2011 un nombre plus élevé de formes différentes, et un nombre plus élevé d'articles en juillet 2010.

L'analyse des spécificités nous a permis de relever une présence dans certains articles du sous-corpus d'informations écologiques essentiellement propres à la Chine. Les événements internationaux ont aussi été rapportés. A l'instar des sous-corpus ENRG_FR et ENRG_US, le sous-corpus ENRG_CN s'est caractérisé par un accroissement de vocabulaire sur la période d'étude retenue. Il convient de noter que cette analyse de spécificités a été rendue plus difficile par les problèmes liés à la segmentation du chinois. Des calculs de cooccurrences ont aussi été menés à partir des formes 核能 /hé néng/énergie nucléaire, sur ENRG_CN. La classification des formes cooccurrentes de ce mot a permis d'illustrer les inquiétudes et les enjeux majeurs de la Chine en pleine phase de développement. L'analyse transversale de ces trois sous-corpus sur la forme-pôle *EPR* a permis de mettre en évidence trois conceptions distinctes de ce sujet. Le sous-corpus français fait apparaître des ensembles de mots reflétant des déboires techniques et des coûts élevés. Dans le sous-corpus américain, les trois mots-clefs suivants ressortent : "business, protectionnisme et paradoxe". Dans le sous-corpus chinois, ce sont les besoins, la volonté d'autonomie et les ambitions internationales qui se dégagent dans le domaine énergétique. En ce qui concerne le thème de l'énergie, les formes liées aux sources et types d'énergie sont très fréquentes dans les sous-corpus ENRG_CN et ENRG_US. Le sous-corpus français nous montre une volonté de sortir du nucléaire et de chercher des sources d'énergie alternatives.

Le septième et dernier chapitre sera consacré à la veille via le corpus parallèle anglais-chinois CLRG. Tout comme cela a été fait dans les chapitres précédents, une analyse textométrique sur la base des occurrences, formes et hapax a été réalisée pour les deux volets CLRG_CN et CLRG_EN. Les aspects linguistiques du chinois fourniront des explications aux différences rencontrées dans le nombre d'occurrences de ces deux volets. Les diagrammes d'accroissement de vocabulaire du corpus CLRG bilingue seront également donnés et feront l'objet d'une analyse. La typologie textuelle et les spécificités de ce corpus seront aussi traitées. La présence du terme EPR dans ce corpus fera aussi l'objet d'une section. Le chapitre se terminera par une analyse transversale de nos deux corpus, à savoir, le comparable ENRG et le parallèle CLRG.

7. Veille parallèle anglais-chinois : énergies et EPR dans CLRG

Pour tenter de préciser et compléter ces résultats obtenus à partir d'importantes masses de données comparables, nous avons construit un corpus parallèle. Rappelons que les corpus parallèles sont constitués de textes sources et de traductions, et permettent d'une part, d'aborder les phénomènes translinguistiques avec beaucoup plus de sécurité, d'autre part, de travailler sur des matériaux constitués de traductions en plusieurs langues, afin de mieux cerner les problèmes de comparaison.

Nous ouvrons ce chapitre par un rappel succinct de la définition du corpus parallèle et comment un corpus parallèle bilingue pourrait apporter des avantages à la veille. Un dépouillement textométrique avec une approche des segments répétés du corpus entier nous livrera ses caractéristiques et informations principales. Par la suite, une veille bilingue sur la forme *EPR* sera appliquée aux textes bilingues (Gale et Church, 1993) afin de mettre en évidence comment les corpus multilingues à caractère comparé et parallèle se complètent dans un processus de veille, de restitution, de prévision et d'anticipation.

Rappelons que le corpus parallèle est « *un corpus qui contient des textes source et leur traduction* » (McEnery et Xiao, 2007). Celui-ci est souvent considéré comme une source traductologique intéressante par l'extraction des couples de lexiques bilingues (Zimina, 2004, 2005). Or, dans un processus de veille, ces lexiques bilingues issus des corpus parallèles ne sont pas uniquement des enrichissements dictionnairiques, ils permettent d'observer comment une langue aborde un sujet particulier par rapport à sa traduction dans une autre langue. Il est encore plus intéressant si nous comparons ces mêmes informations issues du corpus parallèle avec celles des autres corpus ou sous-corpus issus d'autres sources. Cela peut nous révéler des erreurs de traduction et de compréhension du texte d'origine, de l'information déformée voire de la désinformation.

7.1 Présentation du corpus parallèle anglais-chinois

Nous avons opté comme unité d'alignement des articles[211]le paragraphe. Le fichier du corpus bilingue aligné par paragraphe est stocké dans un fichier au format XML de manière hiérarchisée en fonction de la structure des articles, à savoir, titres, dates, URL, paragraphes, etc., lesquels sont rangés dans des balises spécifiques.

Rappelons qu'un corpus multilingue est dit aligné si les unités ou éléments textuels ou para-textuels (paragraphes, phrases, termes, etc.) sont mis en correspondance, langue par langue, dans le corpus multilingue parallèle qui le compose. D'une manière générale, un corpus parallèle est supposé déjà aligné. Mais cela nécessite un choix pour son unité d'alignement et un traitement technique long et coûteux. En effet, ce processus de l'alignement de corpus parallèle consiste plus précisément à effectuer *« la mise en correspondance des différents niveaux d'unités »* (Véronis, 2000) entre deux ou plusieurs volets.

[211] Comme déjà introduit dans le chapitre 4 (section 4.3), le corpus parallèle chinois-anglais a été construit à partir de programmes (se reporter à l'annexe M) entièrement écrits en Python (interface graphique Qt et environnement Eric), y compris ceux de nettoyage, conversion des signes, vérification d'alignement et extraction chronologique.

Afin d'obtenir un corpus proprement aligné, il est nécessaire de consacrer des efforts considérables, extrêmement gourmands en temps. Il s'avère que le corpus récupéré et nettoyé automatiquement, présente encore des décalages d'alignement au niveau des paragraphes, à savoir, 407 articles ayant de 1 à 31 paragraphes décalés, sur un total de 852, répartis comme suit :

- 1 paragraphe ≤ 386 articles ≤ 10 paragraphes
- 11 paragraphes ≤ 16 articles ≤ 20 paragraphes
- 21 paragraphes ≤ 4 articles ≤ 30 paragraphes
- 31 paragraphes = 1 article.

Cet ajustement des paragraphes a été effectué en partie manuellement avec d'une part, l'aide de programmes de détection des articles erronés, et d'autre part, des lectures attentives bilingues des articles concernés. En effet, l'article source n'est pas forcément dans la même langue, c'est-à-dire, la source peut être chinoise ou anglaise et *vice-versa*. La complexité réside dans ce mixage de sens de traduction. Afin de respecter au mieux l'authencité d'informations relayées dans les articles, cet ajustement exige la détermination de la langue source de l'article erroné. Les paragraphes rectifiés en fonction du sens de traduction (chinois vers anglais ou anglais vers chinois), sont selon les types d'erreurs, supprimés ou fusionnés ou regroupés ou ajoutés ou complétés, etc., et ainsi alignés. Quant à la segmentation du volet chinois, elle a été réalisée par Jieba (module Python). Une fois le fichier segmenté, celui-ci est transformé et trié au format Lexico par des programmes spécialement conçus en Python, puis importé dans le logiciel MkAlign. Le résultat est présenté ci-dessous.

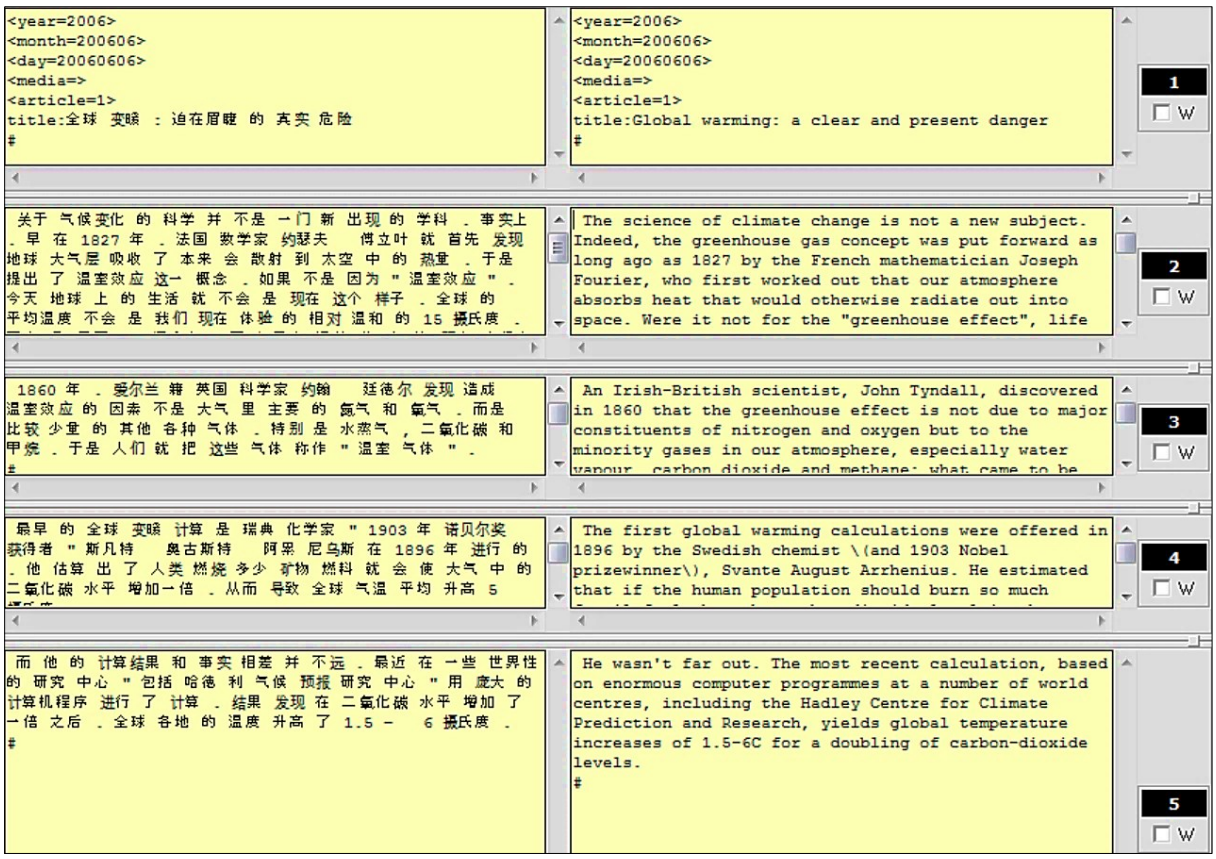


Figure 7.1 CLRG de 2006 à 2014 : extrait du corpus bilingue chinois-anglais aligné chargé dans le logiciel Mkalign

Comme l'illustre la figure 7.1, dans le cadre du logiciel Mkalign, le corpus a pour langue source (à gauche) le chinois et langue cible (à droite) l'anglais. Ces deux volets composés d'unités textuelles, en l'occurrence un paragraphe affiché dans un rectangle, sont ainsi alignés. Cependant, dans le couple gauche-droite de textes, le sens de la langue de départ vers la langue d'arrivée peut être inversé. Ce modèle hybride du sens de traduction est une spécificité du site *Chinadialogue.net*, qui produit des articles bilingues.

CLRG possède 18 720 paragraphes représentant 852 articles, couvrant la période comprise entre le 06/06/2006 et le 27/08/2014, période correspondant à tous les articles disponibles de la rubrique *Climate change & Energy* dans le site au jour de l'exécution de notre programme. La taille du fichier XML est de 15,4 Mo.

Dans l'objectif de cerner les caractéristiques du corpus, nous procédons au dépouillement textométrique.

7.2 Dépouillement du corpus parallèle CLRG

Nous commençons le dépouillement par l'analyse des volumétries des données textuelles bilingues.

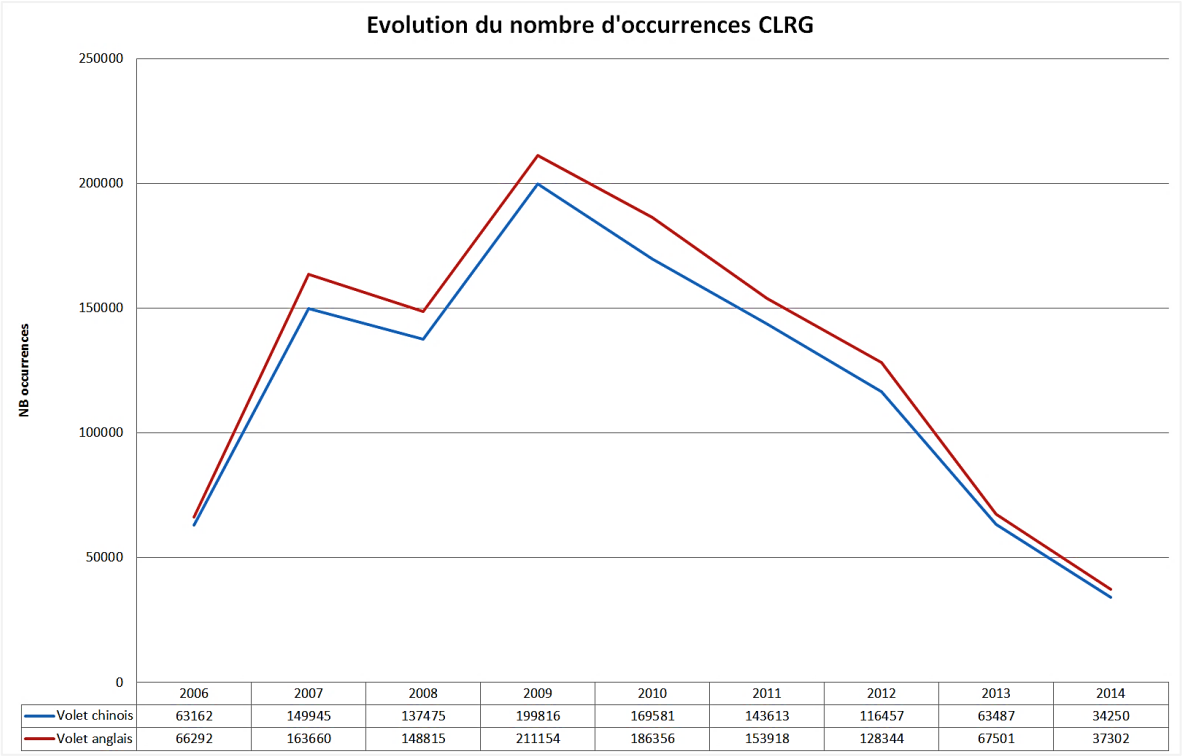
Tableau 7.1 CLRG de 2006 à 2014 : caractéristiques textométriques du corpus par année

	Volet chinois					Volet anglais			
	Occurrences	**Formes**	**Hapax**	**Fmax**		**Occurrences**	**Formes**	**Hapax**	**Fmax**
2006	63 162	8 508	4 516	5 316		66 292	7 829	3 641	3 571
2007	149 945	16 194	8 355	12 228		163 660	13 535	6 025	9 223
2008	137 475	15 104	7 717	10 841		148 815	13 026	5 818	8 267
2009	199 816	18 622	9 299	16 355		211 154	15 326	6 583	11 489
2010	169 581	16 962	8 496	13 652		186 356	14 107	6 040	10 159
2011	143 613	16 923	8 800	11 429		153 918	13 785	6 148	8 684
2012	116 457	14 968	7 817	8 825		128 344	12 329	5 539	6 764
2013	63 487	9 470	4 955	4 519		67 501	8 203	3 815	3 331
2014	34 250	6 573	3 589	2 507		37 302	6 259	3 198	1 853
Total	1 077 786	123 324	63 544	85 672		1 163 342	104 399	46 807	63 341

Le tableau 7.1 ci-dessus fait apparaître des variations du nombre d'occurrences, de formes et d'hapax entre le 6 juin 2006 et le 27 août 2014. Un constat immédiat se dégage, hormis le nombre d'occurrences toujours supérieurs en anglais, les nombres de formes, hapax et Fmax du volet chinois supplantent ceux du volet anglais. Selon la figure 7.2 ci-dessous, le nombre d'occurrences varie de 63 162 (CN) et 66 292 (EN) pour l'année 2006 pour atteindre un premier pic en 2007 à 149 945 (CN) et 163 660 (EN), puis le sommet est atteint en 2009, à savoir respectivement 199 816 (CN) et 211 154 (EN). Par la suite, les données textuelles chutent régulièrement jusqu'à 2014, date du lancement de notre programme. Cette chute peut s'expliquer par l'absence d'événements majeurs relatifs à l'environnement et au climat après l'Exposition Universelle de Shanghai en 2010.

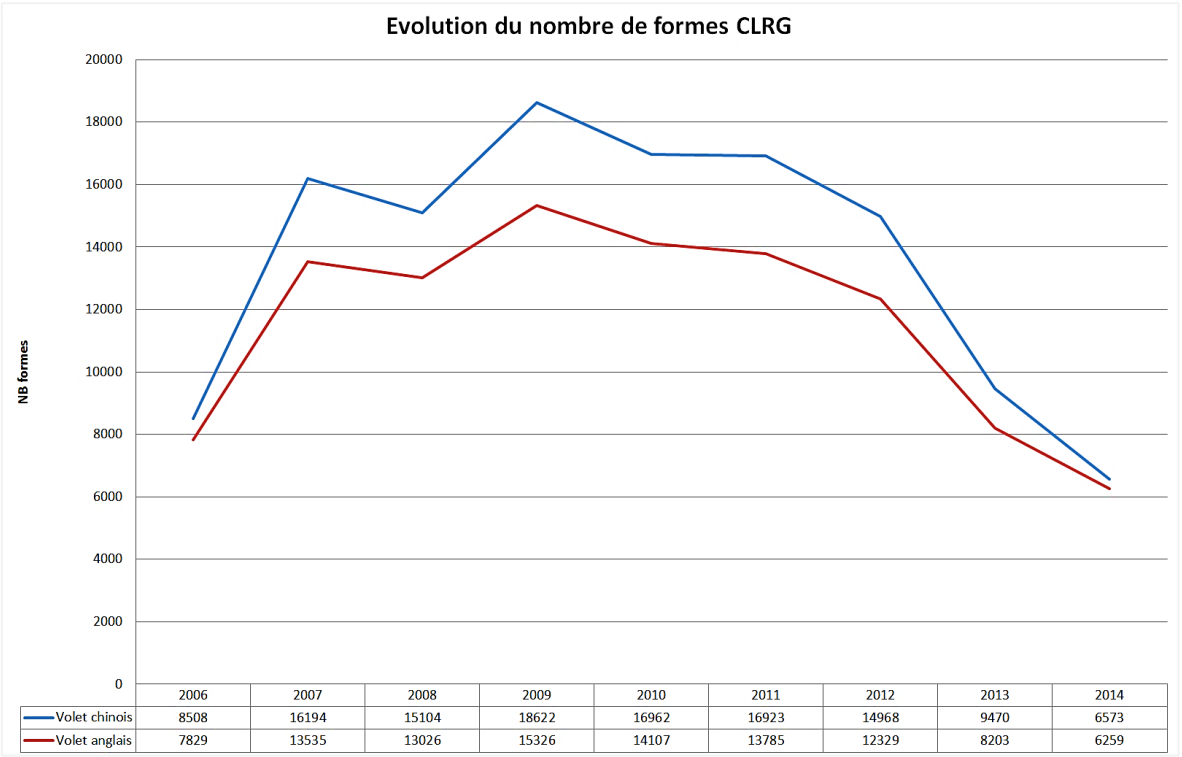
Nous allons nous intéresser au nombre d'occurrences et au nombre de formes de chacun des deux volets du corpus parallèle. D'une manière générale, le nombre d'occurrences est un indicateur de l'intensité de l'emploi des formes dans une quantité textuelle donnée, tandis que le nombre de formes est un indice de la richesse du vocabulaire employé dans un corpus. L'évolution des deux valeurs nous livre des informations intéressantes sur l'émergence textuelle des formes dans le continuum événementiel.

Des informations identiques émanant du corpus parallèle sont rédigées en au moins deux langues. Le fait de tracer les deux valeurs (nombre d'occurrences et nombre de formes) dans l'axe du temps et dans un seul tableau, nous montre immédiatement la différence du mécanisme langagier centrée sur l'emploi et la richesse du vocabulaire.



	2006	2007	2008	2009	2010	2011	2012	2013	2014
Volet chinois	63162	149945	137475	199816	169581	143613	116457	63487	34250
Volet anglais	66292	163660	148815	211154	186356	153918	128344	67501	37302

Figure 7.2 CLRG du 06/06/2006 au 27/08/2014 : répartition annuelle du nombre d'occurrences



	2006	2007	2008	2009	2010	2011	2012	2013	2014
Volet chinois	8508	16194	15104	18622	16962	16923	14968	9470	6573
Volet anglais	7829	13535	13026	15326	14107	13785	12329	8203	6259

Figure 7.3 CLRG du 06/06/2006 au 27/08/2014 : répartition annuelle du nombre de formes

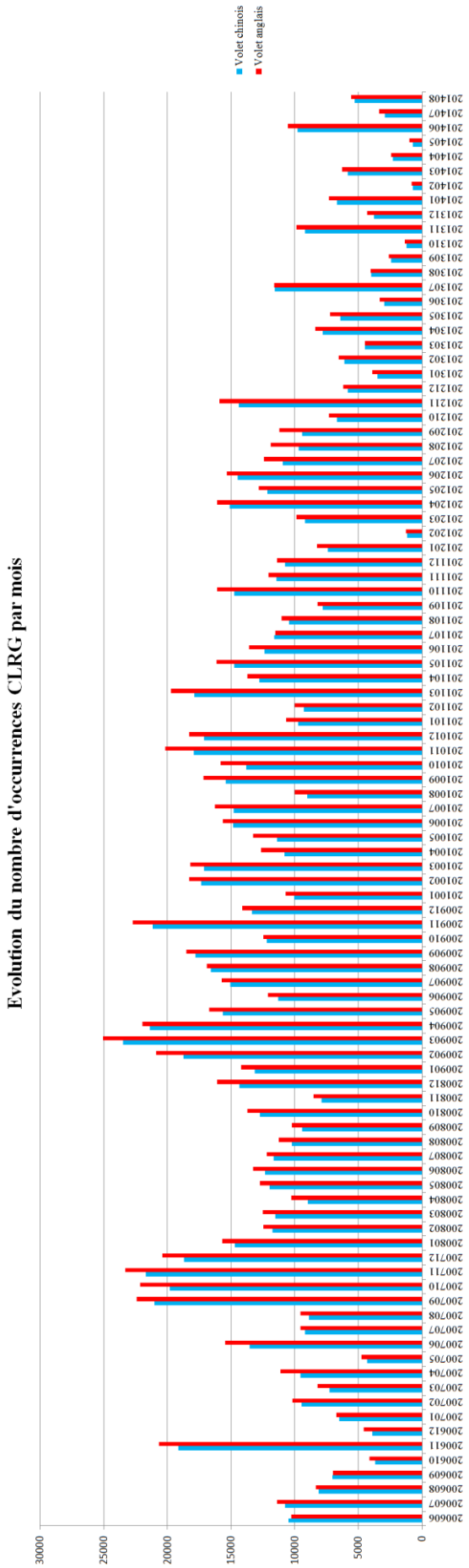


Figure 7.4 CLRG de 2006 à 2014 : répartition mensuelle du nombre d'occurrences

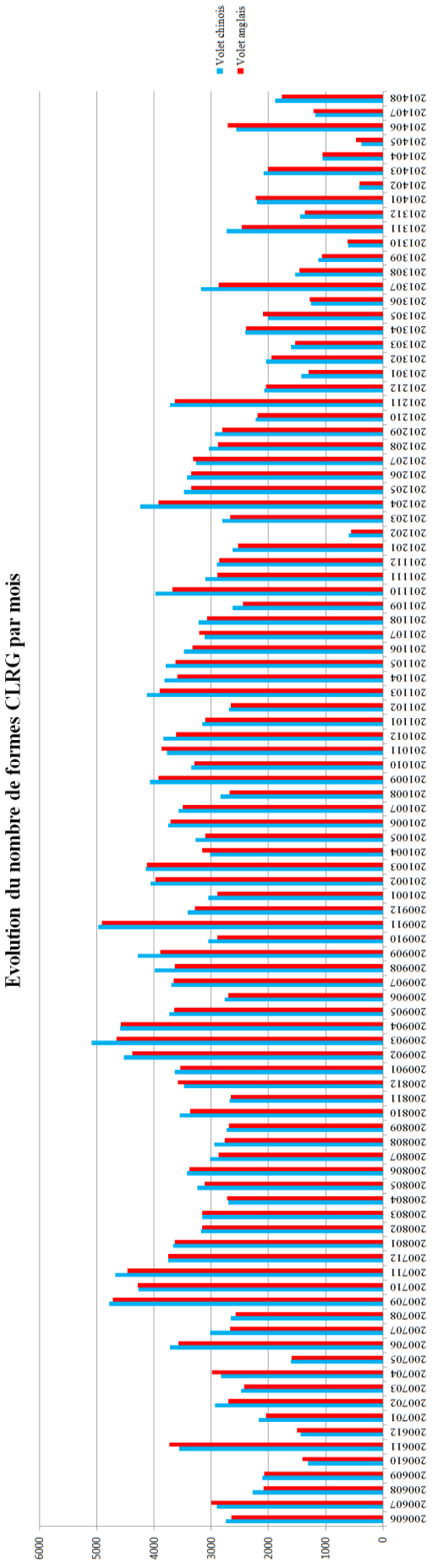


Figure 7.5 CLRG de 2006 à 2014 : répartition mensuelle du nombre de formes

Selon le tableau 7.1 et les figures 7.2, 7.3, les nombres d'occurrences et de formes dans les deux langues atteignent leurs valeurs maximales en 2009. Nous émettons l'hypothèse suivante : les pics constatés sur ces courbes « en dents de scie » sont synchronisés principalement avec les différentes sessions des Conférences des Parties[212](COP) sur le changement climatique. Nous vérifions cette affirmation à l'aide de la répartition mensuelle des occurrences (figure 7.4 ci-dessus) et des formes (figure 7.5 ci-dessus) dans chacun des deux volets. Les tables de données (nombre d'occurrences et nombre de formes) par mois sont disponibles dans l'annexe N, tableau N.1. Cependant, il existe des différences volumétriques ponctuelles dans l'évolution chronologique des productions de textes bilingues. Par exemple, une forte production de textes a été constatée en 2009 (Conférence de Copenhague) et en 2010 (Exposition Universelle de Shanghai).

De manière générale et constante, le tableau 7.1, les figures 7.2, 7.3, 7.4 et 7.5 montrent que le volet anglais présente le plus grand nombre d'occurrences, alors que le volet chinois a le plus grand nombre de formes. Ce phénomène étant très intéressant, des précisions seront apportées ultérieurement. Nous tenterons d'apporter un éclairage sur les questions suivantes :

- Pourquoi le nombre de formes est plus important en chinois ?
- Quelle est la cause du grand nombre d'occurrences en anglais ?
- Quelles sont les formes les plus répétées en anglais et en chinois ?

7.2.1 Apports linguistiques du chinois pour la veille textométrique d'informations

Cette abondance du nombre d'occurrences en anglais s'explique par le fait que la langue chinoise est connue pour sa nature concise et parfois confuse, autrement dit, cette langue est à la fois très compacte et polysémique par ses mécanismes morphosyntaxiques et lexicaux multimodaux (mélange de catégories grammaticales et modales) mais aussi multidimensionnels (absences de conjugaison et déclinaison). Par exemple, la notion 三峡工程/sānxiá gōngchéng/projet des Trois-Gorges ne compte qu'une seule forme ou occurrence. Or, cette notion en anglais ou en français compte au moins trois formes et trois occurrences : projet + des + Trois-Gorges, 3 formes en 3 occurrences, le mot *Trois-Gorges* étant lié par un trait d'union. Mais ce mot composé peut s'écrire parfois sans le trait d'union, dans ce cas, nous avons 4 formes en 4 occurrences : projet + des + Trois + Gorges. D'où, l'importance de la contextualisation qui par la référence chronologique va permettre d'interpréter avec pertinence le segment. Ce premier constat n'explique que partiellement pourquoi le nombre de formes est plus important dans le volet chinois. Nous montrerons plus loin que cette richesse de formes est due au choix de la segmentation de Jieba. Nous tenterons de livrer un début d'explication sur le nombre d'occurrences plus grand en anglais.

[212] Rappel des COP : Bali (COP13) en décembre 2007, Poznań (COP14) en décembre 2008, Copenhague (COP15) en décembre 2009, Cancún (COP16) en novembre - décembre 2010, Durban (COP17) en novembre - décembre 2011, Doha (COP18) en novembre - décembre 2012, Varsovie (COP19) en novembre 2013, Lima (COP20) en décembre 2014, Paris (COP21) en décembre 2015.

Il faut noter que les structures syntaxiques et grammaticales en anglais et en français sont relativement plus riches par rapport au chinois. Les mots grammaticaux ou encore les mots-outils sont souvent indispensables dans le fonctionnement de l'anglais et du français. La syntaxe de ces mots grammaticaux entraine la forte présence de séquences grammaticales répétitives dans l'usage de la langue. Nous montrons la différence syntaxique entre les deux langues par deux textes extraits du corpus CLRG. Les deux textes extraits d'articles du site *Chinadialogue* sont présentés dans un tableau avec leur équivalent anglais l'une après l'autre. Une traduction française en partant du chinois est proposée pour chacune de ces deux textes.

Tableau 7.2 CLRG 2012 : extrait d'un texte pour la comparaison syntaxique

En chinois	En anglais
我们 认为 , 投资 时 必须 首先 考虑 水资源 和 能源 风险 。 在 整个 投资 周期 中 , 水资源 供应 及其 对 供应链 的 潜在 影响 都 应该 是 至关重 要 的 考虑 因素 。	*We believe that water and power risks must be a top priority when planning capital expenditure. It is vital that the availability of water and the potential effect on supply chains is taken into account for the life of the investment.*
Traduction française depuis le chinois :	
Nous pensons qu'il faut avant tout mesurer les risques liés à la ressource hydrique et aux énergies lors des investissements. Dans l'ensemble des phases du cycle d'investissement, la disponibilité de la ressource en eau ainsi que les impacts potentiels de la chaîne logistique relatifs à cette ressource doivent tous deux être les facteurs les plus cruciaux à considérer.	

Le texte chinois du tableau 7.2 ci-dessus est segmenté en mots où le blanc de séparation est marqué par un trait gris « ». Ce texte provient d'un article[213] rédigé en chinois puis traduit en anglais dont le titre[214] en français (titre traduit depuis le chinois) est « L'expansion de la capacité de production d'énergie exacerbe les pénuries d'eau en Chine ». Afin d'illustrer la différence syntaxique de ce texte chinois par rapport à sa traduction anglaise (tableau 7.2), nous procédons à la segmentation de ces deux textes en séquence[215]. Cette structure de représentation est appelée parfois trame[216] (Martinez, 2012). Nous obtenons le tableau suivant (tableau 7.3, ci-dessous). Les deux textes du tableau ci-dessous sont à lire de manière verticale.

[213] L'article est disponible : https://www.chinadialogue.net/article/show/single/ch/5198-Does-China-have-enough-water-to-keep-building-three-power-stations-a-week- (consulté le 06/12/2016)
[214] La version anglaise du titre de l'article est : Does China have enough water to keep building three power stations a week ?
[215] Définition en linguistique : Suite d'unités linguistiques ordonnée de gauche à droite sur l'axe syntagmatique et formant une unité textuelle. http://www.cnrtl.fr/definition/s%C3%A9quence (consulté le 06/12/2016)
[216] Définition en informatique : Bloc d'informations, composé et transmis selon un ensemble de règles constituant une procédure synchrone de contrôle de liaison de données. http://www.cnrtl.fr/definition/trame (consulté le 24/12/2016)

Tableau 7.3 CLRG 2012 : texte bilingue exprimé en séquence

N°	Séquence en chinois	Equivalent français	Séquence en anglais	Equivalent français
1	我们	nous	We	nous
2	认为	penser	believe	croyons
3	,	,	that	que
4	投资	investissement	water	eau
5	时	quand	and	et
6	必须	il faut	power	énergie
7	首先	première / premièrement	risks	risques
8	考虑	considérer	must	devoir
9	水资源	ressource hydrique / ressource en eau	be	être
10	和	et	a	un/une
11	能源	énergies	top	sommet
12	风险	risque	priority	priorité
13	。	.	when	quand
14	在	dans	planning	planification
15	整个	ensemble	capital	capital
16	投资	investissement	expenditure	dépense
17	周期	cycle	.	.
18	中	au milieu de ou dans	It	Il
19	,	,	is	est
20	水资源	ressource hydrique / ressource en eau	vital	vital
21	供应	alimentation	that	cette
22	及其	et sa/son/ses/leur(s)	the	la
23	对	à/relatif à/concernant	availability	disponibilité
24	供应链	la chaîne logistique	of	de
25	的	de	water	eau
26	潜在	potentiel	and	et
27	影响	affecter	the	la
28	都	tous	potential	potentiel
29	应该	devoir / il faut	effect	effet
30	是	être	on	sur
31	至关重要	crucial / essentiel	supply	approvisionnement
32	的	de	chains	chaînes
33	考虑	considérable / (à) considérer	is	est
34	因素	facteur	taken	pris
35	。	.	into	dans
36			account	compte
37			for	pour
38			the	la
39			life	vie

40				of	de
41				the	la
42				investment	investissement
43				.	.

Selon le tableau 7.3 ci-dessus, le découpage séquentiel des deux textes montre immédiatement que les termes grammaticaux tels que *that, the, for, is*, *of* (cases colorées en bleu), sont très répétés dans le fonctionnement de l'anglais. La répétition de ces termes est l'une des caractéristiques de cette langue. En revanche en chinois, ce type de mots grammaticaux est quasi absent et remplacé par des mots de contenu.

Pour étayer notre propos, nous regardons un deuxième exemple extrait de l'article[217] dont le titre[218] traduit en français est «Monsieur PAN Jiazheng, architecte en chef du Barrage des Trois Gorges et un des représentants du pro-développement hydroélectrique en Chine. Sa réflexion sur l'impact négatif du projet des Trois Gorges n'a pas affecté son enthousiasme pour les grandes centrales hydroélectriques. Avec le départ progressif de la première génération de personnes du secteur hydroélectrique, la Chine va-t-elle se livrer à l'introspection de ces projets ? Analyse de Monsieur LI Dun ». La phrase extraite de l'article est présentée dans le tableau 7.4 ci-dessous avec ses traductions anglaise et française.

Tableau 7.4 CLRG 2012 : extrait d'un deuxième texte pour la comparaison syntaxique

En chinois	En anglais
我们 正 处于 一个 相当 多人 " 在 科学 的 名义 下 迷信 技术 , 在 市场 的 名义 下 迷信 金钱 " 的 时代 。	*We live in an era in which many people – in the name of science – have blind faith in technology; and in the name of markets have blind faith in money.*
Traduction française depuis le chinois :	
Nous vivons dans une ère où un nombre assez important de personnes croient aveuglement aux technologies au nom de la science et au culte du Veau d'or (« à l'argent ») au nom des marchés.	

Ce deuxième texte a subi le même traitement que le précédent. Nous obtenons le tableau 7.5 ci-dessous.

[217] L'article est disponible : https://www.chinadialogue.net/article/show/single/ch/5182-Death-of-Three-Gorges-Dam-architect-marks-end-of-era (consulté le 06/12/2016)

[218] La version anglaise du titre de l'article est : The death this summer of controversial figure Pan Jiazheng offers a moment to reflect on China's relationship with big dams.

Tableau 7.5 CLRG 2012 : deuxième texte bilingue exprimé en séquence

N°	Séquence en chinois	Equivalent français		Séquence en anglais	Equivalent français
1	我们	nous		We	nous
2	正	positif		live	vivre
3	处于	dans		in	dans
4	一个	a		an	un/une
5	相当	assez		era	ère
6	多人	beaucoup de gens		in	dans
7	"	"		which	lequel
8	在	dans		many	beaucoup
9	科学	science		people	gens
10	的	de		–	-
11	名义	nom		in	dans
12	下	inférieur		the	le
13	迷信	croire aveuglement		name	nom
14	技术	technologie		of	de
15	，	,		science	science
16	在	dans		–	-
17	市场	marché		have	avoir
18	的	de		blind	aveugle
19	名义	nom		faith	foi
20	下	inférieur		in	dans
21	迷信	croire aveuglement		technology	technologie
22	金钱	argent		;	;
23	"	"		and	et
24	的	de		in	dans
25	时代	ère		the	la
26	。	.		name	prénom
27				of	de
28				markets	marchés
29				have	avoir
30				blind	aveugle
31				faith	foi
32				in	dans
33				money	argent
34				.	.

Avec ce deuxième texte, nous arrivons aux mêmes conclusions que celles du premier texte : le découpage séquentiel des deux phrases en deux langues (tableau 7.5 ci-dessus) montre encore une fois que le fait de répéter les termes grammaticaux tels que *in which*, *in, the* (cases colorées en bleu) constitue un trait saillant dans l'usage de la langue anglaise. Or, la répétition de formes de contenu en chinois est sa caractéristique marquante.

Ce phénomène est attesté par les calculs des segments répétés issus des deux volets de CLRG (*cf.* tableau 7.8 segments les plus répétés ci-après). Selon les résultats de ce tableau 7.8, nous remarquons qu'en chinois, les termes de contenu sont classés en haut du tableau, tandis qu'en anglais, ce sont des mots-outils ou des mots syntaxiques qui sont les plus répétés. Nous en déduisons qu'il s'agit de deux types de répétitions, d'une part de mots grammaticaux en anglais, et d'autre part, de mots de contenu en chinois. Cette forte répétition de mots grammaticaux est la cause du grand nombre d'occurrences en anglais. Plus l'emploi des mots grammaticaux est répétitif, plus le nombre d'occurrences est grand.

Le mécanisme de la formation des mots chinois et le choix du segmenteur Jieba entrainent une prépondérance du nombre de formes et d'hapax dans CLRG_CN. En effet, le corpus parallèle permet de générer et d'enrichir des ressources de lexiques dictionnairiques et traductologiques. Nous montrons un extrait des dictionnaires générés (tableau 7.6 ci-dessous) par CLRG lors de nos traitements textométriques.

Tableau 7.6 CLRG de 2006 à 2014 : extrait des dictionnaires générés par le corpus

N°	Fréquence	Forme	Equivalent français	Fréquence	Forme	Equivalent français
1	85672	的	de	63560	the	la
2	16979	在	dans	36591	of	de
3	13200	和	et	32401	to	à
4	11953	是	il est	32393	and	et
5	11893	了	la	22911	in	dans
6	8453	中国	Chine	21065	a	un/une
7	6237	我们	nous	16322	is	est
8	5583	年	année	13209	that	que ou ce/cet/cette
9	5509	对	à	10906	for	pour
10	5322	也	aussi	9716	s	de quelqu'un
11	5122	将	auxiliaire du futur	8320	are	sont
12	5097	一个	un (+ classificateur)	8223	on	sur
13	4839	中	dans	7314	the	le
14	4515	上	sur	7261	be	être
15	4505	有	avoir	7153	China	Chine
16	4321	都	tous	6785	as	comme
17	4092	就	alors	6756	it	il
18	4049	气候变化	changement climatique	6590	by	par
19	3934	这	ce/cet/cette	5851	have	avoir
20	3815	到	arriver à/jusqu'à	5687	with	avec
21	3803	问题	problème	5584	will	auxiliaire du futur / volonté
22	3660	而	alors/encore	5374	from	de
23	3589	国家	pays	5336	has	a
24	3376	与	et	4796	at	à
25	3269	发展	développement	4640	climate	climat

Le tableau 7.6 ci-dessus illustre les 25 premières formes les plus fréquentes de chacun des volets de CLRG. Selon le constat de ce tableau, il convient de noter que ce sont les termes de contenu qui se répètent le plus en chinois, alors qu'en anglais ce sont les termes grammaticaux.

D'une part, les termes obtenus du corpus parallèle constituent et enrichissent des ressources lexicales bilingues entre le chinois et l'anglais, car il s'agit de textes traduits de l'un à l'autre. Parmi les 25 termes illustrés ci-dessus, nous avons par exemple les correspondances suivantes : 中国/ zhōngguó pour la forme *China* ; 将/ jiāng /auxiliaire du futur pour *will*. D'autre part, des termes d'hapax (termes apparus une seule fois dans le corpus) permettent d'enrichir des termes parfois rares ou non référencés dans les dictionnaires classiques dans chacune des langues du corpus, comme des néologies[219], des emprunts[220] lexicaux, etc. Nous citons quelques exemples ci-dessous :

termes issus du volet chinois :

- 文会 / wén huì / rencontre littéraire
- 堰塞坝 / yàn sāi bà / barrage/lac naturel
- 风投 / fēng tóu / capital risque (venture capital en anglais)
- 鲥鱼 / shí yú / *tenualosa reevesii*, une sous famille des aloses
- 灶烧 / zào shāo / foyer de cuisson à la campagne
- 值得反思 / zhídé fǎnsī / qui mérite de la réflexion

termes issus du volet anglais :

- *bedding* : literie / matériel de couchage
- *branched* : ramifié / bifurqué
- *meaty* : de viande ou qui relève de la viande
- *unfazed* : imperturbable / impassible
- *blankly* : d'une manière qui ne montre aucune compréhension, aucun intérêt ou aucune émotion / complètement / absolument
- *personae* : pluriel de *persona*

Par ailleurs, des notions telles que *environnement écologique* ou *système écologique*, parfois appelées *écosystème* ou encore *éco-système*, se traduisent en chinois :

- par la forme 生态环境/shēngtài huánjìng/environnement écologique, dont le premier segment est 生态/shēngtài/écologie et le deuxième est 环境/huánjìng/environnement,
- et par la forme 生态系统/shēngtài xìtǒng/système écologique, dont le premier segment est 生态/shēngtài/écologie et le deuxième segment est 系统/xìtǒng/système.

Il est facile de constater que dans ce corpus ces deux notions 生态环境/shēngtài huánjìng/environnement écologique et 生态系统/shēngtài xìtǒng/système écologique ne comptent que deux formes et deux occurrences par rapport au français et à l'anglais, nonobstant que la composition lexicale de ces deux notions en chinois soit complexe voire composée morphologiquement. Or, pour exprimer strictement les mêmes informations, le français et l'anglais comptent trois formes et trois occurrences. Il est à noter que ce mécanisme de formation de mots en chinois par bloc de quatre caractères peut être considéré en français et en anglais comme un segment répété, par exemple, 生态环境/shēngtài huánjìng/environnement écologique.

[219] Définition en linguistique : Processus de formation de nouvelles unités lexicales. http://www.cnrtl.fr/definition/n%C3%A9ologie (consulté le 13/12/2016)
[220] Définition en linguistique : Fait pour une langue d'incorporer une unité linguistique, en particulier un mot d'une autre langue. http://www.cnrtl.fr/definition/emprunt (consulté le 13/12/2016)

D'ailleurs, dans le volet chinois du corpus parallèle, un grand nombre de notions chinoises n'ont pas été segmentées par le module Jieba. Nous pouvons en citer quelques-unes :

- 环境污染/huánjìng wūrǎn/pollution de l'environnement :
  环境/huánjìng/environnement + 污染/wū rǎn/pollution, nom + nom/verbe
- 环境退化/huánjìng tuìhuà/dégradation de l'environnement :
  环境/huánjìng/environnement + 退化/tuìhuà/dégradation, nom + nom/verbe
- 生物学家/shēngwù xué jiā/biologiste :
  生物/shēngwù/biologie + 学家/xué jiā/spécialiste, affixe, -iste, nom + affixe/lexème
- 可行性研究/kěxíng xìng yánjiū/étude de faisabilité :
  可行性/kěxíng xìng/faisabilité + 研究/yánjiū/étude, nom + nom ou adjectif + affixe + nom
- 核心技术/héxīn jìshù/technologie centrale ou clé :
  核心/héxīn/cœur ou noyau + 技术/jìshù/technologie, nom + nom
- 王先生/wáng xiānshēng/Monsieur Wang :
  王/wáng/nom de famille + 先生/xiānshēng/Monsieur, nom (de famille) + nom
- 信息技术/xìnxī jìshù/technologie de l'information :
  信息/xìnxī/information + 技术/jìshù/technologie, nom + nom
- 几千年/jǐ qiān nián/quelques milliers d'années :
  几/jǐ/certain, adjectif indéfini + 千/qiān/mille + 年/nián/année, adjectif indéfini + unité + nom (année)
- 密切合作/mìqiè hézuò/collaboration étroite :
  密切/mìqiè hézuò /étroite + 合作/hézuò/collaboration, adjectif + nom ou adverbe + verbe

Ces genres de notions citées ci-dessus augmentent l'apparition des formes individuelles, cependant, celles-ci peuvent être complexes/composées, c'est-à-dire qu'elles peuvent être scindées en deux ou parfois en trois mots chinois, provoquant l'exubérance des nombres de formes et d'hapax dans le volet chinois. C'est la raison pour laquelle le nombre de formes est plus important dans le volet chinois. Il est à noter que nous retrouvons ce type de formation de mots également en anglais. Par exemple, *risk analysis*, *crisis management, etc.* au lieu de *analysis of risk* et *management of crisis*.

Ce phénomène est dû au choix du segmenteur Jieba. En effet, ce type de segmentation dit composé ou complexe est plus facile à comprendre par la clarté des unités sémantiques pour un natif. Tandis que si ces mots avaient été segmentés de manière très fine, c'est-à-dire, très proche des morphèmes lexicaux, des confusions ou incompréhensions risqueraient d'être apportées lors des interprétations des segments pour lesquels les retours aux contextes deviennent indispensables et incontournables.

Les choix de segmentation jouent un rôle primordial dans toutes les veilles et recherches d'informations en chinois, en particulier dans le domaine du filtrage des informations par mots-clés. Plus les textes sont segmentés finement, plus seront coûteuses les interprétations et les restitutions, car des masses d'information seront récoltées, y compris les bruits et les hors sujets.

Maintenant, nous allons étudier la répartition annuelle et mensuelle du nombre d'articles du corpus CLRG de 2006 à 2014.

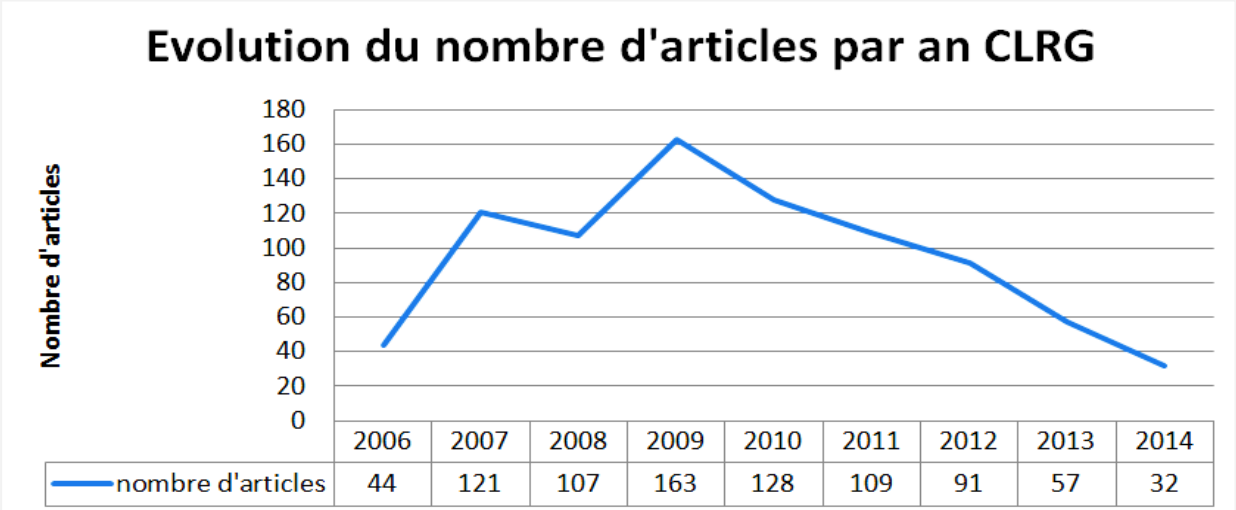


Figure 7.6 CLRG de 2006 à 2014 : répartition annuelle du nombre d'articles

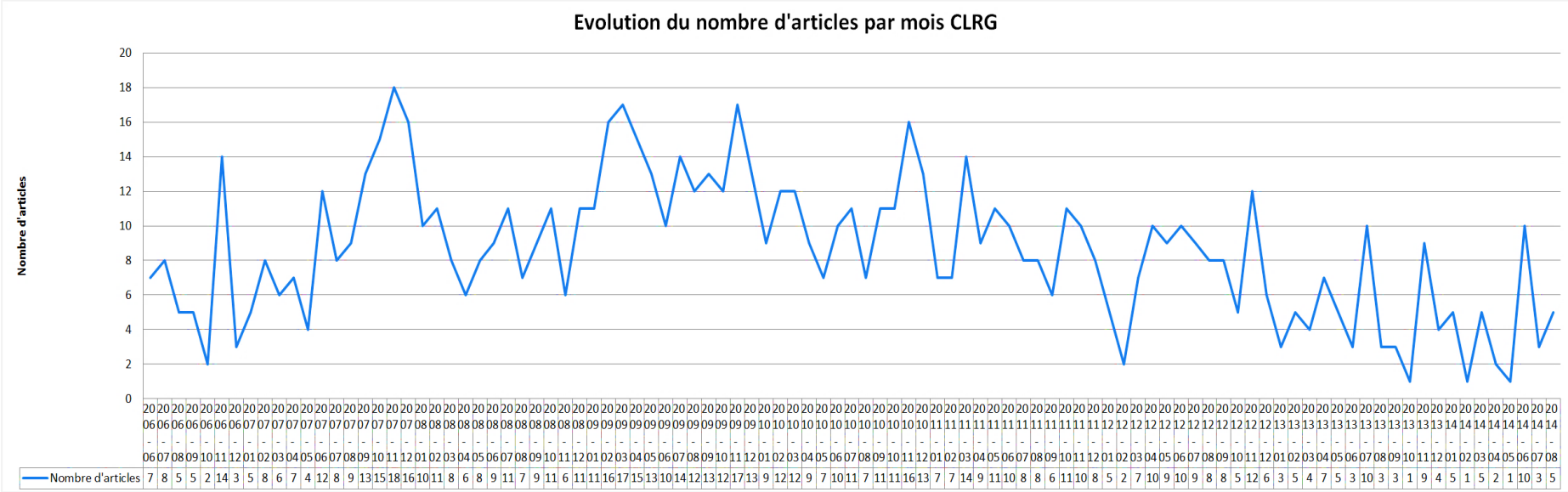


Figure 7.7 CLRG de juin 2006 à août 2014 : répartition mensuelle du nombre d'articles

La courbe annuelle de la figure 7.6 ci-dessus affirme une tendance générale à la baisse en matière de production d'articles après l'année 2009. Cette tendance peut s'expliquer par des déconvenues, par exemple, l'absence de prise de décision concète au terme de la Conférence de Copenhague dont l'un des thèmes majeurs portait sur le réchauffement climatique. Par la suite, la dénomination du thème « réchauffement climatique » change et semble devenir plus restrictive, puisque l'on ne parlera plus que de changement climatique. Ce changement de thème peut également entraîner un impact sur l'organisation des sites des médias.

La forme « en dents de scie » de la courbe de la figure 7.7 ci-dessus exprimant la tendance de production mensuelle d'articles varie en fonction des COP internationales, ainsi que d'autres événements climatiques ponctuels. Cette production périodique ponctuée de « pics » prouve que les ONG suivent de très près les événements climatiques.

7.2.2 Accroissements comparés du vocabulaire de CLRG

Les deux diagrammes d'accroissement de vocabulaire (figures N.16 et N.17 disponibles dans l'annexe N) mettent en évidence l'apparition de nouvelles formes au fur et à mesure de l'avancement dans le temps dans les deux volets de CLRG.

L'ensemble du volet CLRG_CN compte 1 077 786 occurrences et 123 324 formes (selon le tableau 7.1 ci-dessus), le renouvellement de formes se stabilise après 450 000 occurrences. Par la suite, pour chaque intervalle de 50 000 occurrences supplémentaires, le nombre de formes augmente de 2 000 environ.

L'ensemble du volet CLRG_EN compte 1 163 342 occurrences et 104 399 formes (selon le tableau 7.1 ci-dessus), le renouvellement de formes se stabilise après 400 000 occurrences. Par la suite, pour chaque intervalle de 50 000 occurrences supplémentaires, le nombre de formes augmente de 1 000 environ.

Rappelons que, de manière factuelle, dans le corpus parallèle, le nombre de formes du volet chinois est supérieur à celui de l'anglais, mais c'est le cas contraire pour le nombre d'occurrences. Nous pouvons en déduire que le vocabulaire est plus riche en chinois, mais que les textes sont plus longs en anglais. Pour le même corpus, le renouvellement des formes se stabilise plus tardivement en chinois qu'en anglais. Ceci renforce notre première déduction sur la richesse des formes en chinois par rapport à l'anglais. Toutefois, cette richesse du vocabulaire en chinois étayée par la richesse de formes ne reflète pas entièrement la réalité linguistique. En effet, la richesse de formes en chinois se joue dans la combinaison des caractères chinois, ainsi des nouvelles formes se créent par l'agglutination de mots. Or, en anglais, excepté les formes de pluralité et les affixes grammaticaux, les nouveaux mots se distinguent par leurs radicaux linguistiques. Ce phénomène relève du choix du segmenteur Jieba.

Après les analyses volumétriques de CLRG, la typologie textuelle va nous révéler les proximités de chaque partie du corpus.

7.2.3 Typologie textuelle de CLRG

Dans cette section, nous allons présenter entre autres deux AFC distinctes sur le corpus CLRG, l'une pour les années et l'autre pour les mois.

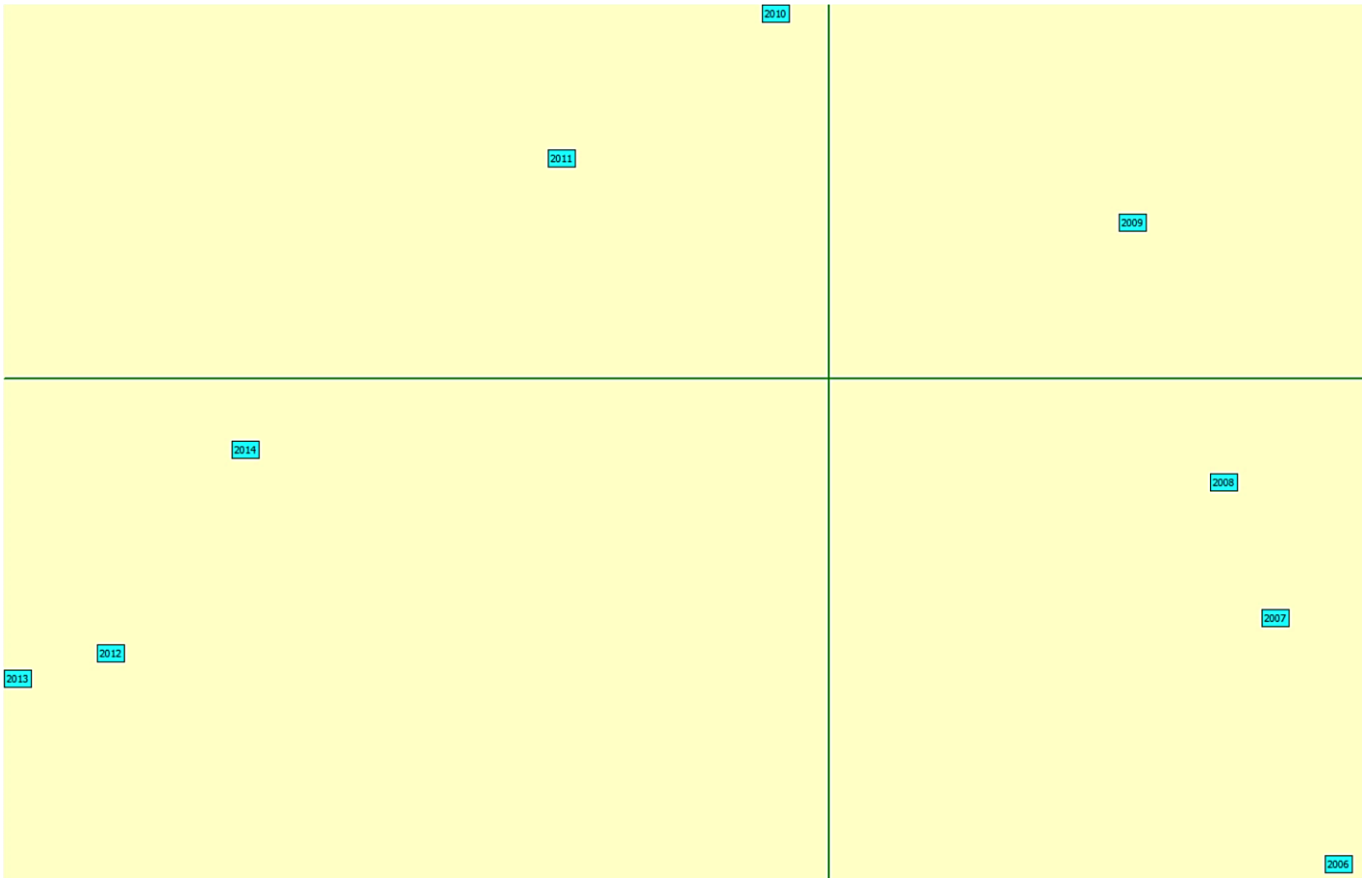


Figure 7.8 CLRG_CN de 2006 à 2014 : analyse factorielle des correspondances par année

D'après la figure 7.8, l'AFC annuelle de CLRG_CN sur l'ensemble de la période présente une parabole avec une série chronologique presque complète, à l'exception des années 2012, 2013 et 2014 où l'ordre n'a pas été respecté. Rappelons que la forme de cette parabole chronologique s'appelle l'effet Guttman (se reporter au chapitre 4, section 4.5.4). Afin de faciliter la lecture de l'AFC, nous regroupons les années par les quatre blocs formés par les deux axes.

Groupement des années par la proximité sur le plan de l'AFC :

- en haut à gauche : 2010 et 2011 ; en haut à droite : 2009
- en bas à gauche : 2012, 2013 et 2014 ; en bas à droite : 2006, 2007 et 2008

Il faut noter qu'une fois de plus, ce sont des groupements obtenus de manière empirique. Il s'agit juste d'une lecture des quadrants de l'AFC dont on n'a pas proposé de motivation d'une importance particulière.

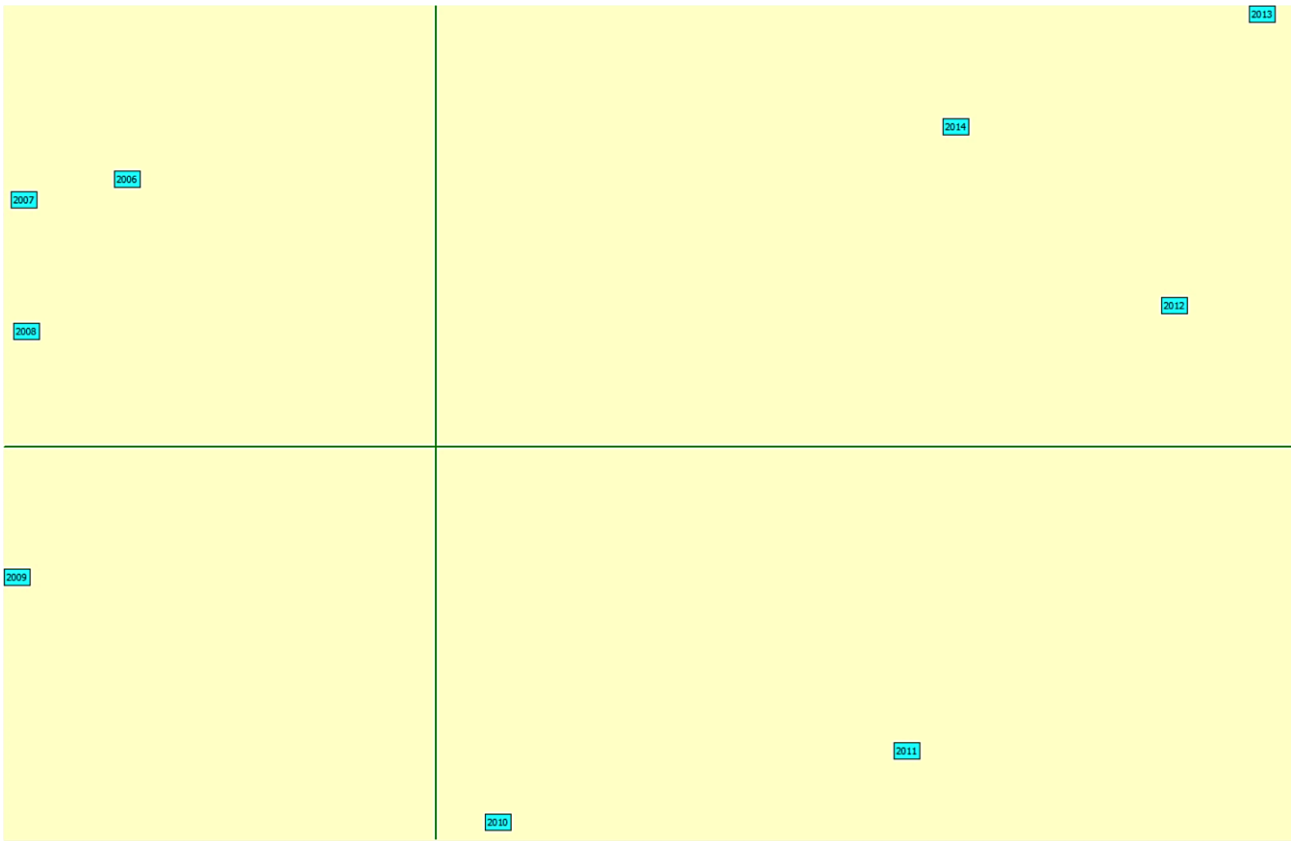


Figure 7.9 CLRG_EN de 2006 à 2014 : analyse factorielle des correspondances par année

Selon la figure 7.9, le même « *mapping* » de CLRG_EN sous la forme parabolique est moins régulier que celui de CLRG_CN, autrement dit, la courbe présentée dans la figure 7.8 (CLRG_CN) respecte mieux la norme canonique d'une parabole. L'évolution chronologique des années sur la parabole atteste l'emploi discursif et récurrent du vocabulaire entre 2006 et 2014. Plus la courbe parabolique est régulière, plus les thèmes du discours du corpus sont récurrents. Nous pouvons donc dire que la récurrence des thèmes est mieux appréhendée dans le volet chinois. Nous pouvons regrouper les années de la façon suivante :

- en haut à gauche : 2006, 2007 et 2008 ; en haut à droite : 2012, 2013 et 2014
- en bas à gauche : 2009 ; en bas à droite : 2010 et 2011

La courbe parabolique du volet chinois formée par les nuages de mots groupés sur le plan de l'AFC est à l'inverse de celle du volet anglais, ceci est dû à la différence des fréquences des covariances des formes lors des calculs de l'AFC dans chacun des deux volets, un phénomène mathématique normal. Il faut noter que cette orientation spatiale dans le premier plan de l'AFC n'a pas de pertinence particulière. En revanche, la dissymétrie des deux courbes atteste que les formes lexicales du volet anglais sont moins récurrentes que celles du volet chinois, autrement dit, l'emploi du vocabulaire chinois est plus chronologiquement répétitif que celui de l'anglais, comme l'explique l'effet Guttman (se reporter au chapitre 4, section 4.5.4). Nous rappelons qu'il s'agit d'un corpus parallèle où les informations relayées dans les deux volets sont à priori identiques. L'un des objectifs de notre comparaison est de relever les comportements langagiers de ces deux langues face aux mêmes informations.

Spécificités évolutives du volet chinois CLRG_CN

Tableau 7.7 CLRG_CN de 2006 à 2014 : extrait et synthèse des spécificités évolutives

2006, 2007, 2008					2009				2010, 2011				2012, 2013, 2014			
Forme	Equivalent français	Frq. Tot.	Fréqu	Coeff.	Forme	Equivalent français	Frq. Tot.	Fréq	Coeff.	Forme	Equivalent français	Frq. Tot.	Fréq	Coeff.	Forme	Equivalent français	Frq. Tot.	Fréqu	Coeff.
0	0	266	244	***	炭	charbon	56	53	36	坎昆	Cancun	91	90	48	中国	Chine	8453	2532	***
2	2	150	134	48	哥本哈根	Copenhague	529	214	32	事故	accident	227	149	30	缅甸	Myanmar	179	129	***
和	et	13200	5061	46	冰盖	inlandsis/calotte polaire/banquise	104	71	29	部落	tribu	92	78	29	页岩	schiste	186	145	***
变暖	réchauffement	673	384	39	喜马拉雅	Himalaya	163	94	29	及	et	630	302	24	2011	2011	226	145	48
二氧化碳	dioxyde de carbone	808	422	31	海地	Haïti	47	43	28	航空业	industrie de l'aviation	73	62	23	气	gaz	123	96	44
1	1	96	85	30	HDI	HDI (Human Development Index)	35	35	27	素食	nourriture végétarienne	47	45	22	煤制气	gaz de houille	57	57	41
一个	un	5097	2026	28	REDD	REDD (Reducing Emissions from Deforestation and Forest Degradat	87	62	27	NGO	ONG	164	106	21	天然气	gaz naturel	475	222	40
全球	mondial	2616	1115	28	葡萄牙	Portugal	55	46	26	法案	projet de loi	178	111	21	煤火	feu de mine	54	54	39
斯特恩	Nicholas Stern (Économiste)	122	98	28	碳	carbone	2388	641	24	漏油	déversement de pétrole	53	47	20	试点	projet pilote	141	99	38
5	5	80	72	27	尼泊尔	Népal	224	105	22	游牧	nomade	45	42	20	企业	entreprise	1442	491	37
建筑	bâtiment	540	297	27	南登	Nandom district (dans le nord du Ghana)	28	28	21	ICT	information and communication technolog	33	33	19	石油	pétrole	1174	407	33
气候变化	changement climatique	4049	1632	26	尼奥	Alto Minho (au Portugal)	28	28	21	2010	2010	383	192	19	密松	Myitsone (barrage de)	75	63	33
南方	Sud	116	90	24	哈多	Hadow (nom de famille)	28	28	21	牧民	gardiens de troupeau	97	70	19	缅	Myanmar	65	58	33
京都	Kyoto	473	259	24	戈尔	Albert « Al » Arnold Gore (45e vice-président des États-Unis)	112	64	20	金太阳	Golden Sun (projet photovoltaïque)	30	30	17	煤炭	charbon	630	255	33
9	9	51	49	22	金融危机	crise financière	92	55	19	来自	venir de / provenance	776	334	17	德国	Allemagne	417	189	32
我们	nous	6237	2382	22	青藏高原	Plateau Qinghai-Tibet	103	60	19	光热	photothermique	37	35	17	油气	pétrole et gaz	134	86	29
6	6	48	46	21	CCS	carbon capture and storage (or carbon capture and sequestration)	193	88	18	三角洲	delta	146	90	17	垃圾焚烧	incinération des déchets	92	68	29
议定书	protocole	468	248	20	经济	économie	2388	601	16	核	nucléaire	122	79	17	项目	projet	2032	611	28
4	4	55	50	20	首页	page d'accueil	442	152	16	2009	2009	390	191	17	电动车	voitures électriques	63	53	28
2007	2007	426	230	20	土壤	sol	190	83	16	切尔诺贝利	Tchernobyl	57	47	17	管道	pipeline	135	84	27
知识产权	propriété intellectuelle	128	91	19	木炭	charbon de bois	25	23	16	哥本哈根	Copenhague	529	243	17	园区	parc	36	36	26
华林	Hualin (prénom)	37	37	19	引爆	détoner ou détonation	42	31	15	核电	électricité nucléaire	380	183	16	拆除	démolition	61	50	25
3	3	58	51	19	江村	Erda (prénom tibétain)	19	19	15	然而	cependant	1129	452	16	苏丹	Soudan	92	62	23
汪	Wang (nom de famille)	35	35	18	摄	photographie	327	120	15	其它	autre	559	251	16	火电	électricité thermique	90	60	22
风力	énergie éolienne	406	216	18	合作	coopération	897	259	15	巴基斯坦	Pakistan	181	102	15	米茨	Willie Smits	30	30	22
2006	2006	289	165	18	旺扎	Lin (Erda) (nom tibétain)	18	18	14	湖	lac	58	45	15	军政府	gouvernement militaire	51	42	22
巴厘岛	Bali	95	71	17	希	Isabel Hilton (CEO and Editor de Chinadialogue)	121	58	14	达赛湖	lac Dalai	26	26	15	2012	2012	294	132	22
了	le (particule grammaticale)	11893	4274	17	办公	bureau	58	37	14	生育	naissance	33	31	15	2015	2015	129	74	21
布什	George Bush	81	62	16	地区	région	1876	474	14	十二五	12ème plan quinquennal	124	77	15	封存	captage (du CO2)	94	61	21
气体	gaz	1098	485	16	碳源	provenance du carbone (CO2)	16	16	13	节能灯	ampoules à base consommation	40	35	15	克钦	Kachin (État Kachin)	37	33	20
environnemen	environnement	41	38	16	气候变化	changement climatique	4049	930	13	生态农业	agriculture écologique	30	29	15	但	mais	3038	806	19
温室	effet de serre	1044	463	16	奥巴马	Obama	232	89	13	EB	Executive Board, EB	30	29	15	潘家铮	Pan Jiazheng (ingénieur)	26	26	19
食品	nourriture	202	122	16	国	pays	399	133	13	CDM	Clean Development Mechanism, CDM	134	80	14	2013	2013	63	45	19
燃烧	combustion	191	114	15	冰川	glacier	717	209	13	核电站	centrale nucléaire	311	153	14	合作社	coopérative	38	33	19
录音棚	studio d'enregistrement	29	29	15	情景	situation	48	31	13	晶科	Jinko Solar (nom de société)	28	27	14	堆年	année (pour un réacteur)	26	26	19
能耗	consommation d'énergie	246	138	15	协议	accord	820	235	13	墨西哥湾	Golfe du Mexique	41	35	14	开采	exploitation	245	107	18
摄	photographie	327	173	15	坦提	Tanti (Tulsi Tanti, fondateur de Suzlon Energy en Inde)	16	16	13	粉煤灰	cendres volantes (de centrales thermiques)	28	27	14	核电	électricité nucléaire	380	148	18
英国	Anglais	777	353	15	经济危机	crise économique	64	38	13	疫苗	vaccin	29	27	13	韩国	Corée	83	52	18
采暖	chauffage	41	37	15	减排	émissions	1573	400	12	福岛	Fukushima	126	75	13	反应堆	réacteur	184	89	18
发展中国家	pays en voie de développemen	915	404	14	融化	fondre	290	101	12	瓜州	Guazhou (district administratif du Gansu)	23	23	13	亏损	perte	55	41	18
世界	le monde	2169	867	14	煤制油	liquéfaction du charbon	20	18	12	让	permettre de	1414	534	13	女性	femmes	41	34	18

La courbe parabolique du volet CLRG_CN formée par les nuages de mots groupés sur le plan de l'AFC (figure 7.8 ci-dessus) montre que l'emploi du vocabulaire chinois est récurrent sur l'axe chronologique, comme l'explique l'effet Guttman. Nous voulons savoir dans notre cas quels sont les éléments déclencheurs de l'effet Guttman ? C'est-à-dire, quels sont les thèmes récurrents du volet et par quelles formes spécifiques sont véhiculés ces thèmes ?

Pour ce faire, une étude sur les spécificités chronologiques pour le volet chinois a été réalisée, dont les résultats sont disponibles dans le tableau 7.7 ci-dessus. En effet, selon les résultats de l'AFC (figure 7.8) et en lisant la courbe de droite à gauche, les 8 années du volet se suivent les unes après les autres, à l'exception des années 2014, 2012 et 2013 (ordre du plan). Le plan AFC est scindé en 4 blocs, blocs séparés par les 2 axes. Ainsi, les 4 blocs nous livrent 4 groupements d'années par leurs proximités, à savoir,

- 2006, 2007, 2008, (en bas à droite)
- 2009, (en haut à droite)
- 2010, 2011 (en haut à gauche)
- 2012, 2013, 2014 (en bas à gauche)

Ensuite, nous procédons aux calculs de spécificités, bloc par bloc, en respectant l'ordre chronologique de la courbe et obtenons les spécificités chronologiques de ce volet. Le tableau 7.7 ci-dessus présente l'extrait des spécificités les plus saillantes de chacun des 4 groupements.

L'analyse du tableau 7.7, extrait des spécificités évolutives du volet CLRG_CN, nous apporte les résultats suivants : 2 thèmes saillants et récurrents, à savoir, *climat* et *énergies.* Il est à noter que l'intitulé de la rubrique du site *Chinadialogue* est *Climate change* & *Energy*. Nous allons expliquer comment évoluent les 2 thèmes entre 2006 et 2014. Avant la Conférence de Copenhague, le thème 气候变化/qì hòu biàn huà/changement climatique (forme colorée en bleu pâle dans le tableau) reste très récurrent dans le discours des ONG de 2006 à 2008. Dès l'ouverture de cette Conférence en 2009, les ONG relatent les informations à la fois sur le *changement climatique* et sur celles de la Conférence de 根本哈根/gē běn hā gēn/Copenhague (forme colorée en mauve). En 2011, l'accident nucléaire de Fukushima s'est produit, alors le thème 核电/hé diàn/électricité nucléaire (forme colorée en rouge) occupe une place importante dans la presse des ONG, sans oublier la Conférence de Cancún sur le *climat* qui a eu lieu en 2010. La coexistence des deux thèmes pendant cette période (2010-2011), *climat* et *énergies (électricité nucléaire)*, a permis le basculement du thème *climat* au thème *énergies* entre 2006 et 2011, en passant du *changement climatique* à l'*électricité nucléaire*. Par la suite, entre 2011 et 2014, les thèmes tels que *énergies* dont l'*électricité nucléaire* dominent le discours de ces ONG.

Le chemin de transition des 2 thèmes est le suivant :

climat (2006-2008) → *climat* + *Copenhague* (2009) → *climat* + *énergies* avec *électricité nucléaire* (2010-2011) → *énergies* avec *électricité nucléaire* (2012-2014).

Nous venons d'exposer le phénomène de thèmes par récurrence, cette courbe parabolique de thèmes récurrents s'appelle l'effet Guttman.

Afin d'avoir une vision plus fine de ces proximités textuelles et de pouvoir comparer avec les résultats d'ENRG, nous procédons également aux AFC sur les deux volets de CLRG par mois.

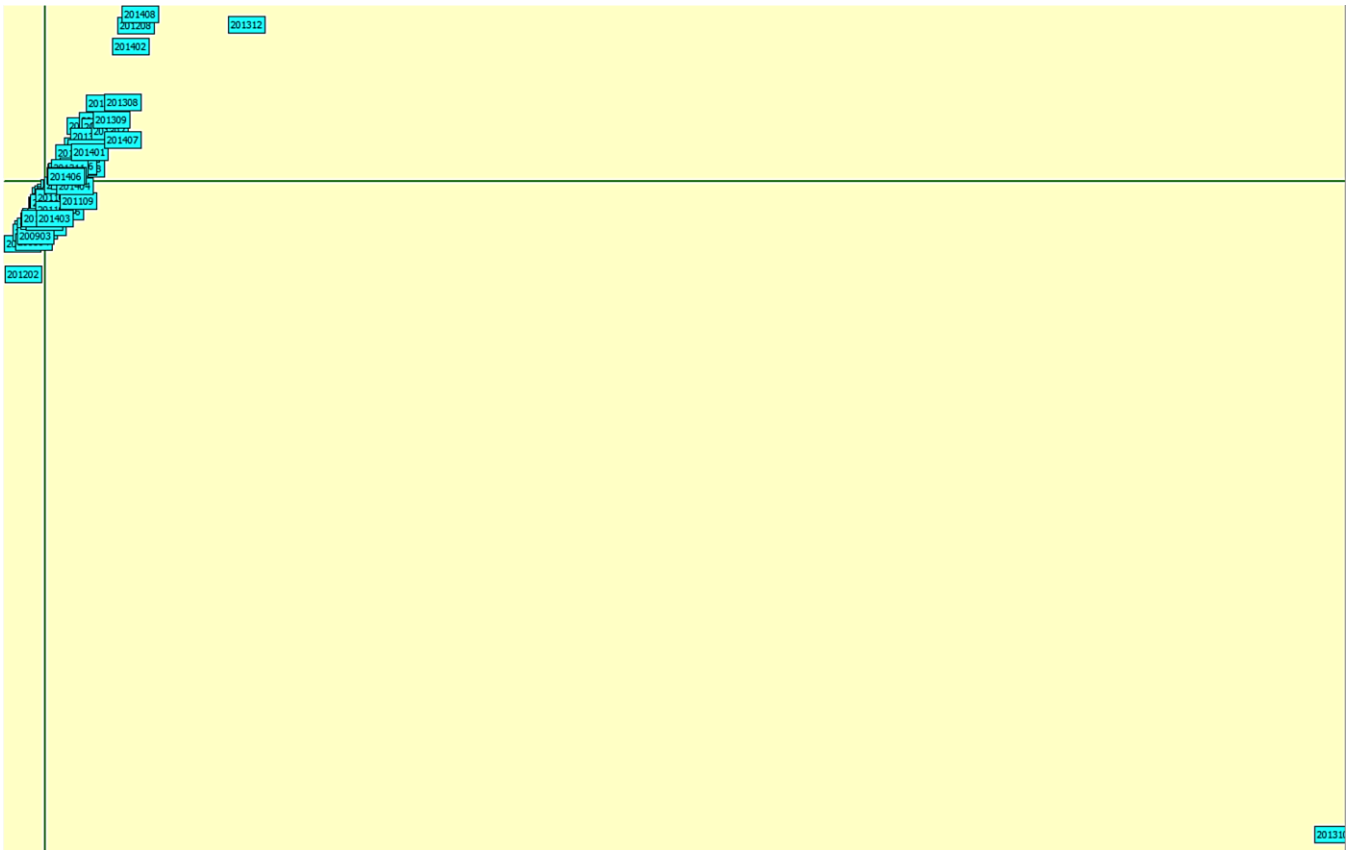


Figure 7.10 CLRG_CN de 2006 à 2014 : analyse factorielle des correspondances par mois

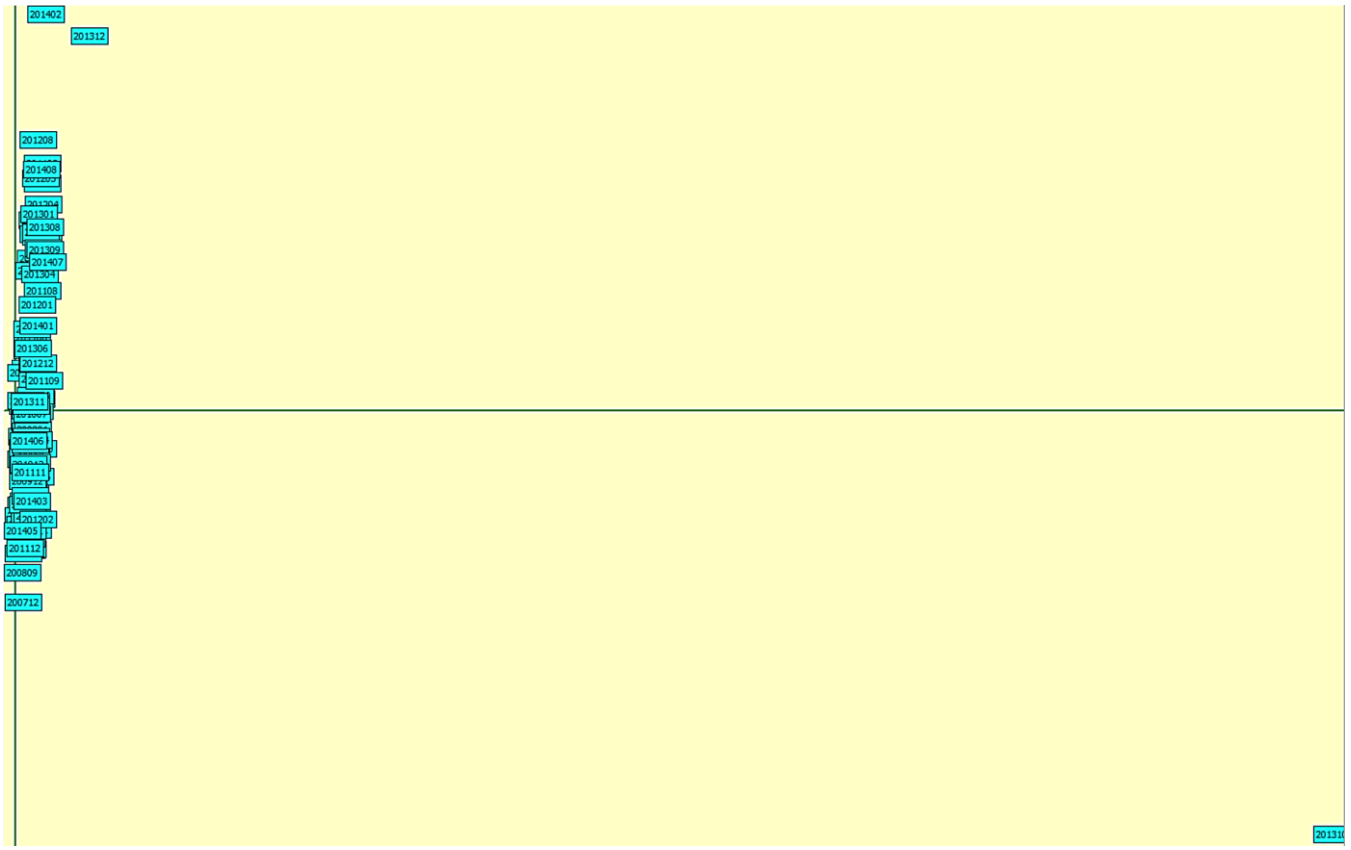


Figure 7.11 CLRG_EN de 2006 à 2014 : analyse factorielle des correspondances par mois

Les AFC par mois (figures 7.10 et 7.11), sur les deux volets attestent une convergence et une superposition d'informations thématisées du corpus à l'exception des mois suivants :

- octobre 2013, complètement écarté du reste pour les deux volets,
- août 2012, décembre 2013, février et août 2014 pour CLRG_CN,
- seulement décembre 2013 et février 2014 pour CLRG_EN.

Toutefois, l'exclusion des deux mois d'août en 2012 et 2014 dans CLRG_CN démontre que la traduction d'une langue à une autre dévie parfois la symétrie statistique des informations bilingues. Ce phénomène est dû à la traduction des notions spécifiques souvent manquantes dans la langue d'arrivée, telles que les toponymes, acronymes, entités nommées, néologismes, termes professionnels, terminologies, etc. Le fait de les traduire sémantiquement (mais pas par transcription phonétique du chinois) entraîne des changements morphosyntaxiques et lexicographiques dans la langue d'arrivée. Par exemple, prenons le mois d'août 2012 :

- le toponyme 三峡/sānxiá/Trois Gorges, forme chinoise évoquée, est souvent traduite par Trois-Gorges en français et *Three Gorges Dam* en anglais. Cette traduction transforme une forme spécifique en trois formes distinctes en anglais, de surcroit, la forme 三/sān/trois est une forme composée lexicalement.
- le nom de la centrale nucléaire *San Onofre* (deux formes distinctes) se traduit par 圣奥诺/shèng ào nuò/San Onofre en un seul mot chinois.
- le terme spécifique 家电/jiā diàn désigne l'électro-ménager en français, *Home Appliances* en anglais, de plus, home est une forme composée (*Home Office* du gouvernement britannique, *Ministry of Home Affairs* (MHA) en Inde).
- le terme *centrale nucléaire* en français, 核电站/hé diàn zhàn/centrale nucléaire ou 核电厂/hé diàn chǎng/usine nucléaire en chinois, se traduit en anglais par *nuclear power plant* ou *nuclear generating station*.

Ces exemples expliquent, entre autres, cette déviation de la symétrie statistique des informations bilingues de l'AFC.

Afin de mettre en évidence les informations principales de CLRG, les calculs de segments répétés ont été appliqués sur les deux volets du corpus sur l'ensemble de la période.

Tableau 7.8 CLRG de 2006 à 2014 : segments les plus répétés

longeur	segment répété	équivalent français	fréquence	longeur	segment répété	équivalent français	fréquence
2	of the	du	8159	2	中国 的	(de) Chine	1252
2	in the	dans le	5877	2	温室 气体	gaz à effet de serre	1016
2	to the	au	2790	2	中的	milieu de	897
2	and the	et le	2558	2	的 是	est/ce que /ce qui	888
2	climate change	changement climatique	2468	2	气候变化 的	(de) changement climatique	830
2	on the	sur	2143	2	的 问题	problème de	828
2	the world	le monde	1853	2	可 再生能源	énergies renouvelables	803
2	to be	être	1839	2	新 的	nouveau/nouvelle	758
2	for the	pour le	1720	2	的 影响	effets / influence	752
2	is a	est un	1586	2	我们 的	notre	733
2	at the	au	1560	2	上 的	sur	693
2	that the	que le	1530	2	可 持续	durable	684
2	from the	du / à partir de	1413	2	更 多	plus	659
2	will be	sera	1365	2	的 国家	pays	630
2	by the	par le	1301	2	都 是	tous sont	626
2	it is	c'est	1237	2	多 的	plus	626
2	with the	avec le	1143	2	% 的	X %	618
2	the United	Union, US, UK	1112	2	是 一个	est un	598
2	is the	est le	1066	2	最大 的	le plus grand	580
2	in China	en Chine	991	2	这样 的	cette	578
2	such as	comme	961	2	也 是	est également	576
2	as a	comme un	954	2	碳 排放	émissions de carbone	573
2	has been	a été	941	2	年 的	année	572
2	of a	d'un	927	2	他们 的	leur	568
2	United States	États Unis	899	2	巨大 的	grande	515
2	the US	les États-Unis	878	2	的 发展	développement de	512
2	in a	dans un	839	2	在 中国	en Chine	512
2	have been	a été	819	2	的 时候	temps	510
2	as the	comme le	809	2	气体 排放	les émissions de gaz	489
3	the United States	les États-Unis	783	2	的 能源	énergie	488
2	need to	avoir besoin de	754	3	温室 气体 排放	les émissions de gaz à effet de serre	479
2	there is	il y a	738	2	重要的	important	465
2	is not	n'est pas	729	2	国家 的	de l'Etat	461
2	more than	plus que	715	2	的 人	personnes	461
2	the country	le pays	705	2	就 会	il sera alors	459
2	to a	à un	703	2	的 碳	charbon de	455
2	It is	Il est	687	2	应对 气候变化	traiter le changement climatique	453
2	would be	serait	639	2	问题 的	(de) problème/question	451
2	of climate	de climat	611	2	世界上	dans le monde	449
2	can be	peut être	610	2	地区 的	de zone / région	444
2	one of	un des	610	2	全球变暖	le réchauffement climatique	443

Le tableau 7.8 ci-dessus illustre les 41 segments les plus répétés de CLRG, nous constatons que la fréquence de segments répétés du volet anglais est beaucoup plus élevée que celle du chinois ; par exemple, la fréquence du segment *climate change* est de 2 468 dans le volet anglais, tandis que dans le volet chinois, la fréquence est de 830. La signification des segments répétés du volet anglais relève peu d'informations intéressantes. En revanche, les segments répétés en chinois nous révèlent les véritables thèmes du corpus. Dès lors, nous pouvons dire que deux types de répétitions se manifestent : d'une part de mots grammaticaux pour l'anglais, et d'autre part, de mots de contenu pour le chinois. Rappelons que la forte répétition de mots grammaticaux est la cause du grand nombre d'occurrences en anglais. Plus l'emploi des mots grammaticaux est intensif, plus le nombre d'occurrences est important. Ce phénomène dissymétrique des segments répétés dans les deux volets est absolument normal, car la structure syntaxique des deux langues est complètement différente. Le fait d'avoir des traductions de l'un à l'autre ne prouve nullement l'emploi symétrique des segments qui se répètent de la même manière dans les deux langues. Cependant, un prétraitement de l'anglais pour éliminer les mots outils donnerait plus de sens à l'étude de ses segments répétés.

La forme *EPR* n'apparait pas en tête du tableau des segments répétés. En effet, elle a été employée seulement 15 fois dans CLRG_CN et 12 fois et dans CLRG_EN, cela réaffirme que cette donnée est un signal faible, car même dans les trois sous-corpus d'ENRG, la présence de l'entité nommée *EPR* est insignifiante. Rappelons que cette forme apparait 219 fois dans ENRG_FR, 0 fois dans ENRG_EN et 23 fois dans ENRG_CN (*cf*. figure 7.20 ci-après).

Il est également à noter que la forme plurielle *EPRs* est apparue 6 fois uniquement dans le volet anglais. Le retour au contexte nous permet d'affirmer que cette forme plurielle est due aux informations relayées sur l'EPR en Chine et en France, à savoir, deux réacteurs EPR sont en cours de construction en Chine, et la politique nucléaire française veut que les anciens réacteurs REP (se reporter à l'annexe J, section 4) soient remplacés progressivement par les EPR. Au vu de ce constat, nous nous intéressons uniquement à la forme *EPR* au singulier.

Nous procédons aux calculs des réseaux cooccurrentiels bilingues des formes 能源/néng yuán/énergie et *Energy* puis à la ventilation de cette forme par les calculs des spécificités sur le corpus entier.

7.3 Réseaux cooccurrentiels comparés

7.3.1 Autour des formes 能源/néng yuán/énergie et *Energy*

Dans le but de comparer les réseaux cooccurrentiels dans les deux corpus ENRG et CLRG, nous avons procédé au calcul de cooccurrences des formes 能源/néng yuán/énergie et *Energy*.

Tableau 7.9 CLRG de 2006 à 2014 : réseaux cooccurrentiels parallèles des formes 能源 /néngyuán/énergie et *energy*

Cooccurrents (source) : 能源 (fq:3015) Co-Freq : 2 | Seuil : 10 Cooccurrents (cible) : *energy* (fq:4475) | Co-Freq : 2 | Seuil : 10

Cooccurrents : (source) 能源

- *Co-Freq : 2*
- *Seuil : 10*

Forme	Équivalent français	Fq	co-Fq	specif	context
消耗	consommer	208	93	25.7	77
合作社	Coopérative	38	24	12.0	17
太阳能	énergie solaire	832	269	38.5	169
封存	captage et stockage	94	39	10.7	27
2030	2030	141	65	19.4	48
煤炭	Charbon / houille	630	284	75.3	157
比例	proportion	180	71	16.5	59
美	États-Unis	170	75	20.7	50
能耗	Consommation d'énergie	246	87	16.4	57
能源供应	approvisionnement en énergie	58	36	16.8	33
政策	politique	1519	390	30.1	300
能效	consommation d'énergie	313	143	39.8	103
加强	renforcer	251	75	10.4	66
分布式	de façon distribuée	54	41	24.1	22
效率	efficacité	265	140	48.3	111
为主	être majoritairement à	74	34	11.0	26
交通	trafic ou circulation	134	49	10.6	45
排放	émission	2904	606	20.6	357
天然气	gaz naturel	475	172	31.8	119
部落	tribu	92	39	11.0	16

Cooccurrents : (cible) energy

- *Co-Freq : 2*
- *Seuil : 10*

Forme	Équivalent français	Fq	co-Fq	specif	context
potential	potentiel	480	161	14.0	149
both	les deux	864	251	12.6	220
bulbs	ampoules	60	41	16.7	31
appliances	appareils	59	45	21.3	39
dependence	dépendance	55	33	11.5	30
15%	15%	78	39	10.0	37
green	vert	650	210	15.7	170
lighting	éclairage	51	38	17.7	33
target	cible	332	129	17.0	109
intensive	intensif	143	88	28.7	79
emissions	émissions	2793	712	16.5	490
power	puissance	2309	936	122.6	599
China	Chine	6988	2005	80.9	1089
coal	charbon	1443	591	80.2	344
bulb	ampoule	25	21	12.3	14
transition	transition	122	55	11.1	55
sector	secteur	495	184	20.9	159
reduction	réduction	520	172	14.2	149
supply	alimentation	445	192	30.9	170
Plan	Plan	224	106	21.6	77

Le tableau[221] 7.9 ci-dessus montre les 20 premières formes (formes affichées/triées par ordre d'apparition dans les phrases du corpus) cooccurrentes parallèles de la forme-pôle 能源/néng yuán/énergie et *Energy* dans CLRG. La colonne contexte indique le nombre de séquences textuelles (phrase, paragraphe ou article) dans lesquelles les formes cooccurrentes sont apparues. Par exemple, la forme cooccurrente 消耗/xiāo hào/consommer a été répétée 208 fois dans 77 séquences textuelles.

[221] Le tableau complet des réseaux cooccurrentiels parallèles des formes 能源/n éngyuán/énergie et *Energy* est disponible dans l'annexe N, tableau N.2.

Le réseau cooccurrentiel du volet chinois nous fournit des éléments ancrés dans une réalité de croissance démographique et économique : besoin de produire de l'électricité à tout prix. Toutefois, l'émission de CO2 fait partie également dans ce contexte des préoccupations énergétiques.
Tandis que les contextes en anglais reposent davantage sur « comment économiser l'énergie » et « comment la Chine va réussir son *Challenge* énergétique. »

7.3.2 Autour des formes 核能 /hé néng/ énergie nucléaire et *nuclear*

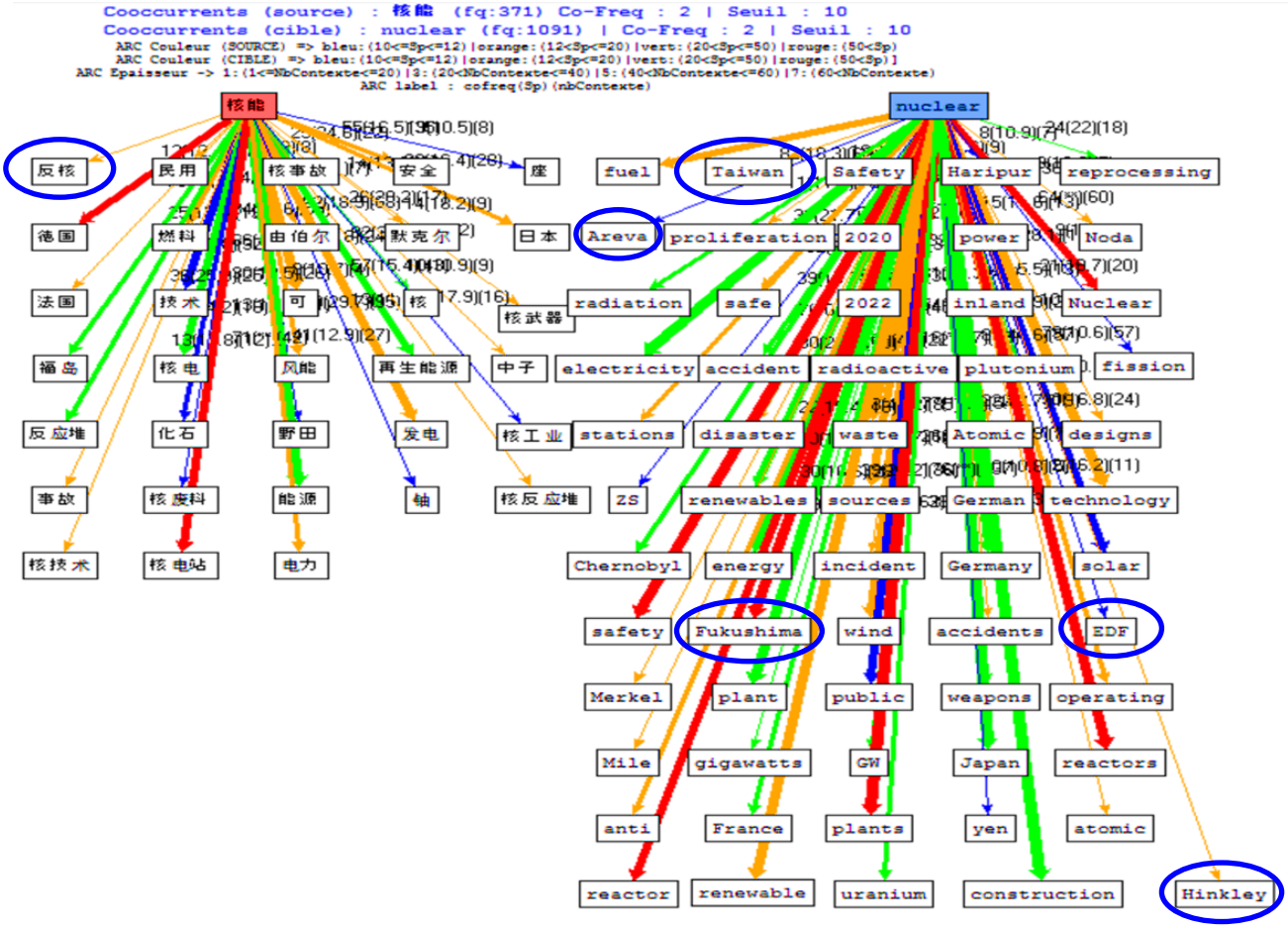


Figure 7.12 CLRG de 2006 à 2014 : réseaux cooccurrentiels parallèles des formes 核能 /hé néng/ énergie nucléaire et *nuclear*

Dans ces deux réseaux cooccurrentiels parallèles (figure 7.12 ci-dessus), réseaux obtenus autour de la forme 核能 /hé néng/ énergie nucléaire en chinois et la forme *nuclear* en anglais, on voit que la forme 反核/fǎn hé /anti-nucléaire est fortement représentée dans le réseau chinois. Tandis que dans le réseau anglais, ce sont les formes relatives aux toponymes comme *Taiwan*, *Hinkley*, aux éponymes des accidents comme *Fukushima*, aux entités nommées comme *EDF*, *AREVA*, qui sont fortement corrélées. Il faut souligner que l'apparition de la forme *Hinkley* (Centrale nucléaire de Hinkley Point en Angleterre) vient enrichir notre fouille d'informations autour de l'évolution de la construction du réacteur EPR dans le monde.

Il convient de noter que dans ce réseau parallèle le nombre de cooccurrences du volet anglais est plus important que celui du chinois. Ceci est dû également au choix de segmentation de Jieba.

7.4 Forme-pôle *EPR* dans les deux volets de 2006 à 2014

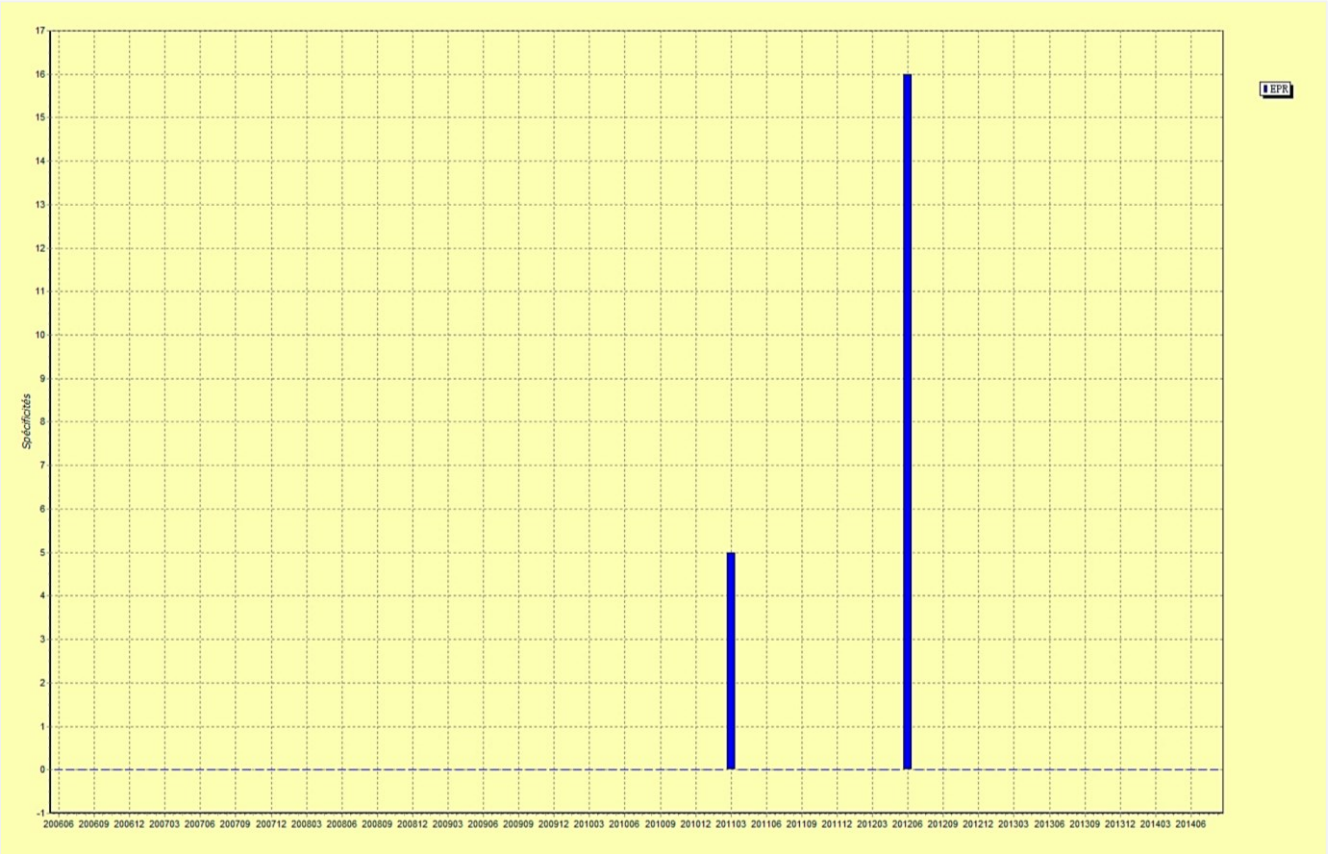


Figure 7.13 CLRG_CN de 2006 à 2014 : ventilation par mois des segments répétés *EPR*

La ventilation des cooccurrences de la forme *EPR* que l'on voit sur de la figure 7.13 confirme dans le volet chinois que la forme *EPR* est spécifique aux mois de mars 2011 et juin 2012, plus particulièrement pour le mois de juin. L'événement de Fukushima s'est produit en mars 2011 et cette ventilation reste très marquée pour ces deux mois ; elle a été réalisée par des calculs portant uniquement sur les textes chinois du volet. Par ailleurs, au mois de mai 2012, la France a élu un nouveau Président de la République dont la ligne politique en matière de nucléaire est différente de celle qui était suivie par le précédent (se reporter à l'annexe H, section Les autres formes poly-cooccurrentes). Nous allons porter une attention particulière au mois de juin 2012 et faire l'hypothèse que la prééminence de cette forme est due, d'une part, au retard annoncé du chantier EPR et à l'augmentation du coût, d'autre part, au contexte de l'arrivée de François Hollande. Ces deux nouveaux facteurs pèsent sur les perspectives de l'avenir du nucléaire en France. Le retour au contexte va nous permettre de vérifier la véracité de cette hypothèse. Nous commençons par une projection de la forme *EPR* dans les deux volets.

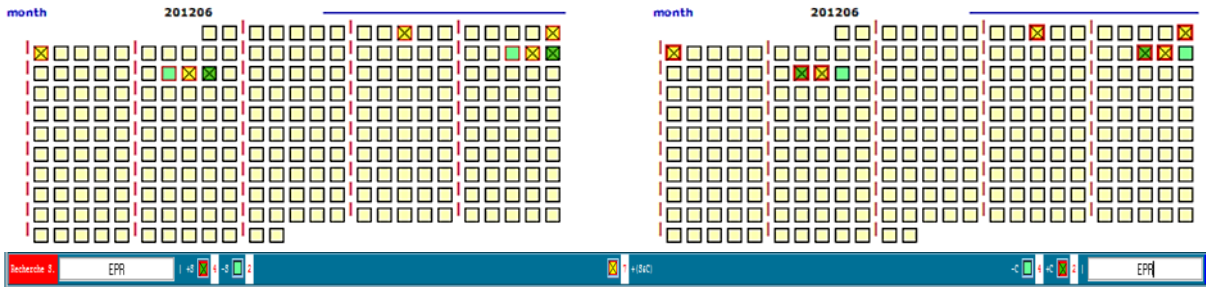


Figure 7.14 CLRG : carte des sections parallèles pour la forme *EPR* du mois juin 2012, à gauche CLRG_CN, à droite CLRG_EN

Légende du volet gauche	Légende du volet droit
carré vert : fréquence 2	carré vert : fréquence 4
carré vert croisé : fréquence 4	carré vert croisé : fréquence 2
carré jaune croisé : fréquence 7	carré jaune croisé : fréquence 7

La carte des sections bilingues de Mkalign (figure 7.14 ci-dessus), où s'illustre la répartition de la forme *EPR* dans les deux volets du corpus, indique que le mois de juin 2012 est le mois où cette forme se répète le plus dans tout le corpus. Le premier article, se trouvant dans la dixième case de la première ligne de la carte des sections, livre des informations très révélatrices que nous avons reproduites ci-dessous. Il est à noter que cet article dont l'extrait ci-dessous est daté du 6 juin 2012 sur *Chinadialogue.net* a été rédigé en anglais comme langue source par le Professeur Steve Thomas[222] puis traduit en chinois et publié sur le même site par Monsieur DONG Jun[223].

> *<month=201206> <day=2012-06-06> <article=707> Title: Nuclear Europe: a dream unwinding (...)# What Fukushima has done for Germany and Italy is to close off the nuclear option forever. All sides now know they must commit fully to energy efficiency and renewables to meet the climate-change challenge. And the environmentalists' claim that nuclear isn't necessary will be properly tested. # Meanwhile in France, the heart of Europe's nuclear industry, it remains to be seen how firm Francois Hollande's position will remain in office. Those with long memories will remember Francois Mitterand coming to power in 1981 on an apparent promise to stop new nuclear, only for a further 10 or so orders to be placed over the following six years. # But the real challenge – regardless of whether Hollande or Sarkozy had won the election – was always going to be what to do about France's existing plants when they reach the end of their lives. Under present plans, these ageing reactors will be retired at a rate of five to six per year from 2017 onwards. The cheaper option for the country's power giant EDF would be to do as the Americans and extend the plants' lifespans from 40 to 60 years, though thanks to post-Fukushima regulatory requirements that existing plants be made more robust for "extreme situations" this is not such a cheap option as it once was. # Such a move would also likely sound the death knell for Areva's problematic European Pressurised Reactor (EPR), the design causing huge delays and cost overruns at Olkiluoto in Finland and Flamanville in France. Both projects are running four years or more late and about 100% over budget. Without new French orders from Areva – a French company – the design would lose all credibility. # On the other hand, if France takes the route of replacing old reactors with EPRs, assuming problems around cost, licensing and construction can be solved, and the EPR remains a viable option, then the cost to EDF of replacing old capacity would be astronomical – far higher than first time around. It is doubtful that France could sustain the logistical and financial challenge of ordering and building four or five EPRs a year for a decade. It would also have to start*

[222] Professeur à l'Université de Greenwich Business School, travaillant dans le domaine de la politique énergétique. http://www2.gre.ac.uk/about/faculty/business/study/ibe/staff/steve-thomas (consulté le 19/09/2015).
[223] Afin de préserver l'authenticité de l'information relayée dans cet article, nous décidons de le traduire en français à partir de sa langue source, et non à partir du chinois, bien que les calculs de ventilation aient été obtenus par les textes chinois.

paying huge sums for decommissioning existing reactors. That leaves France facing some tough choices. (...)

Traduction française d'un extrait de l'article contenant la forme *EPR* depuis l'anglais

Titre : Nucléaire en Europe : le dénouement d'un rêve.

(…)# Ce que Fukushima a apporté comme leçon à l'Allemagne et à l'Italie est d'enterrer l'option nucléaire à jamais. Toutes les parties prenantes savent maintenant qu'elles doivent s'engager complètement vers des énergies efficaces et renouvelables, afin de relever le défi du changement climatique. Par ailleurs, l'affirmation des environnementalistes concernant l'arrêt du nucléaire sera correctement testée. Pendant ce temps, en France, pays cœur de l'industrie nucléaire européenne, on attend de voir si François Hollande restera ferme sur sa position une fois qu'il sera au pouvoir. Ceux qui ont une bonne mémoire se rappellent l'arrivée au pouvoir de François Mitterrand en 1981 avec la promesse de ne plus ouvrir de nouveaux chantiers de centrales nucléaires, suivie d'une dizaine de commandes passées dans les six années suivantes.
Mais le véritable défi, indépendamment de la victoire aux élections présidentielles de Hollande ou de Sarkozy, était de savoir ce que l'on ferait avec les centrales françaises existantes quand elles atteindraient leur fin de vie. Selon le programme nucléaire actuel, ces réacteurs vieillissants devront être retirés à raison de cinq ou six par an à partir de 2017. L'option la moins chère, pour EDF, le géant de l'électricité française, serait de faire comme les Américains, c'est-à-dire de prolonger la durée de vie des centrales de 40 à 60 ans. Grâce au fait que Fukushima a imposé de nouvelles exigences réglementaires pour les centrales existantes, exigences entrainant des travaux de mise aux normes pour résister à des situations extrêmes, cette option n'est pas non plus très bon marché. Une telle initiative sonnerait probablement le glas du réacteur européen à eau pressurisée (EPR) d'Areva, en raison de sa conception qui a causé des retards et des surcoûts financiers à Olkiluoto en Finlande et à Flamanville en France. Les deux projets en cours ont pris quatre ans de retard voire plus et leur coût dépasse le budget initialement prévu de 100 %. Sans nouvelles commandes françaises à la société Areva, la conception de l'EPR perdrait toute crédibilité.
D'autre part, si la France choisit de remplacer ses vieux réacteurs par les EPR en résolvant les problèmes liés aux coûts, licence et construction, alors l'EPR demeure une option viable. A ce moment-là, les coûts supportés par EDF pour remplacer ces vieux réacteurs seraient astronomiques, prix bien supérieurs à ceux annoncés depuis la première augmentation du coût. Il semble peu probable que la France puisse supporter le défi logistique et financier consacré à une telle commande et à une construction de quatre ou cinq EPR par an pendant une décennie. Elle devrait aussi commencer à payer des sommes énormes pour le démantèlement des réacteurs existants. Ce qui laisse la France face à des choix difficiles. (…)

L'extrait de cet article est porteur d'informations sur la situation française face à la politique nucléaire et à son rêve brisé de l'EPR suite aux déboires qui se produisent depuis 2006. Il faut noter que ce passage met en évidence l'aspect critique et dénonciateur de la part des ONG envers la situation nucléaire en France. La spécificité de la forme *EPR* est-elle parallèle dans les deux volets ? La même ventilation dans le volet anglais nous livrera la réponse.

La ventilation de la figure 7.15 ci-dessous du volet anglais montre que la forme *EPR* est uniquement spécifique aux mois de janvier 2008, mars 2011 et juin 2012. A la différence du volet chinois, le mois de janvier 2008 est apparu dans la ventilation des spécificités. Afin de mieux cerner cette apparition, nous projetons la forme *EPR* sur la carte des sections bilingues avec retour aux textes.

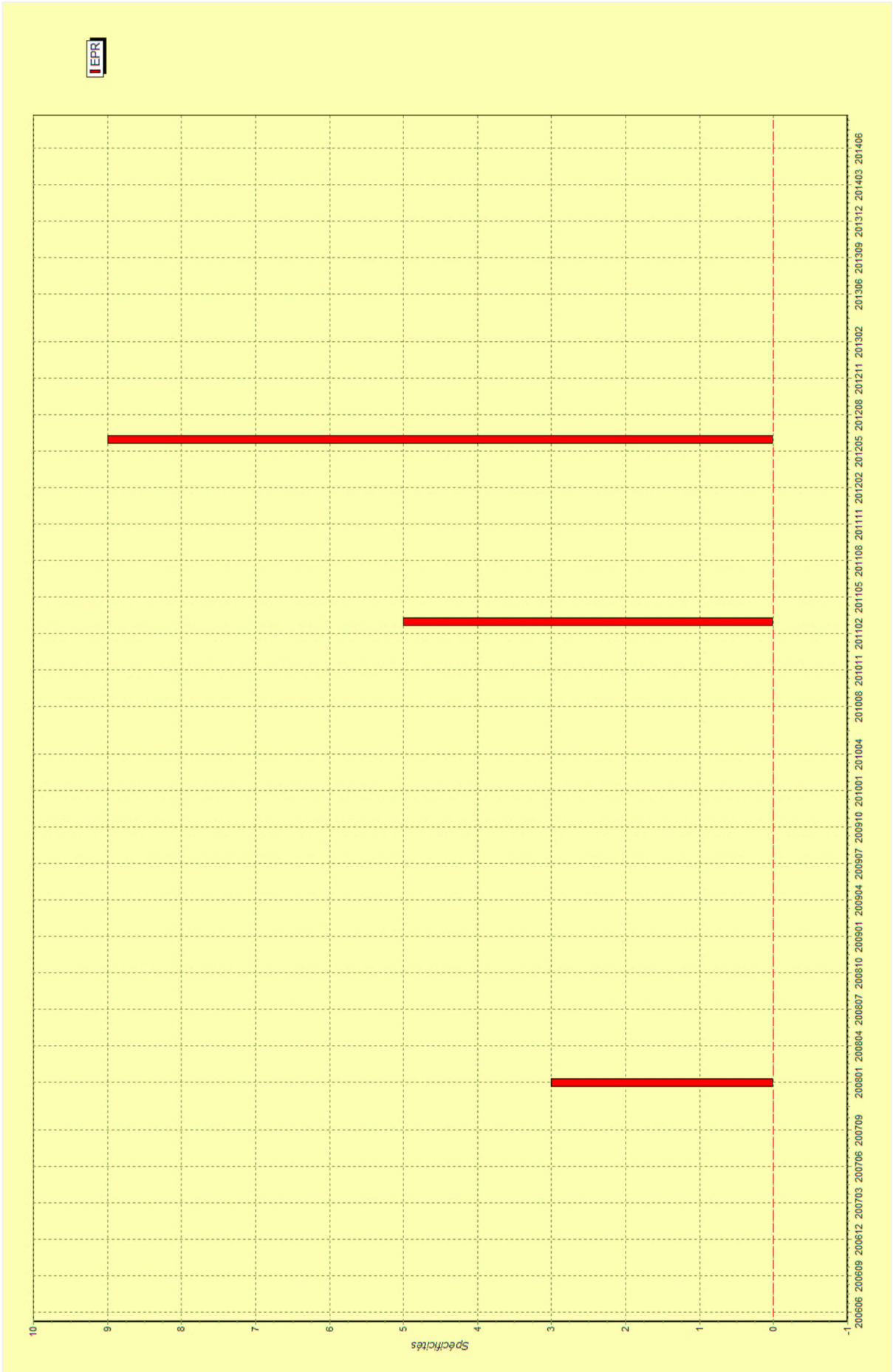


Figure 7.15 CLRG_EN de 2006 à 2014 : ventilation par mois de la forme *EPR*

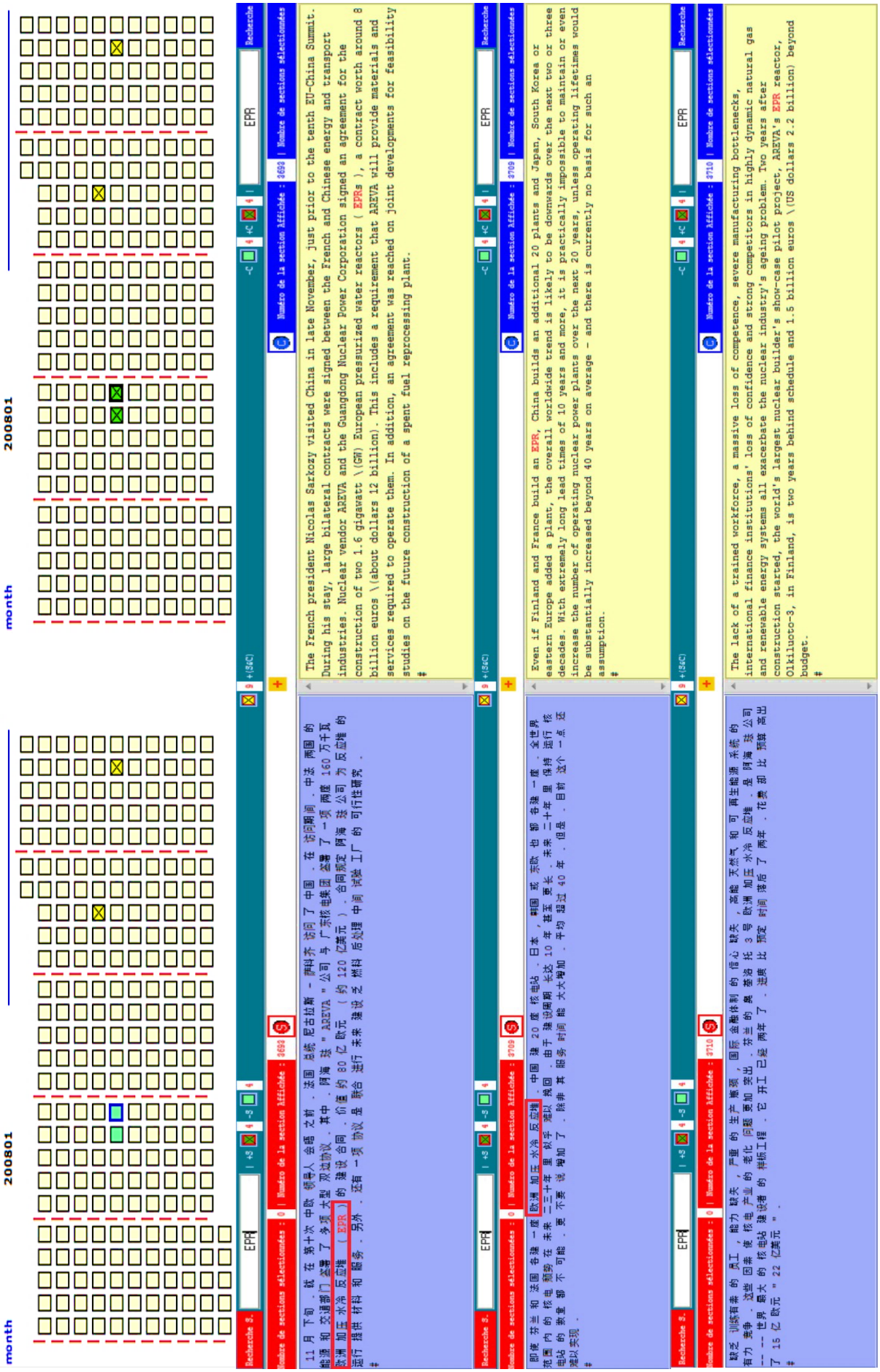


Figure 7.16 CLRG : carte des sections parallèles pour la forme *EPR* de janvier 2008, à gauche CLRG_CN, à droite CLRG_EN

Le retour aux articles est nécessaire pour tenter d'expliquer la raison pour laquelle cette forme *EPR* est spécifique uniquement dans le volet anglais et non dans le volet chinois pendant le mois de janvier 2008[224].

La consultation de l'article montre que la forme *EPR* (entourée en rouge) a été traduite en chinois par 欧洲加压水冷反应堆/Ōuzhōu jiā yā shuǐlěng fǎnyìngduī/EPR, dont les 4 segments sont 欧洲/Ōuzhōu/Europe, 加压/jiā yā/pressurisé, 水冷/shuǐlěng/refroidissement à eau, 反应堆/fǎnyìngduī/réacteur. La traduction littérale est européen pressurisé refroidissement à eau réacteur. En effet, comme l'illustre la partie du milieu gauche de la figure 7.16, de cet article, le sigle *EPR* a été reproduit et explicité en chinois, par la suite, c'est cette dernière forme chinoise qui a été utilisée pour désigner l'EPR dans le volet chinois.

La traduction du dernier passage illustré dans la figure 7.16 est disponible ci-dessous.

> *<day=20080107> <article=169> title: The global nuclear decline*
>
> ***Summary**: The "nuclear renaissance" is a myth. Forecasters continue to overestimate the growth of atomic energy, write Mycle Schneider and Antony Froggatt, not least in China, where the energy mix remains dominated by coal.*
>
> *(...)# The lack of a trained workforce, a massive loss of competence, severe manufacturing bottlenecks, international finance institutions' loss of confidence and strong competitors in highly dynamic natural gas and renewable energy systems all exacerbate the nuclear industry's ageing problem. Two years after construction started, the world's largest nuclear builder's show-case pilot project, AREVA's EPR reactor, Olkiluoto-3, in Finland, is two years behind schedule and 1.5 billion euros (US$2.2 billion) beyond budget. (...)*
>
> *Mycle Schneider (mycle@orange.fr) and Antony Froggatt (a.froggatt@btinternet.com) are both independent consultants on energy and environmental policy. They co-authored the World Nuclear Industry Status Report 2007.*

Traduction française d'un extrait de l'article contenant la forme *EPR* depuis l'anglais

> Titre : Le déclin mondial du nucléaire
>
> Résumé: La "renaissance du nucléaire" est un mythe. Les prévisionnistes continuent à surestimer la croissance de l'énergie atomique, écrivent Mycle Schneider et Antony Froggatt, phénomène pas moins vrai en Chine, où le mix énergétique reste dominé par le charbon.
>
> (…) # L'absence d'une main-d'œuvre qualifiée, une perte massive de compétences, de graves goulots d'étranglement de la fabrication, une perte de confiance des institutions internationales de financement et de sérieux concurrents dans les secteurs très dynamiques de gaz naturel et des énergies renouvelables, le tout exacerbe le problème du vieillissement de l'industrie nucléaire. Deux ans après le début de la construction du réacteur EPR d'AREVA, Olkiluoto-3 en Finlande, le projet vitrine et pilote du plus grand constructeur nucléaire au monde, accuse deux ans de retard et un dépassement de budget de 1,5 milliards d'euros, soit 2,2 milliards de dollars. (…) Mycle Schneider et Antony Froggatt sont deux consultants indépendants sur l'énergie et la politique environnementale. Ils sont co-auteurs d'un écrit intitulé « Rapport de l'état de l'industrie nucléaire mondiale 2007 ».

L'extrait d'article dû à des consultants indépendants pointe du doigt les difficultés de l'EPR, qu'elles soient techniques, financières ou sociales. Le nœud du problème repose d'une part sur la politique énergétique complexe, d'autre part, sur une recherche de solutions techniques et financières qui

[224] Il est à noter qu'un seul article dans lequel l'apparition de la forme *EPR* est asymétrique a été repéré, cette forme y est employée 1 seule fois dans le volet chinois et 3 fois dans le volet anglais.

semble sans fin. Il convient de noter que cet extrait souligne l'aspect critique voire ironique de la part de consultants indépendants.

Apports linguistiques pour la veille

Cette vérification ci-dessus par le retour aux textes bilingues nous conduit à des observations très intéressantes. Les entités nommées ou notions propres telles que l'EPR sont des notions véhiculées par des formes très spécifiques, porteuses d'informations et indicateurs d'événements pour la veille et la recherche d'informations, formes qui sont rarement traduites en chinois. Le fait de les traduire dans une langue lointaine telle que le chinois, déforme la structure morphosyntaxique, lexicale voire parfois sémantique des objets à rechercher. En effet, la traduction chinoise d'EPR, 欧洲加压水冷反应堆 (欧洲/Ōuzhōu pour Europe/européenne, 加压/jiā yā pour pressuriser/pressurisé, 水冷/shuǐlěng pour refroidissement à eau, 反应堆/fǎnyìngduī pour réacteur), peut être segmentée en 4 formes/mots individuellement différentes. Le fait que le segmenteur n'ait pas pu intégrer cette notion nouvelle et récente dans son dictionnaire, conduit à la séparation lexicale d'une entité nommée. Cette situation présente des similitudes avec les exemples français des notions spécifiques et banales comme pomme de terre, porte-clés, etc. Cet exemple souligne les limitations d'une étude de corpus fondée sur les formes brutes, sans analyse initiale autre qu'une segmentation en mots.

7.5 Cooccurrences et poly-cooccurrences parallèles : veille active et veille ciblée *EPR*

Dans le but de comparer les résultats de CLRG avec ceux d'ENRG, nous procédons aux calculs et analyses des cooccurrences et poly-cooccurrences de l'EPR dans les deux volets de CLRG, d'abord sur toute la période 2006-2014, puis sur les 3 années retenues de 2010 à 2012.

Nous commençons par une présentation des réseaux cooccurrentiels parallèles (figures 7.17 et 7.18 ci-dessous) issus du corpus parallèle entier toujours avec les paramètres par défaut.

Il est à noter que dans le cas du corpus parallèle, les réseaux poly-cooccurrentiels parallèles sont obtenus par des calculs statistiques appliqués individuellement à chaque volet. Les textes de chaque volet relatent les mêmes contenus d'information et chaque volet correspond à une langue, ce qui a pour conséquence d'obtenir un réseau poly-cooccurrentiel propre à chaque volet.

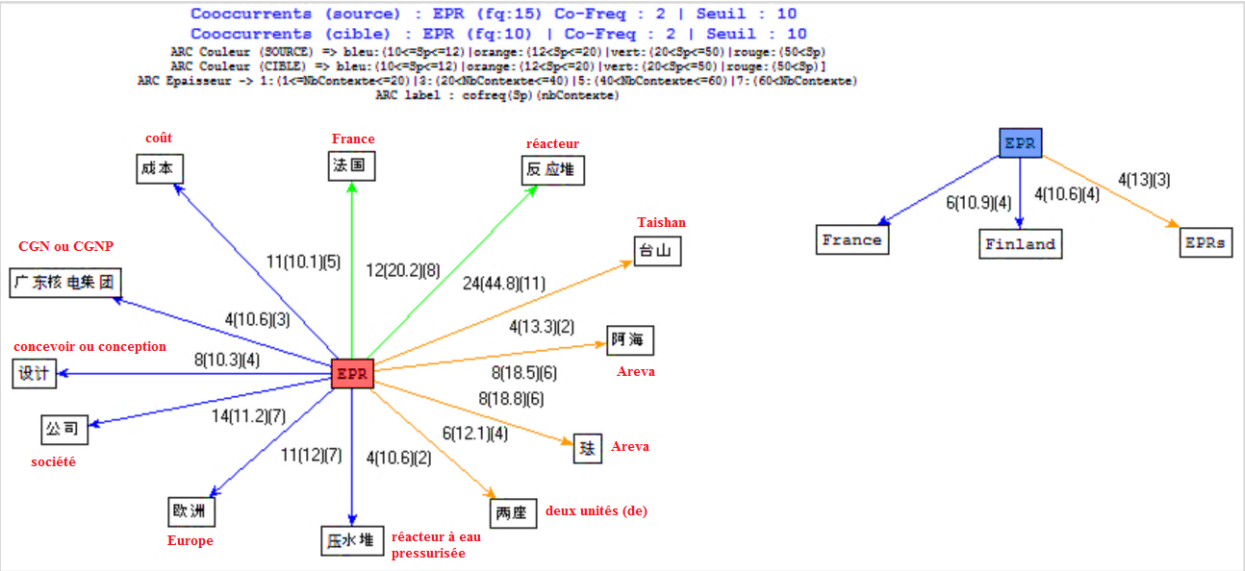


Figure 7.17 CLRG de 2006 à 2014 : réseaux cooccurrentiels parallèles autour de la forme *EPR*

Dans le volet chinois (figure 7.17 ci-dessus) pour la forme-pôle *EPR* nous avons 12 cooccurrents associés, tandis que seulement 3 ont été sélectionnés dans le volet anglais[225].

Hormis le cooccurrent *France v.s* 法国/fà guó/France qui a été repéré pour les deux volets, tous les autres cooccurrents sélectionnés sont différents. La notion *France* dans le volet chinois (ligne verte) est plus spécifique qu'en anglais (ligne bleue).

La tentative de remise en contexte de l'ensemble de ces cooccurrents des deux volets, considérés comme des mots-clés par volet nous livre les informations suivantes :

Volet chinois : le *réacteur à eau pressurisée* de la *société Areva* de *France*, *réacteur* de *conception Europe* dont *deux unités* sont en cours d'installation à *Taishan* par *CGN* avec un *coût*.

Les études précédentes issues d'ENRG nous ont déjà permis d'expliciter des faits se rapportant à l'EPR dans le monde. Ces informations fragmentaires issues des cooccurrents de CLRG_CN tracent une synthèse authentique et efficace de la situation de l'EPR. En revanche, il peut exister d'autres interprétations notamment le mot 两座/liǎng zuò/*deux unités*.

Volet anglais : La *France* et La *Finlande* construisent des *EPRs*.
La France et la Finlande construisent chacune un seul EPR, alors que la Chine en construit deux.

Ce calcul des cooccurrences parallèles corrobore encore une fois nos observations. Le système linguistique du discours chinois est plus répétitif, tandis que ceux de l'anglais et du français ont une tendance forte à employer les anaphores, armes ultimes pour éviter les mots qui se répètent. Cette hypothèse va être de nouveau illustrée par nos analyses.

Forme-pôle *EPR* dans la période 2010-2012

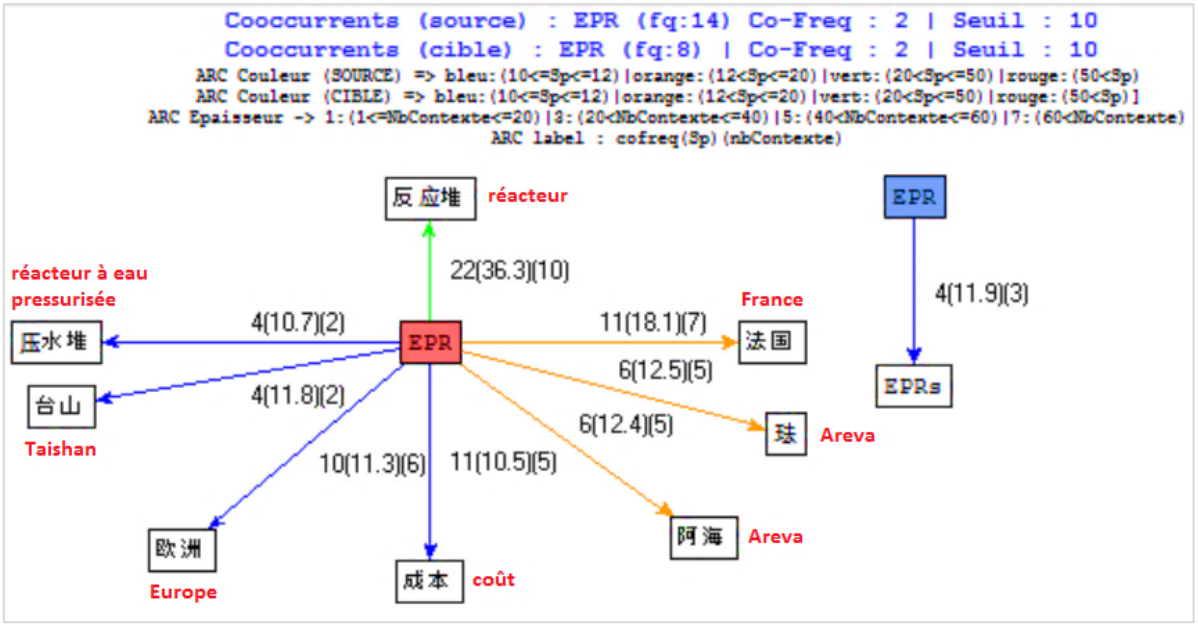


Figure 7.18 CLRG de 2010 à 2012 : réseaux cooccurrentiels parallèles autour de la forme *EPR*

[225] Il est à noter que la traduction du nom Areva en chinois 阿海珐/Ā hǎi fà a été segmentée en deux formes distinctes en raison de l'absence de ce mot dans le dictionnaire du segmenteur Jieba.

Sur les trois années (figure 7.18 ci-dessus), dans le volet chinois, 8 cooccurrents associés à la forme-pôle *EPR* sont apparus, pour seulement 1 cooccurrent dans le volet anglais.

Volet chinois

Par rapport à la période complète, 4 cooccurrents sont absents : *CGN*, *deux unités*, *société*, *conception*. Toutefois, l'interprétation possible de ce réseau cooccurrentiel ci-dessus ne modifie pas notre restitution d'informations sur la construction de l'EPR en Chine.

Volet anglais

L'information liée au réseau cooccurrentiel s'appauvrit par rapport à la période complète. Dans ce cas, la restitution est délicate sans retour au contexte.

Les réseaux cooccurrentiels nous apportent des formes bien définies, porteuses d'informations et d'indicateurs d'événements sur la forme-pôle recherchée. Nous allons voir ce que peuvent apporter les réseaux poly-cooccurrentiels.

Répétition de mots en langue chinoise, un fait linguistique

A travers les deux réseaux poly-cooccurrentiels, l'un sur 8 ans (2006-2014) et l'autre sur 3 ans (2010-2012), nous constatons que les deux formes *EPR* et *coût* décèlent manifestement une information intéressante. Dans ce corpus, les articles relaient les mêmes informations, mais ces articles sont rédigés d'abord dans l'une des deux langues puis traduits dans l'autre langue, deux systèmes langagiers différents. Or, les textes de ces deux volets ne produisent pas les mêmes réseaux cooccurrentiels au travers des récits des ONG. Le volet chinois génère constamment plus de formes cooccurrentes que le volet anglais. Ceci pourrait s'expliquer par le fait que le récit en chinois est moins gêné par la répétition de formes. Nous tenterons de vérifier cette hypothèse.

Une projection simultanée par année des deux formes, ***EPR*** et 成本/chéng běn/coût(s)/*cost(s)* sur la carte des sections (figure 7.19 ci-dessous, partie supérieure), permet d'accéder directement aux unités textuelles (paragraphes) qui les contiennent. Nous rappelons que dans la carte des sections lorsque les couleurs affichées sont atténuées, cela signifie que la répétition des formes sélectionnées est moindre par rapport aux couleurs plus foncées. La flèche rouge indique les paragraphes où les formes ***EPR*** et 成本/chéng běn/coût(s)/*cost(s)* sont co-présentes, paragraphes en chinois et en anglais cités en entier ci-dessous. C'est grâce à la consultation attentive de chacun des paragraphes en couleur que nous avons pu repérer ces paragraphes clés. L'extrait de l'article est révélateur puisqu'il montre, de manière flagrante, lors de la traduction de ce passage, que la répétition de mots en chinois n'est pas un véritable problème pour les locuteurs natifs.

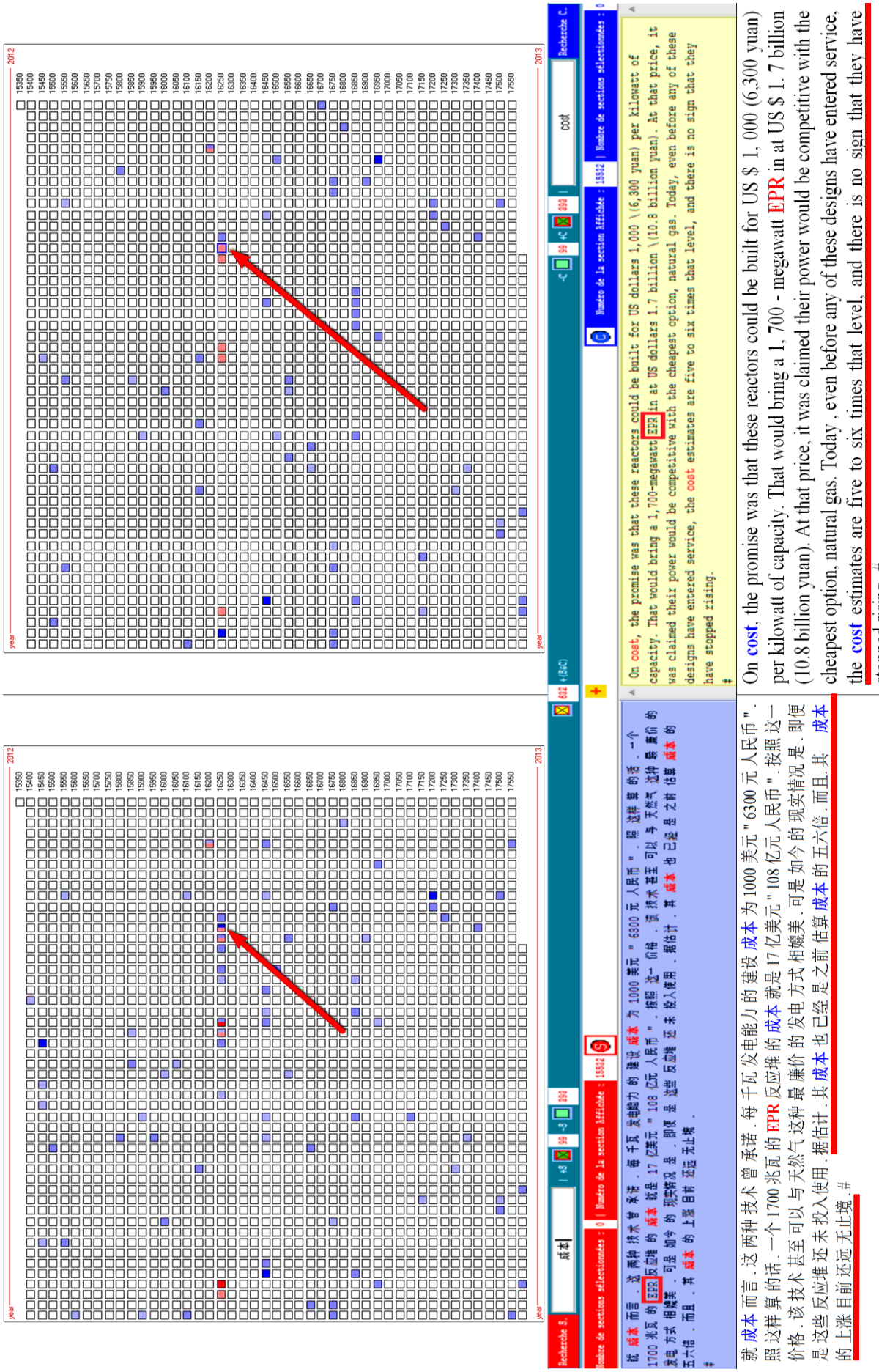


Figure 7.19 CLRG 2012 : carte des sections parallèles pour les deux formes *EPR* en rouge et 成本/chéng běn/coût/cost en bleu, un carré = un paragraphe, à gauche CLRG_CN, à droite CLRG_EN

Traduction française

> Au sujet des coûts financiers, la promesse avait été faite que ces réacteurs pourraient être construits pour 1 000 $ US (soit 6 300 yuans) par kilowatt. Cela reviendrait pour un EPR d'une puissance de 1 700 mégawatts à 1,7 milliards $ US (soit 10,8 milliards yuans). A ce prix, l'EPR serait compétitif par rapport au gaz naturel, l'option la moins chère (de toutes les énergies). Aujourd'hui, avant même que cette technologie soit mise en service, les estimations de coût sont multipliées par 5 voire par 6 et aucun signe ne montre que cette tendance s'inverse.

La construction des EPR, notamment en Finlande et en France, cumule une succession de retards par rapport aux dates annoncées de mise en service. Ces ajournements, en raison des défauts de construction et des renforcements des normes de sécurité et de sûreté, engendrent des contraintes de construction supplémentaires. Les surcoûts financiers considérables suite aux retards ont pour effet d'augmenter le prix de l'électricité fourni par ce type de centrale. Les EPR en Chine ont bénéficié des retours d'expériences européennes et sont mis en service avant ceux du Vieux continent.

Sur la figure 7.19 ci-dessus, partie inférieure, nous pouvons constater que pour un récit d'une même information, la forme 成本/chéng běn/coût/*cost* a été répétée 6 fois lors de la traduction de ce paragraphe, tandis qu'en anglais, la forme *cost* a été employée 2 fois. Cet exemple illustre bien ce phénomène de répétition en chinois. Pour les gens qui ne sont pas très familiers avec la langue chinoise, nous vous présentons un exemple d'anaphores. En français, on dit que «… les estimations de coût sont multipliées par 5 voire par 6 et aucun signe ne montre que cette tendance s'inverse» (phrase soulignée en rouge dans la figure 7.18). En chinois, cette phrase devient : « …son coût est déjà 5 à 6 fois supérieur à son coût initial et l'augmentation du coût semble sans fin ».

En effet, ce phénomène de répétition de mots est assez récurrent chez les locuteurs natifs, en particulier à l'écrit, car le fait de répéter les mots importants permet d'éviter les confusions de la compréhension et d'expliciter les sens du texte et/ou des notions. C'est pour cette raison que le réseau cooccurrentiel chinois est plus riche que celui de l'anglais. Nous rappelons que le phénomène du rapport occurrences/formes entre le chinois et l'anglais traité dans la section 7.2 nous montre qu'il s'agit de mots grammaticaux ou de mots-outils qui sont très répétitifs en anglais, alors qu'en chinois, c'est la répétition de notions qui est plus saillante.

Nous venons d'apporter un éclairage à la répétition de mots en langue chinoise qui est *de facto* un fait linguistique.

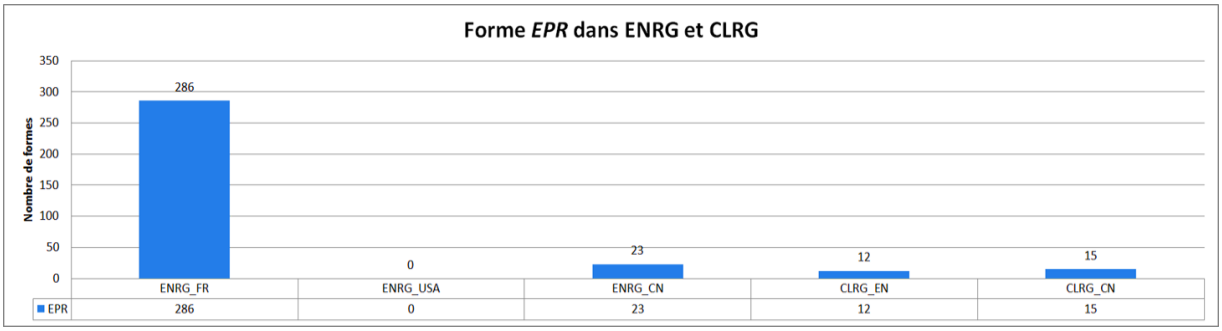


Figure 7.20 ENRG et CLRG de 1999 à 2014 : forme *EPR* dans les deux corpus complets

En dépit de la quantité faible de l'emploi de la forme *EPR*, considérée comme un type de signaux faibles (*cf.* chapitre 5, section 5.4.10, Veille active et veille ciblée par poly-cooccurrences évolutives de l'EPR), la répétition de certaines notions en chinois permet toutefois de détecter des informations clés. Dans l'étude présente, on se donne a priori la forme EPR, bien qu'elle ne ressorte pas naturellement du dépouillement du corpus. On ne la détecte pas comme un signal faible. On connaît à l'avance le signal à détecter. Une fois donnée cette forme, les associations en découlent, sont-elles encore des signaux faibles ? Au terme de nos études comparables et parallèles, la notion EPR représente un signal faible. Comme le montre la figure 7.20 ci-dessus, elle a été très peu employée dans CLRG, seulement un peu plus dans le sous-corpus français. Il est à noter que notre approche n'est pas de détecter ce signal faible, mais d'explorer les informations corrélées à ce signal qui sont ensevelies dans la masse des textes.

7.6 Réseaux poly-cooccurrentiels parallèles sur CLRG

Les paramètres par défaut ont été maintenus dans tous les calculs de poly-cooccurrences.

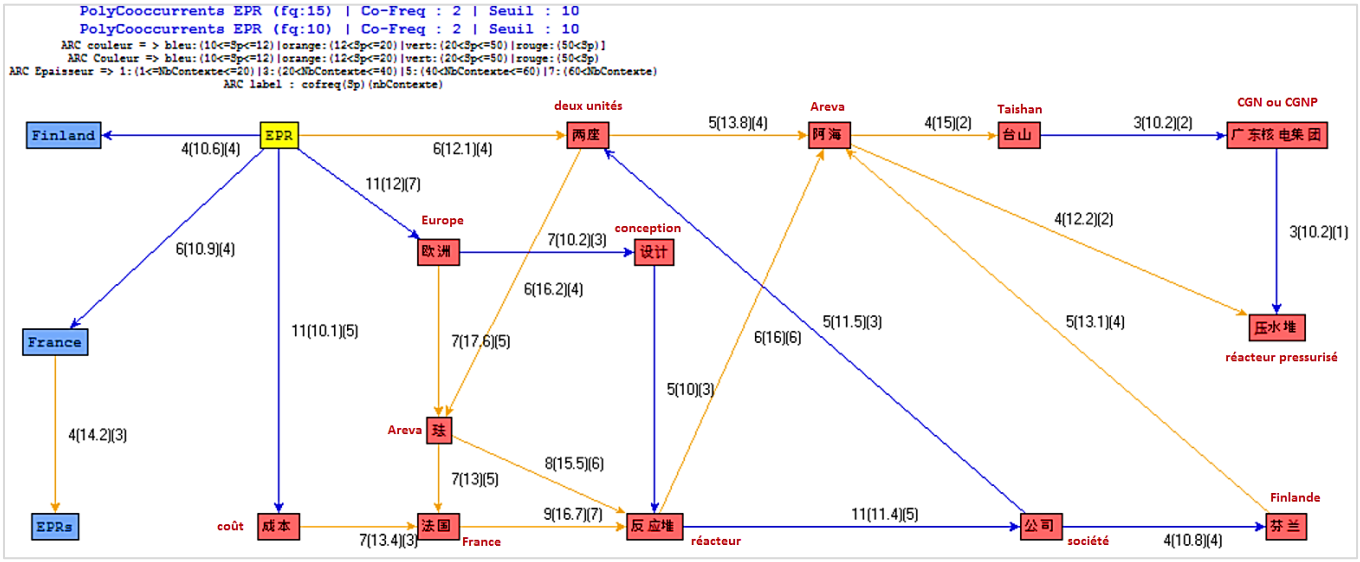


Figure 7.21 CLRG de 2006 à 2014 : réseau poly-cooccurrentiel parallèle de la forme *EPR*

Selon la figure 7.21 ci-dessus, la forme-pôle *EPR* nous apporte 13 poly-cooccurrents associés en chinois (en rouge), mais seulement 3 poly-cooccurrents (en bleu) en anglais.

Une nouvelle forme poly-cooccurrente est apparue dans la branche chinoise par rapport aux réseaux cooccurrentiels de 2006-2014 (figure 7.17 ci-dessus), à savoir, la forme 芬兰 /fēn lán/Finlande.

Toutes les autres formes poly-cooccurrentes parallèles restent identiques aux cooccurrents de CLRG.

Pourquoi cette apparition du mot *Finland* dans le CLRG 2006-2014 ?

Nous rappelons que le mécanisme de l'attraction lexicale des poly-cooccurrences a donné davantage de résultats que le calcul des cooccurrences. Dans ces deux types de calculs, nous avons maintenu les paramètres par défaut du logiciel Trameur (co-fréquence 2, seuil 10, contexte . !?). Le paramètre « contexte . !? » signifie que l'unité de calcul des poly-cooccurrences se base approximativement sur chaque phrase du corpus. En effet, la poly-cooccurrence fonctionne par l'attraction lexicale du renouvellement de formes dans chacune des unités phraséologiques, elle récupère les nouvelles formes qui sont répétées le même nombre de fois que la forme-pôle choisie dans la phrase ou la séquence qui

suit. L'absence des enchainements de mots dans les résultats cooccurrentiels s'explique par l'absence de répétition de nouvelles formes dans la phrase suivante. Ce phénomène est dû au mécanisme anaphorique de cette langue. L'utilisation des anaphores, appelée en pragmatique linguistique, la deixis, le déictique ou l'emploi déictique, due au bon usage de la plupart des langues occidentales serait la véritable cause de ce phénomène. C'est la raison pour laquelle des nouvelles formes peuvent apparaître dans les réseaux poly-cooccurrentiels.

Nous faisons l'hypothèse suivante : la nouvelle technologie de l'EPR a suscité un certain engouement avant le démarrage de leurs constructions en 2009 et en 2010 en Chine. Nous vérifions cette hypothèse par retour aux textes à l'aide de la carte des sections parallèles classée par année. Il s'avère que cette hypothèse est fausse. En effet, nous constatons en 2010 un événement lié au captage de CO2 (CCS) en rapport avec des sociétés finlandaises de ce domaine, telles que *Pöyry*, et un autre lié aux sociétés de production d'électricité, telles que *Fortum*. Lors de la traduction de ces entités nommées, la forme *Finland* n'a pas été employée en anglais. A cause de la répétition du mot 芬兰/fēn lán/Finlande en chinois, le captage de formes poly-cooccurrentes en chinois est plus dense et plus large que le réseau cooccurrentiel.

A la demande du gouvernement britannique, la société finlandaise *Pöyry* a réalisé une étude à propos des coûts de réduction concernant la capture des émissions de carbone et le stockage du dioxyde de carbone. *Fortum*, basé en Finlande, était un des principaux services publics producteur d'électricité pour les pays nordiques. Les services publics appartenant à l'Etat avaient le monopole de toute la production d'électricité scandinave. Or en 2010, le marché de l'électricité s'ouvre vers d'autres moyens de financement.

Nous examinons maintenant la période retenue 2010-2012.

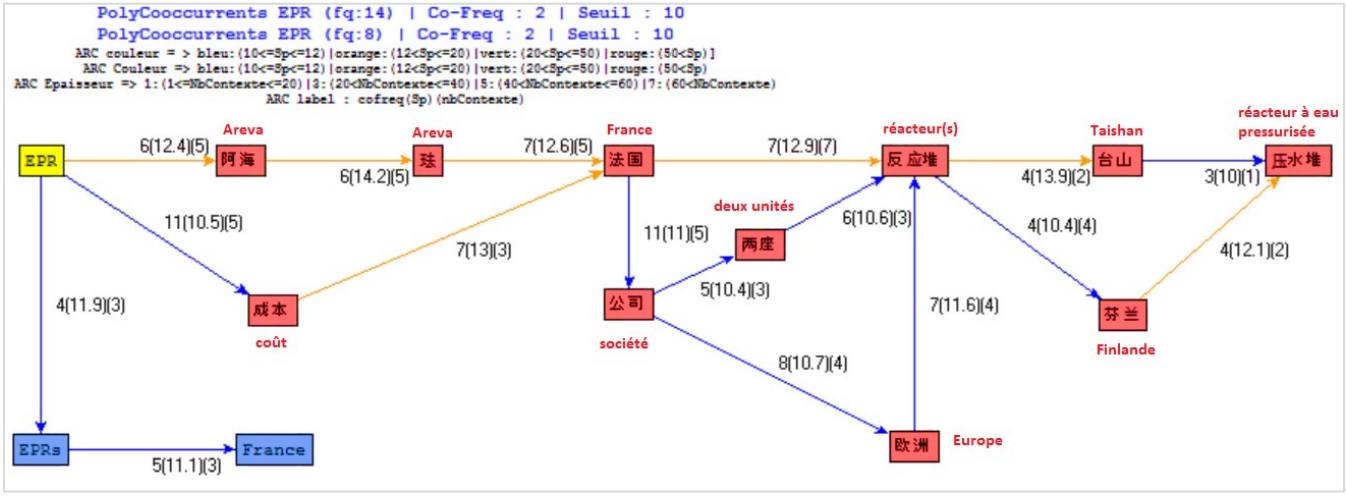


Figure 7.22 CLRG de 2010 à 2012 : réseau poly-cooccurrentiel parallèle de la forme *EPR*

D'après la figure 7.22, sur les trois années, dans le volet chinois, 11 poly-cooccurrents associés à la forme-pôle *EPR* sont apparus, contre seulement 2 poly-cooccurrents dans le volet anglais.

Par rapport à la période complète (2006-2014), 2 poly-cooccurrents de la période 2010-2012 ont disparu de la branche chinoise, à savoir, *CGN* et *conception*, tandis que la branche anglaise voit la seule disparition de *Finland*. En revanche, la forme *Finland* existe toujours dans le réseau poly-cooccurrentiel chinois. Encore une fois, cette disparition des 3 poly-cooccurrents dans les deux

branches ne change pas notre restitution d'informations sur la construction d'EPR en Chine. Pour tenter d'expliquer ce phénomène, un retour au contexte est nécessaire.

La forme *Finland* est attachée aux formes *Pöyry* et *Fortum*. La première forme *Pöyry* se trouve dans un article du 02/03/2010 et la seconde *Fortum* dans un autre article du 30/07/2010, deux articles datant de 2010.

Dans le volet chinois, on trouve une fois la forme *Finland* dans chacun des deux articles, mais pas dans le volet anglais. Un extrait de l'article daté du 02/03/2010 est disponible dans le tableau 7.10 ci-dessous. Nous pouvons constater que les deux formes *Poÿry* (贝利/bèi lì/ Poÿry) et *Finland* (芬兰/fēn lán/Finland) sont présentes dans le volet chinois. Or, dans le volet anglais, la forme *Poÿry* est présente mais la forme *Finland* est absente. Quand cette forme a-t-elle disparu ?

Tableau 7.10 CLRG 2010 : paragraphe parallèle de l'article daté du 02/03/2010 avec les formes *Poÿry* (贝利/bèi lì/ Poÿry) et *Finland* (芬兰/fēn lán/Finland)

Volet chinois	Volet anglais
Présence simultanée des deux formes ***Poÿry*** et ***Finland***.	La forme ***Poÿry*** présente mais absence de la forme ***Finland***.
那么成本究竟有哪些？据能源顾问公司芬兰贝利集团公司为英国政府做的一份研究显示，2015 年燃煤或燃气电厂每吨二氧化碳的碳捕获减量成本将有可能在$41（￥280）到$57（￥389）左右。而地下咸水层二氧化碳储存及监控成本将在每吨$1.6（￥10.9）到$3.2（￥21.8）左右。政府间气候变化专门委员会关于 CCS 所做的一份专题报告认为，250 公里长的输运管线每年输运 5 百万吨二氧化碳，则每吨二氧化碳的运输成本为$1（￥6.8）到$6（￥41）左右。可是，实际数据受具体地点及土地成本的影响很大。	*So what are the costs? According to a study by energy consultancy **Pöyry** for the UK government, the abatement cost for carbon capture in 2015 is likely to be US$41 (280 yuan) to US$57 (389 yuan) per tonne of carbon dioxide for capturing from coal or natural-gas power stations, and the cost for storing and monitoring carbon dioxide in saline aquifer US$1.60 (10.9 yuan) to US$3.20 (21.8 yuan) per tonne of carbon dioxide.*

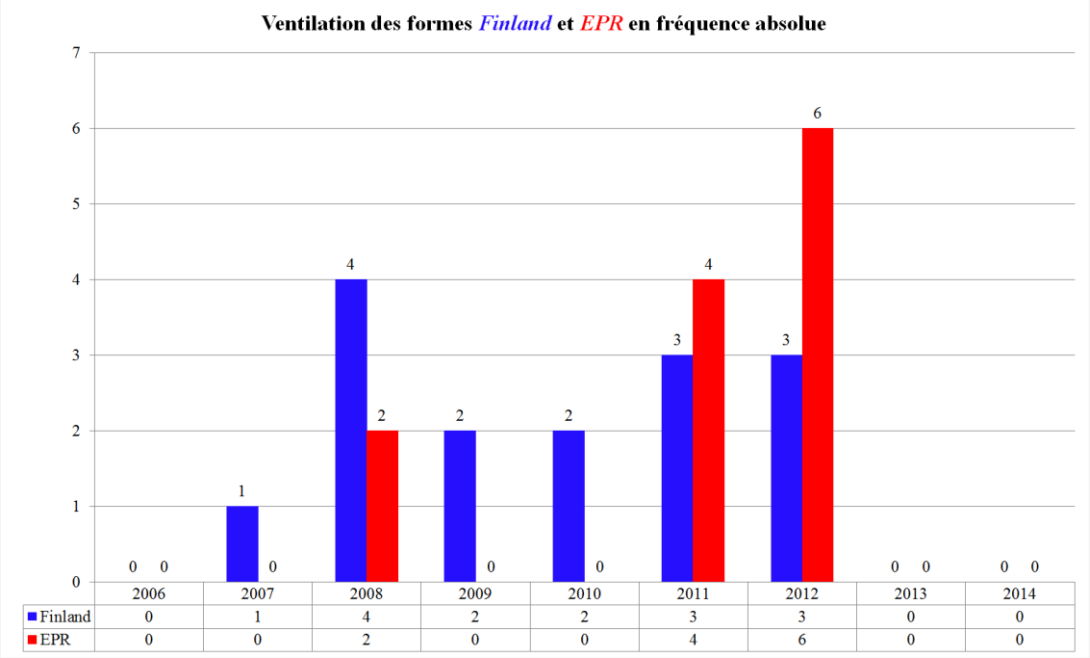
Ce phénomène de répétition est reproduit dans l'article daté du 30/07/2010, article où les auteurs chinois ont précisé le pays d'origine de la société *Fortum* (富腾/Fù téng/**Fortum**) dans la version chinoise, mais pas dans la version anglaise, comme le montre le tableau 7.11 ci-dessous.

Tableau 7.11 CLRG 2010 : paragraphe parallèle de l'article daté du 30/07/2010 avec les formes 富腾/Fù téng/Fortum et *Finland* (芬兰/fēn lán/Finland)

Présence simultanée des deux formes ***Fortum*** et ***Finland***.	La forme ***Fortum*** présente mais absence de la forme ***Finland***.
实际上，斯堪的纳维亚所有的发电产业，无论可再生还是常规的，都曾经全部属于 Vattenfall（瑞典）、富腾（**Fortum**，芬兰）和 StatKraft（挪威）等国有公司。我们进入了一个没有此类项目长期银行融资先例的市场。瑞典还有一个新的支持系统，即以英国模式为基础的绿色证书体系（这是一个国家交易机制，用认证书证明来自可再生能源的电力的买卖情况）。	*Virtually all Scandinavian power generation, renewable or conventional, had been owned by state-owned utilities – Vattenfall, **Fortum**, StatKraft. We went into a market that had no history of long-term bank financing for this kind of project.*

Donc, l'apparition de la forme *Finland* dans le réseau poly-cooccurrentiel anglais de 2006 à 2014 est due à la répétition de cette forme avant l'année 2010, comme illustré par la figure 7.23 ci-dessous. En

effet, la chaine de poly-cooccurrences anglaises est interrompue en 2010 en raison de l'absence de la forme *Finland*. La ventilation en fréquences absolues des formes *EPR* et *Finland* montre que la forme *Finland* a été employée 4 fois en 2008, cette répétition a déclenché le mécanisme de captage des poly-cooccurrences fédérées autour de l'EPR.



	2006	2007	2008	2009	2010	2011	2012	2013	2014
Finland	0	1	4	2	2	3	3	0	0
EPR	0	0	2	0	0	4	6	0	0

Figure 7.23 CLRG_EN de 2006 à 2014 : ventilation par année de la forme *Finland* et *EPR* en fréquence absolue

7.7 Synthèse des analyses du corpus parallèle CLRG

L'expérience des cooccurrences et poly-cooccurrences parallèles sur l'EPR semble attester que le réseau cooccurrentiel est plus sûr et efficace pour la restitution d'informations contextuelles, tandis que le réseau poly-cooccurrentiel permet de capter plus d'informations, y compris les bruits.

Rappelons que le discours du corpus parallèle est celui des ONG et des institutions indépendantes. A travers les études bilingues que nous avons menées autour des thèmes changement climatique et énergies, et également grâce aux analyses des segments répétés du corpus, nous sommes amenés à constater un vocabulaire plus spécialisé, thématisé, et plus proche de nos critères de recherches que dans le corpus comparable (*cf.* tableau 7.8, Segments les plus répétés des deux volets du corpus CLRG de 2006 à 2014).

Les AFC sur tous les mois du corpus CLRG attestent l'extrême convergence des informations dégagées et véhiculées par les formes employées dans les discours des ONG, à la différence des instances étatiques ou privées.

L'approche des ONG apporte une clarté technique et une complémentarité informationnelle indispensables à la demande de la transparence de la société actuelle face à la prolifération des informations gigantesques, diverses et multicanales, mais aussi parfois filtrées.

L'un des avantages de ce corpus parallèle est, d'une part l'enrichissement des ressources de lexiques dictionnairiques et traductologiques, d'autre part l'approche critique et revendicatrice des ONG par les mots qu'elles emploient.

7.8 Analyses transversales des deux corpus : poly-résonances croisées et faits translinguistiques

Les explorations textuelles centrées sur la veille trilingue que nous avons réalisées nous mènent vers de grands horizons informationnels grâce à la textométrie. Nous tenterons de retracer dans les lignes qui suivent une vision panoramique de nos découvertes «trans-heuristiques» et de les valoriser pour la veille d'informations multilingues. Cette méthode heuristique reposant sur le croisement des poly-résonances textuelles multilingues et des analyses événementielles et informationnelles, est appelée poly-résonances croisées.

***Energie(s), nucléaire* et *EPR* dans les corpus ENRG et CLRG**

Tableau 7.12 ENRG et CLRG de 1999 à 2014 : *énergie(s)* et *EPR* dans les corpus

	énergie	***énergies***	***nucléaire***	***EPR***
ENRG_FR	nucléaire	renouvelables	toponymes, types ou sources d'énergie (vent, solaire etc.), catastrophes, activités nucléaire, électricité	s'enliser, déboire
ENRG_US	renouvelables, propres, sources, climat, solaire, efficacité, puissance		toponymes, accidents, sources d'énergie (charbon, vent, etc.)	*Business is business.*
ENRG_CN	transition énergétique, solaire, éolienne, production, sylviculture, émissions, propres		toponymes, inventaire des réacteurs	besoins, ambitions
CLRG_CN	consommation, énergie solaire, charbon, approvisionnement d'énergie, émissions		toponymes et éponymes des accidents, antinucléaire, déchets, retraitement	dénouement d'un rêve (逝去的梦, Un rêve disparu)
CLRG_EN	économiser, appareils, réductions, Plan (quinquennal)		toponymes, et éponymes des accidents, déchets, retraitement, Areva, EDF, Taiwan	dénouement d'un rêve (*A dream unwinding*)

Les résultats analytiques issus des réseaux cooccurrentiels et poly-cooccurrentiels de tous les corpus sur le continuum complet ont été synthétisés dans le tableau 7.12 ci-dessus.

Dans un premier temps, nous nous apercevons que l'énergie en France est dominée par le nucléaire, où la forme *nucléaire* n'est qu'un méronyme de l'énergie. Dans les deux autres sphères, celle-ci demeure un hyponyme ontologique. Le deuxième constat est celui sur la forme *EPR*, qui se décline par la perception de sa construction dans les trois sphères. Le troisième constat concerne la Chine et les États-Unis qui se signalent par leur volonté politique de construire un avenir plus *propre* en matière énergétique et environnementale en opérant une *transition énergétique*, en particulier la Chine. Concernant la forme-pôle *nucléaire*, les formes telles que *déchets et retraitement* sont plus saillantes dans le corpus parallèle CLRG, cela s'explique par le genre catégoriel du discours dans lequel les sujets associés aux retraitements radioactifs restent l'un des chevaux de bataille des ONG.

Faits translinguistiques

Au travers des études de nos deux corpus, des faits linguistiques relativement simples et évidents semblent apparaitre, comme la notion de répétition et d'anaphore. Mais en réalité, ce n'est pas aussi simple, encore moins, si on les observe de manière translinguistique. D'une langue à l'autre, le phénomène change. Nous avons donc décidé de quantifier certaines de ces répétitions. Rappelons que la répétition de mots grammaticaux est une remarque saillante pour la langue anglaise, tandis que la répétition de termes de contenu est une spécificité réservée à la langue chinoise (*cf.* section 7.2.1).

Nous montrons un exemple concernant le français et l'américain (*cf.* figure 7.24 ci-dessous). Dans le corpus comparable, nous avons retenu 4 couples de formes, *énergie* vs *energy*, *climat* vs *climate*, *nucléaire* vs *nuclear*, *réchauffement* vs *warming*.

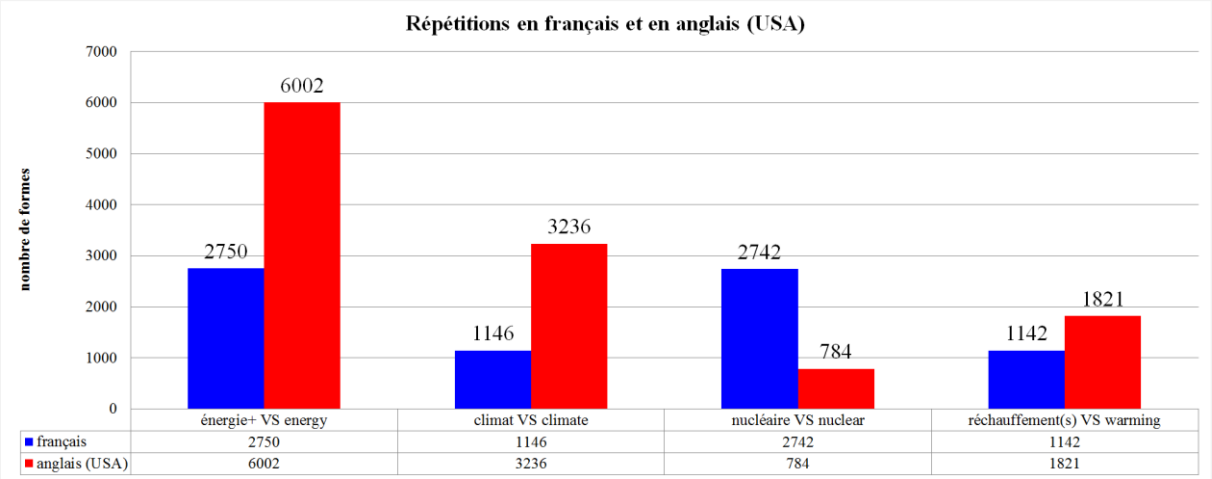


Figure 7.24 ENRG_FR et ENRG_US de 1999 à 2012 : quantification des répétitions en français et en anglais

Nous obtenons les constats suivants : pour les couples de formes *énergie* vs *energy*, *climat* vs *climate*, *réchauffement* vs *warming*, le nombre de répétitions est plus important en anglais qu'en français. Quant au couple *nucléaire* vs *nuclear*, c'est le français qui domine largement, à savoir presque trois fois et demi. La question est de savoir, si la forme *climate* en anglais peut jouer un rôle d'adjectif. En effet, l'association des termes comme *climate change* ou *climate system* reste relativement récurrente en anglais. Le même phénomène s'est produit pour la forme *nucléaire* en français avec un nombre largement supérieur. Ceci s'explique par le fait que la forme *nucléaire* peut être à la fois nom et adjectif.

Pour étayer notre propos, nous avons choisi un autre terme relativement difficile à transformer en adjectif, à savoir la forme *réchauffement(s)* vs *warming*. On obtient plus de formes en anglais.

Nous pouvons donc en déduire que d'une part le phénomène de la répétition est plus prononcé en anglais qu'en français, d'autre part, le mécanisme des anaphores est plus employé en français.

Et en chinois ? Que se passe-t-il autour des répétitions et des anaphores ?

Pour ce faire, nous avons préféré poursuivre ces recherches à partir du corpus parallèle CLRG où les données textuelles sont moins éclectiques et plus comparables, à l'aide des résultats relatifs aux segments répétés et poly-cooccurrences.

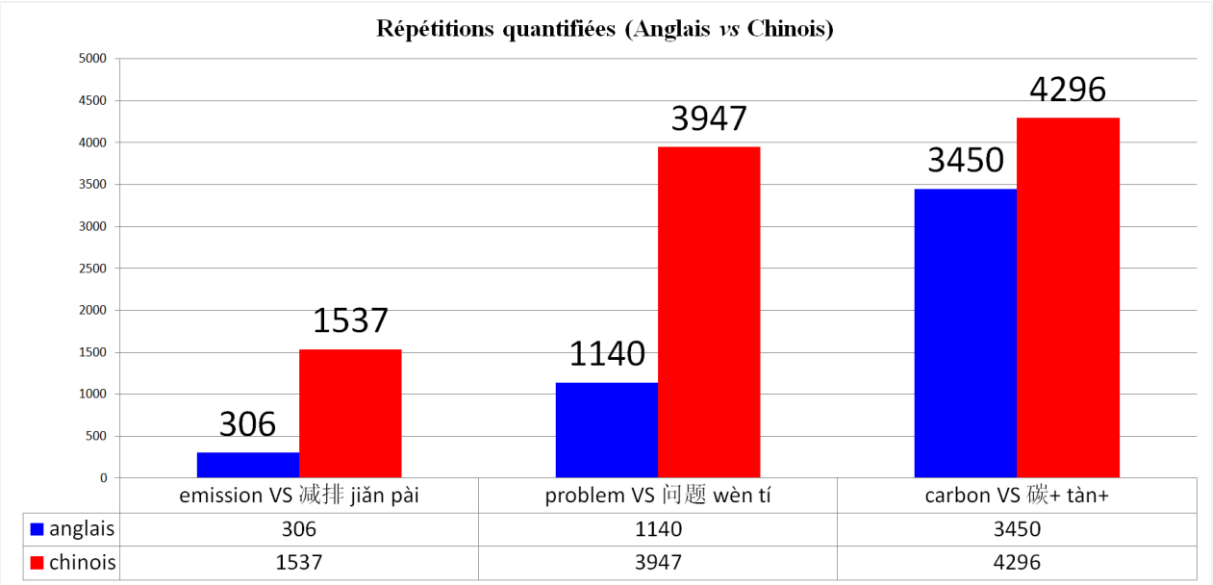


Figure 7.25 CLRG de 2006 à 2014 : quantification des répétitions en anglais et en chinois

Rappelons que parmi les segments les plus répétés dans chacun des volets, obtenus à partir de notre corpus CLRG (*cf.* section 7.2.3, tableau 7.8), les termes de contenu sont classés en haut du tableau, tandis qu'en anglais, ce sont des mots-outils ou des mots grammaticaux qui sont les plus répétés.

Dans les réseaux cooccurrentiels parallèles issues de CLRG (*cf.* section 7.5, figure 7.17) avec la forme *EPR*, seulement 3 formes sont corrélées en anglais. Or, dans le volet chinois, nous avons obtenu beaucoup plus de formes, telles que « coût, réacteur, Areva, CGN » etc.

Ce phénomène de répétitions en chinois est fréquent dans ce corpus, un exemple caractéristique est disponible dans la section 7.5, figure 7.19, phrase soulignée en rouge. Dans cet exemple où l'entité nommée *EPR* est fortement corrélée avec la forme 成本/chéng běn/coût/*cost*, nous constatons que dans le volet chinois, la forme 成本/chéng běn/coût/*cost* a été répétée 6 fois, tandis que dans le volet anglais, elle apparait seulement 2 fois. Cet exemple illustre bien ce phénomène de répétition de termes de contenu en chinois.

Pour aller plus loin, nous avons choisi trois termes parfois difficiles à transformer en anaphore afin de mieux observer ce phénomène de répétition.

Ces trois couples de formes sont *emission* vs 减排/jiǎn pái/réduire les émissions de CO2, *problem* vs 问题/wèn tí/ problème et *carbon* vs 碳/tàn carbone. Dans la figure 7.25 ci-dessus, nous remarquons que la répétition est constamment plus forte en chinois qu'en anglais.

Pour la néologie 减排/jiǎn pái/réduire les émissions de CO2, il est à noter qu'en anglais la simple forme *emission* couvre des champs sémantiques beaucoup plus larges que pour le chinois, on aurait pu s'attendre à plus de répétitions dans le volet anglais.

Dès lors, nous pouvons dire qu'il y plus de répétitions en chinois, un peu moins en anglais, encore moins en français.

Ce phénomène est dû sans doute essentiellement au mécanisme des anaphores ou au mécanisme déictique qui n'est pas le même en français et en anglais.

Dans le dessein d'une exploitation plus diversifiée de la linguistique, en l'occurrence la sémantique et la micro-sémantique pour mieux modéliser la veille et la recherche d'informations, nous proposons une autre méthode, inspirée du mécanisme de l'ontologie et des jeux de la sémantique interprétative de Rastier (Rastier, 1987, 2011 ; Rastier et Pincemin, 1999 ; Valette, 2004, 2008 ; Valette et al, 2006 ; Valette et Rastier, 2006 ; Valette et Slodzian, 2008) et exposée dans le tableau 7.13 ci-dessous.

7.9 Méthode analytique par Objets-Traits-Entrées (OTE) en classe sémantique et ontologique

Tableau 7.13 Méthode analytique par Objets-Traits-Entrées (OTE) en classe sémantique et ontologique

Classe sémantique et ontologique		Traits communs et saillants (connotations et sens)			Entrées
Objets (ou notions)	**classes et niveaux**	**Langue A**	**Langue B**	**Langue C**	**brèches communes permettant d'élargir les champs de la recherche d'informations**
	hyperonymie				
	hyponymie				
	méronymie				
	métonymie				
	pantonymie				
	paronymie				
	rétronymie				
	synonymie				
	(…)				

Cette méthode analytique par Objets-Traits-Entrées (OTE) en classe sémantique et ontologique, présentée dans le tableau 7.13 ci-dessus, consiste à classer les formes souvent statistiquement repérées dans le corpus par leurs classes et niveaux sémantiques et ontologiques. Par la suite, nous analysons ces formes en extrayant les traits sémantiques saillants et communs, langue par langue, afin de déduire des éventuelles entrées ou brèches communes permettant d'élargir les champs de la recherche d'informations.

Le tableau 7.14 ci-dessous est une application concrète de cette méthode OTE à partir de nos résultats disponibles dans l'annexe L. En effet, les explications et justifications du remplissage des cases du tableau ont été étayées et élaborées à partir des résultats issus du tableau L.1 (*cf.* annexe L).

Sémantique multilingue analytique de nos résultats

Tableau 7.14 Application de la méthode analytique par Objets-Traits-Entrées (OTE) en classe sémantique et ontologique

Classe sémantique		Traits communs et saillants (connotations et sens)			Entrées
objets (ou notions)	**classes et niveaux**	**français**	**anglais**	**chinois**	**brèches communes permettant d'élargir les champs de la recherche d'informations**
énergie	hyperonymie	nucléaire	climat, puissance, efficace	production, propre, émissions, charbon, hydraulique, géothermie, biomasse, sylviculture	**éléments naturels (feu, eau, vent, terre, mer, atome, arbre, etc.)**
énergies	hyperonymie	éolienne, solaire, marémotrice (houlomotrice)			
nucléaire	hyponyme et méronyme de l'énergie	politique, sortir	politique, tabou, un moyen pour réduire les émissions (GES)	politique, énergie propre, un moyen pour réduire les émissions (GES)	**réacteur(s), matière(s) et élément(s) radioactif(s)**
EPR	méronyme du nucléaire	retards, coût	*business*	expansion énergétique	**réacteur nucléaire de troisième génération**

Comme le démontre le tableau 7.14 ci-dessus, la méthode OTE en classe sémantique et ontologique permet d'une part de synthétiser nos résultats issus de nos recherches, d'autre part d'ouvrir des entrées ou brèches, c'est-à-dire, de nouveaux points d'entrées permettant d'élargir les champs de la recherche d'informations. Selon ce même tableau, nous partons des notions *énergie*/*énergies* en trois langues pour arriver à cerner ces mêmes notions par les éléments nouveaux tels que feu, eau, vent, terre, mer, atome, arbre, sylviculture etc. A propos de la notion *nucléaire*, les formes comme *réacteur(s)* ou *radioactif(s)* vont permettre de trouver des informations corrélées avec l'énergie de l'atome. Quant à *EPR*, l'anaphore *réacteur nucléaire de troisième génération* sera un nouveau point d'entrée pour les explorations et extractions de connaissances relatives à la troisième génération du nucléaire.

Conclusion du chapitre

Dans ce chapitre, notre analyse textométrique a permis de relever un nombre d'occurrences légèrement supérieur dans le volet anglais par rapport au volet chinois. Cela s'explique par les aspects linguistiques de la langue chinoise qui est une langue plus compacte et polysémique que l'anglais. Par ailleurs, le nombre de formes et d'hapax se sont révélés supérieurs dans le volet chinois, ce phénomène est dû au mécanisme de formation des mots en chinois. La répétition de mots grammaticaux est une remarque saillante pour la langue anglaise, tandis que la répétition de termes de contenu est une spécificité réservée à la langue chinoise. De manière générale, nous constatons que le mécanisme langagier reposant sur la répétition de notions est plus prononcé en chinois, un peu moins en anglais, encore moins en français. En ce qui concerne la forme *EPR*, la ventilation par mois des segments répétés associés à *EPR* a montré que cette forme est spécifique au mois de mars 2011 (survenue de la catastrophe de Fukushima) et au mois de mai 2012 (élection à la Présidence de la République de François Hollande dont la ligne écologiste (Laurent, 2011) est différente de celle de Nicolas Sarkozy). L'analyse des cooccurences et poly-cooccurrences de la forme *EPR* dans le corpus parallèle a permis de montrer aussi que le chinois a un système discursif qui évite moins la répétition de notions que le système discursif anglais ou français.

Enfin, l'analyse transversale des deux corpus a mis en lumière que le sous-corpus français voit la forme nucléaire comme un méronyme de l'énergie, alors qu'il ne s'agit que d'un hyponyme ontologique dans les autres sous-corpus. Les résultats des analyses des formes liées à *EPR* du sous-corpus français témoignent des déboires et des échecs de ce réacteur. En revanche, dans le sous-corpus américain, les analyses illustrent la volonté exportatrice des États-Unis et nous font penser à la célèbre formule "*Business is business*". Quant au sous-corpus chinois, les résultats apparaissent liés à des ambitions politico-économiques et à des besoins sociétaux et révèlent également une volonté de réduire la pollution.

Conclusion générale

L'objectif de notre recherche était d'élaborer une série de méthodes se rapportant à la veille textométrique multilingue. Nous avons choisi de les appliquer à deux ensembles de textes, dont l'un porte sur le discours de presse et l'autre sur celui des ONG. Ces corpus sont centrés sur deux thèmes d'actualité : les énergies et l'environnement. Plusieurs méthodes informatiques, statistiques et linguistiques ont été articulées à partir de ces deux corpus chronologiques s'étalant sur les années 1999 à 2014.

Afin de mieux entreprendre nos analyses multilingues, nos deux études portent sur deux types de corpus en trois langues différentes, celles qui dominent les échanges économiques mondiaux :

- un corpus *comparable* ENRG trilingue, français, anglais et chinois, constitué de trois sous-corpus, ENRG_FR, ENRG_US et ENRG_CN,
- un corpus *parallèle* CLRG bilingue anglais-chinois, dont les volets sont dénommés respectivement CLRG_EN et CLRG_CN.

Les corpus comparables et parallèles sont deux catégories de données textuelles pour les études scientifiques : la première catégorie a pour avantage, lors de sa constitution, une liberté de choix relatifs aux sources, genres de discours, périodes, langues, etc. et une accessibilité large et facile à toutes ces informations, nonobstant les hétérogénéités des données complexifiant les traitements et les analyses interprétatives ultérieures. Les corpus comparables permettent notamment d'obtenir des complémentarités lexicales et informationnelles par le biais de leurs convergences et divergences sémantiques. Tandis que la seconde catégorie, celle des corpus parallèles, très utile pour l'extraction des lexiques bilingues dictionnairiques et traductologiques, se limite à des informations énoncées en deux ou plusieurs langues où la symétrie informationnelle reste prépondérante. De ce fait, il est difficile de transmettre le style rhétorique de la langue de départ vers la langue d'arrivée, en raison des aléas du sens de traduction et du niveau hétérogène des traducteurs employés par les structures bilingues ou multilingues concernés. L'approche idéale est de combiner les deux types de corpus malgré les coûts conséquents d'investissements à la fois matériels et intellectuels.

Les deux corpus ont constitué le continuum textuel discursif dans lequel nous avons examiné l'émergence événementielle des formes textuelles liées aux énergies, en trois langues, dans le continuum informationnel dans les domaines de l'environnement et de l'énergie. Cet espace-temps est à la fois « trans-heuristique » et trans-catégoriel, à savoir, d'une part, il permet de découvrir les différents objets associés à notre recherche, tels que les énergies, les réacteurs nucléaires, l'environnement, les scandales chinois, les conséquences des catastrophes, etc., et d'autre part, il exploite des sources d'informations diverses, telles que la presse, les sites d'information, les médias et les ONG.

Conclusion

Nous avons commencé notre recherche par un rappel théorique et une brève description de l'état de l'art consacrés plus particulièrement aux domaines associés à notre recherche : veille, intelligence économique, traitements automatiques des langues et textométrie. Puis, nous avons présenté les types d'énergies et l'environnement dans les contextes socio-économiques des trois pays retenus, France, États-Unis et Chine, ainsi que les spécificités de leurs presses respectives, les caractéristiques et technicités de leurs langues et les algorithmes de la segmentation du chinois. Par la suite, nous avons mis en pratique nos méthodes textométriques et analytiques transdisciplinaires sur les différents corpus. Elles permettent d'abord de mieux structurer les tâches de veille technologique devenues partie intégrante de l'activité des grandes entreprises modernes et d'élargir à l'ensemble des acteurs de l'économie mondiale la préoccupation de recherche d'informations à caractère industriel et commercial. Dans un deuxième temps, les mêmes méthodes mettent en évidence la manière dont sont perçus et formulés, dans les différentes langues et cultures, les principaux concepts de chaque domaine, énoncés plus haut. Enfin, elles aboutissent à la constitution de ressources traductologiques, dictionnairiques et rhétoriques bilingues adaptées à chaque secteur de recherche.

Nous rappelons quelques points essentiels évoqués au cours de nos recherches.

L'intelligence économique (I.E) se distingue par l'aspect de pilotage qu'elle met en œuvre sur le plan de la conception générale dirigée d'un projet à enjeux économiques. Il s'agit de concevoir, d'ordonner et de planifier les séquences pratiques en amont de la veille, suivies de l'affectation de connaissances de « diorama[226] » aux informations stratégiques. Ainsi des projets de veille se définissent, qui permettent de désigner les objectifs et les processus de traitements concrets. La veille textométrique multilingue, pour laquelle nous avons proposé une série de traitements informatiques, statistiques et analytiques, manipule plus directement la matière textuelle. Il s'agit d'une méthode empirique qui détermine, dans un corpus donné, les caractéristiques textuelles et informationnelles intrinsèques des différentes parties des échantillons textuels prélevés, sur lesquelles les exégèses multilingues, spécialisées et contextuelles sont indispensables, afin de mettre en évidence la véracité et l'authenticité informationnelle et événementielle véhiculées dans les discours.

L'I.E et la méthodologie de veille ont permis de déterminer le contexte général de nos thèmes de recherche. Les objectifs et processus de veille multilingue, au-delà de nos corpus construits, sont également abordés. Par la suite, nous avons mené des études de veille et de restitution d'informations via les mots-clés, réseaux cooccurrentiels et poly-cooccurrentiels multilingues, dont chacun comporte des contenus fragmentaires d'informations, fragments parfois détachés de contextes ou parfois fédérés à partir des informations principales d'un événement. La distribution de ces suites de formes cooccurrentes et poly-cooccurrentes, calculée par leur aspect quantitatif, est soumise à des contraintes contextuelles. La carte des sections de la textométrie illustre la distribution d'une unité textuelle dans le corpus et permet d'accéder à son contexte local. Quant à l'aspect qualitatif, nous avons tenté de comprendre et de restituer des informations associées à ces deux types de réseaux, dans l'objectif d'interpréter et de consolider nos analyses par le travail de création de dictionnaires d'événements que nous appelons le contexte général.

Cette intégration méthodologique située à l'intersection de diverses disciplines tant littéraires que scientifiques, axée sur l'approche multilingue, a pour but d'analyser de plus près la manière dont chacune des langues choisies relatent les informations. Dans un premier temps, des informations se décèlent à travers des traitements et analyses appliqués sur un premier corpus monolingue construit,

[226] Définition du TLFi : Tableau ou suite de tableaux de grandes dimensions, en usage surtout au XIX^{e} siècle, qui, diversement éclairé(e), changeait d'aspect, de couleur et de forme, était agrémenté(e) ou non de premiers plans en relief et donnait aux spectateurs l'illusion du mouvement. http://www.cnrtl.fr/definition/diorama

que nous appelons résonances textuelles monolingues. Dans un deuxième temps, les mêmes traitements sont réitérés dans les autres corpus (monolingues ou multilingues). Par la suite, les informations émanant de tous ces différents corpus en différentes langues sont comparées transversalement afin d'obtenir des résonances textuelles multilingues. Ce processus est appelé poly-résonance textuelle.

Parmi les divers traitements sur les sous-corpus français et anglais, les analyses factorielles des correspondances ont d'abord été appliquées afin d'établir une typologie générale, de façon à mieux cerner la convergence et la répartition des informations événementielles véhiculées dans les mots spécifiques de nos échantillons textuels. Ainsi, des formes spécifiques désignant des fragments d'informations surgissent au fil du temps par leurs lexiques caractéristiques, parfois singuliers et propres à chaque période analysée.

Par la suite, nous avons recouru aux méthodes des cooccurrences et poly-cooccurrences évolutives. Ces méthodes ont permis de fédérer les éléments fondamentaux et les points d'ancrage des deux objets majeurs de la recherche, *énergie(s)* et *EPR*, ce dernier pouvant être considéré comme un véritable signal faible. Ces deux types de calcul ont mis en évidence des réseaux de vocabulaires récurrents et spécifiques à ces deux objets, ainsi que des indicateurs discursifs révélateurs sur l'axe du temps. Nos efforts ont été consacrés à la mise en contexte, créations de dictionnaires d'événements, et ont permis la restitution, la prévision et l'anticipation d'événements informationnels concernant nos objets recherchés. C'est bien la dimension dynamique de ce travail qui nous intéresse, la projection en direction de l'avenir permettrait de dessiner des perspectives dans de nombreux secteurs d'activités, tels que l'intelligence économique multilingue, la veille multilingue, l'ingénierie multilingue, la linguistique appliquée du chinois, etc.

Les mêmes traitements statistiques et informatiques ont été pratiqués sur tous les sous-corpus, notamment sur ENRG_FR, ENRG_CN et sur l'ensemble de CLRG. De surcroît, l'outil *segments répétés* se révèle particulièrement efficace pour l'étude des textes chinois. Cet outil permettant de repérer les séquences textuelles via leurs fréquences, récupère partiellement ou parfois intégralement les informations diffusées dans les formes non lexicalisées des dictionnaires informatiques.

Pour la première étude, trois sous-corpus de veille à caractère comparable ont été constitués de manière automatique grâce à des programmes informatiques spécifiques, dont les données textuelles trilingues ont été recueillies via quatre plateformes d'informations différentes, *Le Monde*, *New York Times*, *QQ.com* et *Sina.com.cn*, se trouvant dans trois pays distincts, à savoir la France, les Etats-Unis et la Chine. Les articles de ces trois sous-corpus trilingues couvrent une période comprise entre le 24 septembre 1999 et le 17 avril 2012 pour ceux en français et en anglais, et du 23 mars 2008 au 23 avril 2013 pour ceux en chinois, dates correspondant à l'exécution des programmes.

Dans le but de mieux mettre en pratique les méthodes de veille textométrique multilingue, les trois sous-corpus ont subi des traitements séparés en fonction de leurs proximités linguistiques et de leurs différences d'écritures, d'une part langues française et anglaise, d'autre part langue chinoise. Les soucis d'homogénéité et de comparabilité nous ont conduits à une sélection de sous-corpus annuels centrés sur les années 2010, 2011 et 2012, dans lesquels les mois de ces trois années sont analysés par clivages typologiques, clivages obtenus par les AFC.

Pour la deuxième étude, un corpus parallèle bilingue anglais-chinois a été construit de manière automatique. L'alignement par paragraphe des deux volets CLRG_EN et CLRG_CN a été effectué manuellement. Ces bi-textes allant du 6 juin 2006 au 27 août 2014 proviennent du site

Chinadialogue.net, un site indépendant et à but non lucratif, spécialiste de l'environnement, basé à Londres et à Beijing (Chine).

Obtenus dans la perspective de compléter l'approche additionnelle de la veille multilingue et d'apporter une certaine clarté au discours bilingue de la plaidoirie environnementale des ONG, les résultats de l'étude bilingue des segments répétés parallèles ainsi que leurs analyses montrent que, pour une même information énoncée et décrite en deux langues, la répétition événementielle et thématique est plus saillante en chinois.

L'entité nommée *EPR* contribue à une veille anticipative à haute valeur informationnelle. Les connaissances liées à celle-ci, extraites de nos deux corpus à l'aide des réseaux cooccurrentiels et poly-cooccurrentiels, ont été comparées. Les informations obtenues sont identiques et correspondent aux chantiers de construction actuels de l'EPR dans le monde. Celles-ci nous permettent de prévoir et d'anticiper les ambitions énergétiques internationales de la Chine.

La méthodologie employée dans notre étude a révélé un certain nombre d'avantages, d'inconvénients et de particularités pour la veille textométrique multilingue. Elle a également mis en évidence une volonté d'informer de situations préoccupantes et une force de propositions propre au discours des ONG.

Le travail mené dans ce cadre nous semble déboucher sur des apports à différents domaines, dont la linguistique.

Apports pour la veille transcontinentale

La veille trilingue utilisant les analyses textométriques appliquées aux corpus ENRG atteste de manière documentée que la presse des trois sphères manipule une sémantique commune (*protection de l'environnement*, *énergies renouvelables, réduction des émissions, changement climatique*) exprimée par des lexiques codés en trois langues. Notre travail a pu prouver plus particulièrement l'authenticité de cette sémantique qui était au départ inhérente à chacune des trois sphères de communication. Notre prévision et notre anticipation ont livré des informations stratégiques (*EPR* à *Hinkley Point*) pour le secteur énergétique. Ces travaux ont produit des contributions scientifiques précieuses à l'élaboration méthodologique de la veille transcontinentale. Cette sémantique est associée à l'une des préoccupations planétaires majeures de notre ère, le *changement climatique*, thème représenté par des formes lexicales récurrentes, générant une véritable prise de conscience écologique mondiale. De plus, le Vieux continent traditionnellement protecteur de l'environnement est en train de conduire la Chine et les États-Unis au partage de cette même conviction environnementale (*COP21*). Ce constat est attesté par les segments adjacents à la forme *émission* fréquemment répétés. L'un de nos objectifs était d'établir, d'évaluer et de prouver ce lieu commun, réflexion générale communément admise par l'opinion publique.

Les États-Unis et la Chine manifestent un intérêt grandissant pour les *énergies*, dites *nouvelles* ou *renouvelables* telles que le *soleil*, le *vent*, la *biomasse*, etc. Ces deux pays cherchent une énergie plus *puissante*, *efficace* et *propre*. Le pouvoir politique chinois est également préoccupé par la *transition énergétique*. Ces constats quantitatifs sont confirmés par les formes fédérées dans les réseaux cooccurrentiels et poly-cooccurrentiels autour du thème des *énergies renouvelables*.

Cependant, les résultats qualitatifs mettent en évidence la question de la sortie ou non du nucléaire, une question embarrassante en France pour le gouvernement et l'opinion publique. Cette interrogation est accentuée par les déconvenues rencontrées autour de la mise en œuvre de l'EPR dont les retours

d'expériences continuent de profiter au secteur nucléaire chinois dans le cadre de son expansion énergétique nationale et internationale, consolidée par les technologies américaines (*AP1000*).

Après les études de notre corpus comparable et de notre corpus parallèle, nous avons effectué des analyses transversales sur ces deux corpus qui nous apportent des éléments concrets. A partir de ces analyses nous avons proposé une nouvelle méthode permettant de créer des ouvertures utiles à la recherche sémantique multilingue. Nous remarquons une quantité notable de formes chinoises relatives aux nouvelles énergies telles que *géothermie*, *solaire*/*photovoltaïque*, *éolien*/*offshore*, *biologie*/*sylviculture*/*biogaz*, *gaz de schiste* etc. mais aussi des notions, comme *émission*, *consommations*, *approvisionnement d'énergies*, *économiser*, *réductions*, *plan quinquennal*. Ces ensembles forment un champ sémantique exprimant un besoin croissant et massif d'énergie, situation relativement critique compte tenu de la grande consommation d'électricité en Chine. Cependant, par prévision et anticipation, nous pouvons affirmer que ces *nouvelles énergies* doivent être davantage *planifiées* et *propres*, l'empire du Milieu est désormais contraint de respecter les conventions préconisées par les organisations internationales (Carfantan, 2014).

Apports pour la sémantique multilingue

Ces constatations viennent confirmer l'intérêt de la sémantique multilingue, cadre dans lequel la notion d'énergie révèle trois perceptions déclinées par leurs différentes connotations au travers des trois continents, à savoir, *productions* et *propre* pour les Chinois, *climat, puissance* et *efficacité* pour les Américains, *nucléaire* pour les Français.

La différence marquée entre singulier et pluriel de la forme *énergie* en français renvoie à un contexte spécifique dans lequel le singulier désigne presque toujours l'énergie *nucléaire*, tandis que le pluriel correspond plutôt aux termes *énergies renouvelables*, *énergies propres*, etc. employés dans les autres sous-corpus.

Quant à la veille bilingue par CLRG, elle nous fait découvrir, par des formes et des tournures critiques voire ironiques, mais souvent dénonciatrices, une forte conviction des ONG face à ces problèmes embarrassants.

Apports pour l'ingénierie et les moteurs de recherche en chinois

Dans notre recherche, plusieurs entités nommées du chinois ont été mal segmentées, problème récurrent dans les traitements automatiques de cette langue. La question de la segmentation en mots du chinois, pilier de tout traitement automatique des données textuelles, a également été approfondie dans ce travail. Un comparateur de segmentation chinoise a été spécialement conçu afin d'illustrer et de recenser les différences des résultats produits par les segmenteurs automatiques. Nos études antérieures (Shen, 2009) et nos recherches actuelles démontrent que la segmentation en idéogrammes isolés ne constitue pas une solution efficace pour la recherche d'informations en chinois tandis qu'une segmentation relativement plus élaborée en unités lexicales, comme celle réalisée dans le corpus comparable, apporte du bruit, bruit manifesté par des lexèmes (morphèmes lexicaux) non-pertinents.

Nous avons pu montrer qu'une segmentation, appelée *intelligente*, tenant compte au maximum des notions propres et/ou complètes, telles que les toponymes, les éponymes, les entités nommées, les notions spécifiques, les néologismes, les terminologies, etc., constituerait une solution optimale pour la veille en chinois. En effet, notre recherche a apporté la preuve que dans la veille chinoise, plus la collecte des termes spécifiques est complète, plus grande sera la rapidité de l'accès aux informations.

Ce phénomène de formation de mots chinois souligne que l'amélioration des moteurs de recherche nécessite une indexation constante des mots, termes, noms, entités nommées, etc.

La dissymétrie des réseaux cooccurrentiels et poly-cooccurrentiels issus des deux volets de CLRG démontre que la fonction discursive et les mécanismes syntagmatiques, paradigmatiques d'un discours entre deux langues lointaines, telles que anglais *versus* chinois (voire français *versus* chinois), ne peuvent être transposés ni par le voisinage informationnel, ni par la traduction.

Apports pour les études de figures de rhétorique et styles littéraires

De manière empirique, la prégnance de constats sémantiques à travers les segments répétés, réseaux cooccurrentiels et poly-cooccurrentiels dans les textes en chinois nous ont permis d'établir, ce qui n'avait pas été mis en lumière précédemment, que les anaphores, forte tendance du bon usage du français et de l'anglais, sont une figure de style très peu pratiquée par les locuteurs chinois. Notre étude de la textométrie de corpus particuliers débouche sur un fait linguistique général qui différencie le chinois et les deux langues d'origine européenne.

Le présent travail montre que les linguistes possèdent tous les atouts et les outils pour mettre leurs compétences méthodologiques et analytiques au service de la veille et de l'intelligence économique. Les entreprises bénéficieraient ainsi de leurs aides précieuses en les impliquant davantage dans leurs perspectives stratégiques et commerciales.

Signaux faibles, une révélation anticipative

Les signaux faibles peuvent parfois être une révélation anticipative. Comme nous l'avons constaté, l'EPR est un signal faible dans nos corpus, en raison de sa très basse fréquence dans les discours trans-médiatiques à la fois issus de la presse et des ONG. Les démarches basées sur les segments répétés polarisés par la forme-pôle *EPR*, ainsi que son emploi déictique, *réacteur nucléaire de troisième génération*, ont permis de localiser les informations enfouies parmi la masse des articles.

Apports méthodologiques

Au-delà de la textométrie multilingue, notre étude a également mis en perspective des métriques pragmatiques susceptibles de contribuer aux différentes phases de la fouille d'informations multilingues : les stratégies de veille multilingue proposent un système fiable, flexible et fonctionnel ainsi qu'un processus intégrant les complémentarités issues de l'intelligence complétive ; la méthode prospective «évaluations et analyses par critères» ouvre des brèches efficaces d'investigation ; la méthode Objets-Traits-Entrées (OTE) élargit les champs de la recherche d'informations. Ces principes retenus par l'étude semblent être parfaitement exploitables dans les poly-résonances textuelles.

Perspectives d'ouverture

Lors de nos études antérieures (Shen et Salem, 2009), nous avons vu que la mise en correspondance de différents genres de discours, tels que presse, blog et forum, apporte des révélations particulièrement intéressantes. Ces constats concernent notamment l'authenticité de l'opinion publique à travers ces trois catégories de supports et la manière dont s'exprime le peuple en Chine, un pays où selon les clichés, la censure est constamment appliquée.

Une première ouverture, à l'issue de ce travail, consisterait à élargir nos méthodes à des analyses plus poussées sur les corpus provenant de sources multiples, telles qu'un grand nombre de journaux, toutes lignes éditoriales confondues, de réseaux sociaux et forums, des ONG, etc. réunis autour de thèmes

proches et comparables. Cette compilation de sources diverses et variées permettra de déterminer les indices discursifs, explicites, mais aussi implicites, constituant l'horizon d'un travail futur de veilles, restitutions, prévisions et anticipations multimodales.

Une deuxième ouverture nous conduirait à aller au-delà de la veille trilingue. Dans le cadre de cette recherche, seules trois langues ont été sélectionnées, le français, l'anglais et le chinois. Les véritables enjeux du multilinguisme seraient de pouvoir envisager le traitement d'un plus grand nombre de langues vivantes en appliquant les méthodes de la textométrie.

Une troisième ouverture serait d'étendre le champ de comparaisons des connaissances et outils utilisés, en diversifiant les méthodes de veille et de fouilles d'informations. L'incorporation de nouvelles méthodologies et techniques transdisciplinaires créera ainsi d'importantes complémentarités. Par exemple : l'accès aux volumes gigantesques de données serait amélioré grâce aux algorithmes des mégadonnées (Big data), des séquences textuelles récurrentes seraient captées par les outils d'extraction, comme par exemple, l'extraction par patrons morphosyntaxiques.

Enfin, nos méthodes devraient s'appliquer dans tous les secteurs soumis à la mondialisation qui exigent notamment l'anticipation et la veille stratégique. La conjoncture politique internationale impose une plus grande vigilance dans les domaines du renseignement, de l'espionnage et de la cyber-surveillance, compte tenu des enjeux concernant la transformation et la prolifération des informations : la sécurité est donc plus que jamais capitale.

Ainsi s'achève notre présente étude sur une devise du stratège Sun Tzu: dans « L'Art de la Guerre (600 av. J-C) », Chapitre 10 :

> *« De l'étude du terrain, le dixième chapitre de L'Art de la Guerre est basé sur l'art de la topographie. Le besoin en renseignement du chef ne se restreint pas à l'adversaire. »*

Il est primordial de connaître le terrain et d'en avoir une bonne perception pour l'élaboration de la manœuvre.

> *« Qui connaît l'autre et se connaît ne sera point défait ; qui connaît Ciel et Terre volera de victoire en victoire. »* - Sun Tzu

Glossaire

Les abréviations entre parenthèses précisent le domaine auquel s'applique la définition.

Abréviations :

sp Méthode des spécificités
sr Analyse des segments répétés
ling Linguistique
stat Statistique
sa Segmentation automatique
tal Traitement automatique des langues
tm Textométrie
vs Veille stratégique

accroissement de vocabulaire - (stat) relation du nombre de formes par rapport au nombre d'occurrences au cours du corpus. Ce calcul permet d'observer l'apparition de nouvelles formes au fur et à mesure que l'on avance dans le corpus.

analyse factorielle - (stat) famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des « facteurs » résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

anaphore - (grammaire) procédé consistant à rappeler un mot ou groupe de mots précédemment énoncé par un terme grammatical. (rhétorique) procédé visant à un effet de symétrie, d'insistance, etc., par répétition d'un même mot ou groupe de mots au début de plusieurs phrases ou propositions successives.

caractère - (sa) signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.

carte des sections - (sa) représentation graphique d'un caractère délimiteur sous forme de carré pour chaque délimiteur rencontré dans le texte.

caractère délimiteurs / non-délimiteurs - (sa) distinction opérée sur l'ensemble des caractères, qui entrent dans la composition du texte permettant aux procédures informatisées de segmenter le texte en occurrences (suite de caractères non-délimiteurs bordée à ses extrémités par des caractères délimiteurs).

concordance - (sa, tm) ensemble de lignes de contexte restreint se rapportant à une même forme-pôle.

cooccurrence - (sa) présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, etc.) des occurrences de deux ou plusieurs formes données issues du calcul de spécificités.

cooccurrence évolutive - (sa, tm) présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, etc.) des occurrences de deux ou plusieurs formes données issues du calcul de spécificités évolutives.

cooccurrent émergent - (vs, tm) unité résultant du calcul de cooccurrences évolutives qui est unique pour la période analysée du corpus.

cooccurrent stable - (vs, tm) unité résultant du calcul de cooccurrences évolutives qui est récurrente sur plusieurs périodes analysées du corpus.

corpus - (ling) ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.

corpus - (tm) ensemble de textes réunis et sauvegardés au format électronique, se servant de base à une étude assistée par les outils informatiques.

délimiteur de séquence - (sa) sous-ensemble des caractères délimiteurs de forme correspondant aux ponctuations faibles et fortes (en général – le point, le point d'interrogation, le point d'exclamation, la virgule, le point-virgule, les deux points, les guillemets, les tirets et les parenthèses). Le blanc se sert généralement de caractère délimiteur d'occurrence.

empan textuel - (tm) partie informatique du corpus. Chaque occurrence correspond à une coordonnée informatique, une position dans le corpus. Dans ce système un empan correspond à une position x1 à une position x2.

entité nommée - (tal) noms de personnes, organisations, entreprises, lieux, produits recherchés dans le texte et auxquels on attribue une étiquette désignant la nature de l'entité.

étiquetage - (tal) processus qui consiste à associer une étiquette indiquant une information (linguistique ou informationnelle) à un segment (par exemple, un mot, un groupe de mots, une phrase, un paragraphe, etc.) d'un corpus.

éponyme - (ling) (Celui, celle, ce) qui donne son nom à quelque chose ou à quelqu'un, à qui l'on se réfère, que l'on vénère.

expression régulière (ou rationnelle) - (sa, tal) suite de caractères en informatique décrivant un ensemble de chaînes de caractères possibles selon une syntaxe précise. Elle est largement utilisée dans les programmations et les éditions textuelles électroniques.

forme (ou forme graphique) - (sa, tm) archétype correspondant aux occurrences identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrences. En anglais type. En chinois, 词形 ou 字形（Pinyin 拼音：cí xíng ou zi xíng）.

fouille textuelle - (vs, tal) ensemble de techniques automatiques permettant la recherche d'informations et la découverte de connaissances nouvelles dans des bases de données textuelles.

fréquence - (sa) (d'une unité textuelle) nombre de ses occurrences dans le corpus.

fréquence d'un segment - (sr) (ou d'une polyforme) nombre des occurrences de ce segment dans l'ensemble du corpus.

fréquence maximale - (sa) fréquence de la forme la plus fréquente du corpus. En français, le plus souvent la préposition « de », en anglais, l'article « the », en chinois, la particule grammaticale « 的 de ».

fréquence absolue - (sa) fréquence en chiffre absolu d'une unité textuelle dans le corpus, sans se rapporter à d'autres facteurs.

fréquence relative - (sa) fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties, rapportée à la taille du corpus.

hapax - (sa) forme dont la fréquence est égale à un dans le corpus (hapax du corpus) ou dans une de ses parties (hapax de la partie).

holonymie - (ling) relation sémantique entre mots d'une même langue, partitive et hiérarchisée : un holonyme A d'un mot B est un mot dont le signifié désigne un ensemble comprenant le signifié de B.

hyperonymie - (ling) rapport d'inclusion entre des unités lexicales, considéré comme orienté du plus général au plus spécifique. (C'est l'inverse de l'hyponymie.) [Chien est dans un rapport d'hyperonymie avec basset, caniche, etc.]

hyponymie - (ling) rapport d'inclusion entre des unités lexicales, considéré comme orienté du plus spécifique au plus général. (C'est l'inverse de l'hyperonymie.) [Chien est dans un rapport d'hyponymie avec carnivore, animal, etc.]

hypothético-déductif, -ive, (adjectif) qui part d'une proposition, dont la vérité sera jugée à posteriori, et en déduit toutes les propositions qui en sont la conséquence logique. *Intelligence, méthode, science hypothético-déductive. Il resterait à examiner (...) ce qui se produit au niveau des opérations propositionnelles, où le langage des sujets se modifie de façon si caractéristique en même temps que le raisonnement des sujets devient hypothético-déductif* (J. Piaget, *Le Structuralisme,* Paris, P.U.F., 1968, p. 81).*On distinguera quatre* types de systèmes notionnels. a) *Systèmes hypothético-déductifs, élaborés par une pure théorie, possédant des caractères formalisables, et où les concepts sont clairement fonctionnels (ex. : mathématiques, logique)* (A. Rey, *La terminologie : noms et notions,* Paris, P.U.F., 1979, p. 45).

lemme - (sa) forme canonique du mot d'où sont dérivées les formes fléchies. En chinois, 词条（Pinyin 拼音：cí tiáo）ou 词目（Pinyin 拼音：cí mù）.

lemmatisation - (sa) regroupement sous une forme canonique (lemme) des occurrences du texte.

lexical - (ling) qui concerne le lexique ou le vocabulaire

lexicométrie - (tm) ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire d'un corpus de textes.

lexique - (ling) ensemble des mots d'une langue.

méronymie - (ling) la méronymie est une relation sémantique entre mots, lorsqu'un terme désigne une partie d'un second terme. Par exemple, bras est un méronyme de corps, de même que toit est un méronyme de maison. De façon plus théorique, la méronymie est une relation partitive hiérarchisée : une relation de partie à tout. Un méronyme d'un mot M est un mot dont le signifié désigne une sous-partie du signifié de M. La relation inverse est l'holonymie.

multilingue - selon le TLFi, est Qui est rédigé en trois langues ou davantage.

multilinguisme - subst. masc. État d'un individu ou d'une communauté linguistique qui utilise concurremment trois langues différentes ou davantage.

occurrence - (sa) apparition d'une unité linguistique dans le discours; suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs de forme. En anglais apparition of a token. En chinois, 词条数/cí tiáo shù.

partie - (d'un corpus de texte) fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

partition - (stat, tm) (d'un corpus de texte) division d'un corpus en parties constituées par des fragments de textes consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.

phrase - (sa) fragment de texte compris entre deux séparateurs de phrase et recevant un sens indépendant.

poly-cooccurrence - (sa) attractions lexicales au-delà de la cooccurrence binaire.

relation - (tal) scénarios ou événements impliquant une entité nommée ou plus.

segment - (sr) toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur de séquence est un segment du texte.

segment répété - (sr) suite de formes dont la fréquence est supérieure ou égale à 2 dans le corpus.

segmentation - (sr) opération qui consiste à délimiter des unités minimales dans un texte.

série textuelle chronologique - (sa) dimension chronologique de corpus permettant de mettre en évidence des variations qui surviennent au cours du temps dans l'emploi du vocabulaire, de mettre en évidence des moments importants dans l'évolution de celui-ci.

séquence - (sa) suite d'occurrences du texte non séparées par un délimiteur de séquence.

seuil – (stat) quantité arbitrairement fixée au début d'une expérience visant à sélectionner parmi un grand nombre de résultats, ceux pour lesquels les valeurs d'un indice numérique dépassent ce seuil (de fréquence, en probabilité, etc.).

signal - désigne un message qui peut se présenter sous des formes diverses : quantitative ou qualitative, linguistique ou non linguistique, écrite ou orale. Il est émis délibérément et volontairement par une source qui peut être, selon le cas, une personne physique ou morale, ou encore un dispositif technique. Lorsqu'un signal est émis par une personne, sans que celle-ci en ait l'intention, ni même conscience, on utilise plutôt le vocable "signe". Par exemple, dans une réunion de travail, un participant peut sursauter en entendant une parole. Ce sursaut est un signe pour qui sait le voir et l'interpréter (Lesca et Lesca, 2011).

signal d'alerte précoce - désigne un signal annonciateur de changements dans l'environnement de l'entreprise et de nature à influer de façon significative sur le devenir de celle-ci. Il résulte généralement de l'interprétation d'un signal faible (Lesca et Lesca, 2011).

signal faible - désigne un "outil" d'aide à la décision. Il se présente *a priori* comme une "donnée" d'apparence anodine, mais dont l'interprétation que l'on en fait peut déclencher une alerte. Cette alerte indique que pourrait survenir un événement susceptible d'avoir des conséquences considérables (en termes d'opportunité ou de menace). Après interprétation le signal n'est plus qualifié de faible mais de signal d'alerte précoce (Lesca et Lesca, 2011).

source (vs) origine d'une transmission d'informations.

spécificité - (sp) indice de sur-emploi ou de sous-emploi dans la ou les partie(s) sélectionnée(s) par rapport à l'ensemble du corpus. Un exposant, seuil, rend compte du degré de significativité de l'écart constaté (un exposant égal à x, indique que la probabilité d'un écart de répartition supérieur ou égal à celui que l'on a constaté était, au départ de l'ordre de 10-x).

spécificité évolutive - (sp) calcul de spécificités (accroissements spécifiques dans Lebart & Salem, 1994) d'une partie par rapport à l'ensemble des périodes précédentes (en excluant momentanément du corpus les périodes postérieures).

spécificité négative - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique négative de la partie j si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée est inférieure aux seuil fixé au départ.

spécificité positive - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive de la partie j si sa sous-fréquence est « anormalement élevée » dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous- fréquence constatée est inférieure au seuil fixé au départ.

taille - (sa) (d'un corpus) sa longueur mesurée en occurrences (de formes simples).

textométrie - (tm) ensemble des méthodes et des outils informatiques permettant d'opérer des analyses statistiques et de faciliter des explorations qualitatives d'un corpus.

topographie textuelle - (tm) représentation graphique des phénomènes langagiers mis en évidence par l'étude statistique afin d'apprécier leurs positions dans le texte.

type généralisé (TGen) - (tm) sous-ensemble d'occurrences d'un texte défini à l'aide des expressions régulières.

veille stratégique - (vs) collecte, analyse et interprétation des informations nécessaires à une entreprise pour élaborer un plan d'action en réponse à une situation économique ou en vue d'améliorer sa compétitivité.

ventilation - (sa) (des occurrences d'une unité dans les parties du corpus) La suite des n nombres (n=nombre de parties du corpus) constituée par la succession des sous- fréquences de cette unité dans chacune des parties, prises dans l'ordre des parties.

vocabulaire - (sa) ensemble de formes attestées dans un corpus de textes.

Sigles et Acronymes

ACRIE	Accompagnement, Conseil, Recherche en Intelligence économique
ADEME	Agence de l'environnement et de la maîtrise de l'énergie
AFC	Analyse Factorielle des Correspondances
AFDIE	Association Française pour le Développement de l'Intelligence Economique
AFNOR	Agence Française de NORmalisation
AIE/AIEA	Agence Internationale de l'Energie (en anglais IEA International Energy Agency) ou Agence Internationale de l'Energie Atomique
ANDRA	Agence Nationale pour la gestion des Déchets Radioactifs
ASN	Autorité de Sûreté Nucléaire
BNEF	Bloomberg New Energy Finance
BRICS	Brésil, Russie, Inde, Chine, Afrique du Sud
CAEA	China Atomic EnergyAuthority
CBEEX	China Beijing Environment Exchange
CCI	Chambre de Commerce et d'Industrie
CCS/CSC	Carbon Capture Stockage an anglais, Capture et Stockage du Carbone en français
CCNUCC	Convention-Cadre des Nations Unies sur les Changements Climatiques
CCUS	Carbon Capture, Use and Storage
CDM/MDP	Clean Development Mechanism en anglais, Mécanisme de Développement Propre en français
CEA	Commissariat à l'Energie Atomique et aux énergies alternatives
CERDI	Centre d'Etudes et de Recherches sur le Développement International
CGNP/CGN	China Guangdong Nuclear Power Corporation (abrégé CGN ou CGNPC)
CIGéO	Centre Industriel de stockage GéOlogique
CIGREF	Club Informatique des Grandes Entreprises Françaises
CNNC	China National Nuclear Corporation
CNOOC	China National Offshore Oil Corporation
CNRTL	Centre National de Ressources Textuelles et Lexicales
COP	Conférence des parties (CP, en anglais Conference of the parties)
CPIC	China Power Investment Corporation
CSA	Comité de la Sécurité Alimentaire mondiale
CSC	Capture et Stockage du Carbone

Sigles et Acronymes

D2IE	Délégation Interministérielle à l'Intelligence Economique
DIGEC	Direction du Gaz de l'Electricité et du Charbon
DOE	Department of Energy
EPR	European Pressurised Reactor
EQ/eq	Équivalent-CO2
ESLSCA	Ecole Supérieure Libre des Sciences Commerciales Appliquées
GEIDE	Gestion Electronique de l'Information et des Documents Existants
GES	Gaz à Effet de Serre
GFII	Groupement Français de l'Industrie de l'Information
GNL	Gaz naturel liquéfié
GWe	GigaWatt electrical
HCTISN	Haut Comité pour la Transparence et l'Information sur la Sécurité Nucléaire
ICTCLAS	Institute of Computing Technology Chinese Lexical Analysis System
IE	Intelligence Economique
IEA	International Energy Agency
IEEE	Institute of Electrical and Electronics Engineers
IFOP/Ifop	Institut français d'opinion publique
IGCC	Integrated Gasification Combined Cycle
IGCC	Institute on Global Conflict and Cooperation **(**University of California San Diego**)**
IHEST	Institut des Hautes Etudes pour la Science et la Technologie
INES	International Nuclear Event Scale
INRIA	Institut National de Recherche en Informatique et en Automatique
IRENA	Agence Internationale des énergies renouvelables (International Renewable Energy Agency)
IRSN	Institut de Radioprotection et de Sûreté Nucléaire
ITER	International Thermonuclear Experimental Reactor
MDP/CDM	Mécanisme de Développement Propre
MEP	Ministry of Environmental Protection
MW	Méga Watt (unité de puissance qui traduit le potentiel de production de l'installation)
NNSA	National Nuclear Safety
NPTC	Nuclear Power Technology Corporation (Chine)
NRC	Nuclear Regulatory Commission
NSC	Nuclear Safety Centre
NTIC	Nouvelles Techniques de l'Information et de la Communication

OCDE	Organisation de coopération et de développement économiques
OIT	Organisation Internationale du Travail
OMC	Organisation Mondiale du Commerce
ONG	Organisation Non Gouvernementale
ONU	Organisation des Nations unies
OPECST	Office Parlementaire d'Evaluation des Choix Scientifiques et Technologiques
PBMR	Pebble Bed Modular Reactor
RCG	Réacteur à Caloporteur Gaz
REN21	Renewable Energy policy Network for the 21st Century
REP	Réacteur à eau sous pression
RNR	Réacteur à neutrons rapides
RPC	République Populaire de Chine
SCIM	Smart Common Input Method
SEPA	State Environment Protection Agency
SFEN	Société Française d'Energie Nucléaire
SNPTC	State Nuclear Power technology Company (Chine)
SOV	Sujet + Objet + Verbe (langue SOV)
SVO	Sujet + Verbe + Objet (langue SVO)
TAL	Traitement Automatique des Langues
THTR	Thorium High Temperature Reactor (Thorium réacteur à haute température)
TKIP	Temporal Key Integrity Protocol (Protocole temporel par clé intégrale)
TLFi	Trésor de la langue française informatisé
TVO	Teollisuuden Voima Oyj (producteur d'électricité finlandais)
VSST	Veille Stratégique Scientifique et Technique
WEP	Wired Equivalent Privacy (Clé privée par équivalence filaire)
WISE	World Information Service Energy
WPA	WI-FI Protected Access
WWF	World Wildlife Fund

Bibliographie

ADAM, Jean-Michel (1997). Unités rédactionnelles et genres discursifs : cadre général pour une approche de la presse écrite. *Pratiques*, juin 1997, n°94, 18 p.

AGUILAR, François-Joseph (1967). *Scanning the business environment*. Mac Millan, New York, 239 p.

ALEXEEVA, Olga, ROCHE, Yann (2014). La Chine en transition énergétique : un virage vers les énergies renouvelables ? *VertigO*, revue électronique en science de l'environnement, décembre 2014, vol. 14, n°3, 31 p, http://vertigo.revues.org/15540 (consulté le 13/04/2015).

ANSOFF, Harry Igor (1975). Managing strategic surprise by response to weak signals. *California Management Review*, vol. 18, n°2, pp. 21-33.

ANSOFF, Harry Igor (1989). *Stratégie du développement de l'entreprise*. Paris, éditions d'organisation, 288 p.

BAUMARD, Philippe (1991). *Stratégie et surveillance des environnements cooccurrentiels*. Paris, Masson, 192 p.

BAUQUIS, Pierre-René (2001). Un point de vue sur les besoins et les approvisionnements en énergie à l'horizon 2050. *Liaison énergie francophonie*. Québec, Institut de l'énergie des pays ayant en commun l'usage du français, n°52, pp. 5-15.

BAUQUIS, Pierre-René (2006). Quels axes pour une politique énergétique française ? *Revue de l'énergie*. Paris, éditions techniques et économiques, n° 571, pp. 149-158.

BEDUNEAU-WANG, Laurent, SHAN, Meng, GALHARRET, Sophie, VENDRYES, Thomas (2010). L'Union européenne face à la Chine : quelle politique environnementale? *Terra Nova*. Note publiée le 2/04/2010, 13 p, http://www.tnova.fr/note/lunion-europ-enne-face-la-Chine-quelle-politique-environnementale (consulté le 13/04/2015).

BELL, Allan (1991). *The language of News Media*. Cambridge, MA, USA, Wiley-Blackwell, coll. Language in Society, 277 p.

BENZECRI, Jean-Paul (1968). La place de l'a priori. *Encyclopedia Universalis*, tome 17, pp. 11-23.

BENZECRI, Jean-Paul et coll. (1973). *L'analyse des données*. Paris, Dunod, tome 1 : la taxinomie, tome 2 : l'analyse des correspondances, 615 p.

BENZECRI, Jean-Paul (1977). Analyse discriminante et analyse factorielle. *Les cahiers de l'analyse des données*, Paris, Dunod, tome 2, n°4, pp. 369-406.

BENZECRI, Jean-Paul et coll. (1981). *Pratique de l'analyse des données : linguistique et lexicologie*. Paris, Dunod, 585 p.

BERTEL, Evelyne, NAUDET, Gilbert (2004). *L'économie de l'énergie nucléaire*. Les Ullis, EDP Science-France, coll. génie atomique, 2004, 445 p.

BERTIN, Annie (1997). L'expression de la cause en ancien français (Vol. 219). Librairie Droz.

BERTIN, Annie (2001). Maintenant: un cas de grammaticalisation?. *Langue française*, 42-64.

BERTIN, Annie (2002). L'émergence du connecteur en effet en moyen français. *Linx*. Revue des linguistes de l'université Paris X Nanterre, (46), 37-50.

Bibliographie

BERTIN, Annie (2003). Les connecteurs de cause dans l'histoire du français: Contradictions du changement linguistique. *Verbum*, (3), 263-276.

BESSON, Bernard, POSSIN, Jean-Claude (2001). *Du Renseignement à l'Intelligence Economique*. Paris, Dunod, coll. Stratégies et management, 2ème édition, 240 p., 1ère édition 1996.

BESSON, Bernard, POSSIN, Jean-Claude (2002). *L'Audit d'Intelligence Economique : mettre en place et optimiser un dispositif coordonné d'intelligence collective*. Paris, Dunod, coll. Fonctions de l'entreprise, 2ème édition, 29/08/2002, 205 p.

BOIZARD, Odile (2005). Veille ou intelligence économique, faut-il choisir ? Retour d'expérience. *Information Sciences for Decision Making*, n° 21, 13 p., http://isdm.univtln.fr/PDF/isdm21/boizard.pdf (consulté 20/12/2014).

BOLLIER, David (2010). The promise and peril of big data. *Communications and Society program*. Washington, DC, the Aspen Insitute, 61 p.

BONNAFOUS, Simone, TOURNIER, Maurice (1995). Analyse de discours, lexicométrie, communication et politique. In : *Langages, 29e année*, n°117, Paris, Larousse, pp. 67-81.

BONNAURE, Pierre (2011). Le grand retour du nucléaire. *Futuribles*. Paris, janvier 2011, n°370, pp. 31-44.

BONNEVAL, Laure, LACROIX-LANOË, Cécile (2011). L'opinion publique européenne et nucléaire après Fukushima. *Fondation Jean-Jaurès*. Paris, 26 septembre 2011, note n°101, 17 p. http://www.ifop.com/media/pressdocument/355-1-document_file.pdf (consulté le 17/12/2014).

BOQUET, Yves (2009). La démographie chinoise en mutation. *Espace, populations, sociétés*. Lille, Université Lille 1, dossier pédagogique, 2009/3 : les populations de la Chine, pp. 551-568, http://eps.revues.org/3869 (consulté le 28/03/2015).

BOUR, Ludovic et al (2012). *Nouveaux usages de la veille : 5 pratiques en émergence*. Livre blanc nouveaux usages de la veille. Paris, GFII, Groupe de travail Intelligence Economique et Economie de la connaissance, juin 2012, 56 p., http://www.gfifr/uploads/docs/Livre%20blanc%20Nouveaux%20usages%20de%20la%20veille.pdf (consulté le 14/02/2015).

BOURIGAULT, Didier et SLODZIAN, Monique (1999). Pour une terminologie textuelle. *Terminologies nouvelles*, 1999, vol. 19, no 1999, p. 29-32.

BOURRY, Chantal (2012). *La vérité scientifique sur le nucléaire*. Paris, éditions rue de l'Echiquier, 204 p.

BOWKER, Lynne, PEARSON, Jennifer (2002). *Working with Specialized Language : a practical guide to using corpora*. New York: Routledge, october 2002, 256 p.

CARAYON, Bernard (2003). *Intelligence économique, compétitivité et cohésion sociale*. Documentation française, coll. des rapports officiels, juillet 2013, 173 p.

CARAYON, Bernard (2004). Une nouvelle politique publique pour répondre à la guerre économique. *Constructif*. Paris, éditeur Fédération Française du Bâtiment, mai 2004, n°8, http://www.constructif.fr/bibliotheque/2004-5/une-nouvelle-politique-publique-pour-repondre-a-la-guerre-economique.html?item_id=2548 (consulté le 11/01/2015).

CARFANTAN, Jean-Yves (2014). *Le défi chinois : les nouvelles stratégies d'un géant*. Seuil, 281 p.

CARON-FASAN, Marie-Laurence (2001). Une méthode de gestion de l'attention aux signaux faibles. *Revue Systèmes d'information et Management*. Paris, Eska, vol. 6, n°4, pp.73-89.

CASTAGNOLI, Sara et al (2011). Designing a Learner Translator Corpus for Training Purposes. *In N.*

Kübler, Corpora, Language, Teaching, and Resources : From Theory to Practice. Berne, éditeur Peter Lang, pp. 221-248.

CHAMBON, Jean-Louis et al (2010). *La Chinamérique : un couple contre-nature ?* Le cercle Turgot, Paris, Eyrolles, coll. éditions d'Organisation, 281 p.

CHANG, Shiyan, ZHAO, Lili, TIMILSINA, Govinda, ZHANG, Xiliang (2012). Development of biofuels in China: technologies, economics and policies. Policy Research Working Paper 6243. Washington, DC, World Bank, http://elibrary.worldbank.org/content/workingpaper/10.1596/1813-9450-6243 (consulté le 1/03/2015).

CHAUDET, Didier (2013). La politique chinoise en Asie Centrale : quand Beijing regarde à l'Ouest. *Le Huffingtonpost*, publié le 26/09/2013, http://www.huffingtonpost.fr/didier-chaudet/politique-internationale-chinoise_b_3974806.html (consulté le 6/03/2015).

CHAVARDES, Daniel (2004). Situation et problématique de l'énergie en Chine, perspectives de développement électronucléaire. Pékin, 22 septembre 2004, http://www.uarga.org/downloads/Documentation/Perspect_energ_chin.htm (consulté le 10/04/2015).

CHAVARDES, Daniel (2009). Le développement de l'énergie nucléaire en Europe, en Amérique et en Asie. Evolution de l'opinion publique américaine vis à vis de l'énergie nucléaire. Octobre 2009, http://www.janus.co.jp/Portals/0/images-en/expert_columns/pdf/J.4.2.3.pdf (consulté le 08/04/2015).

CHAVARDES, Daniel (2010). Le développement de l'énergie nucléaire en Europe, en Amérique et en Asie. Comparaison de l'énergie nucléaire en Chine et en Inde. Mai 2010, http://www.janus.co.jp/Portals/0/images-en/expert_columns/pdf/J.4.2.6.pdf (consulté le 08/04/2015).

CHEN, Keh-Jiann, LIU, Shing-Huan (1992). Word identification for Mandarin Chinese Sentences. Institute of Information Science. Academia Sinica. *Actes de Coling-92*, Nantes 23-28 août 1992, pp. 101-107.

CHEN, Yuyu, EBENSTEIN, Avraham, GREENSTONE, Michaël, LI, Hongbin (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *PNAS, Proceedings of the national Academy of Sciences of the United States of America,* edited by William C. Clark, Harvard University, 6 august 2013, Cambridge, MA, vol.110, n°32, pp.12936-12947, http://www.pnas.org/content/110/32/12936 (consulté le 7/03/2015).

CHIAO, Yun-Chuang (2004). *Extraction lexicale bilingue à partir de textes médicaux comparables : application à la recherche d'information translangue.* Thèse pour le doctorat en Informatique Médicale, Université Pierre-et-Marie-Curie, Paris 6, 30 juin 2004, directeur Pierre Zweigenbaum, 196 p.

CICUREL, Francine (1994). Les scénarios d'information dans la presse quotidienne, le Français dans le monde. *Numéro spécial Recherches et applications, « médias, faits et effets ».* Paris, septembre 1994, pp. 91-102.

CONSEIL de l'EUROPE (2005). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Editions Didier, coll. FLE, 187 p.

CORI, Marcel, LEON, Jacqueline (2002). La constitution du TAL. *ATALA*. Paris, vol. 43, n°3, pp. 21-55, https://halshs.archives-ouvertes.fr/file/index/docid/158854/filename/CoriLeon.PDF (consulté le 12/01/2015).

CORI, Marcel (2004). Traitement automatique des langues et formalisation en linguistique. Université Paris X Nanterre, 29 octobre 2004, http://www.tal.univ-paris3.fr/plurital/cours3-2004/slides-cours3-2004-1.pdf (consulté le 28/03/2015).

Bibliographie

CORNISH, Francis (2005). Compléments nuls vs. pronoms objets manifestes en anglais en tant qu'anaphoriques : syntaxe, sémantique ou pragmatique ? Workshop « *Reference : how much linguistics semantics and how much pragmatics ?* », 21ème Congrès Scandinave de Linguistique à l'Université Scientifique et Technique de Trondheim, Norvège, organisé du 1 au 2 juin 2005, 13 p.

COUTENCEAU, Christian, BARBARA, François, EVERETT, William et al (2009). *Guide pratique de l'Intelligence Economique*. Paris, Eyrolles, 156 p.

COUTENCEAU, Christian, BARBARA, François, CHAPUIS-THUAULT, Véronique et al (2014). *L'intelligence économique au service de l'innovation*. Paris, Eyrolles, 408 p.

DAVID, Bruno (2004). Guerre en Irak : les journalistes à découvert. *Actes du 10ème colloque franco-roumain en Sciences de l'Information et de la Communication : supports, dispositifs et discours médiatiques à l'heure de l'internationalisation,* Bucarest, juin 2003, 2004, 8 p, https://halshs.archives-ouvertes.fr/sic_00000896/document (consulté le 02/05/2015).

DEJEAN, Hervé, GAUSSIER, Eric (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, Paris, Université de la Sorbonne Nouvelle, Alignement lexical dans les corpus multilingues, numéro spécial, 2002, 22 p.

DELANOE, Alexandre (2010). Statistique textuelle et séries chronologiques sur un corpus de presse écrite. Le cas de la mise en application du principe de précaution. *JADT 2010, 10th International Conference on Statistical Analysis of Textual Data*, pp. 561-572.

DELBOSC, Anaïs (2011). 12ème plan quinquennal chinois : marché(s) du carbone en vue. *Point Climat*, *CDC Climat*. Paris, juin 2011, n°5, 6 p.

DELENGAIGNE, Xavier, MONGIN, Pierre, DESCHAMPS, Christophe (2011). *Organisez vos données personnelles : L'essentiel du Personal Knowledge Management*. Paris, Eyrolles, éditions d'Organisation, coll. Livres outils, Efficacité professionnelle, 248 p.

DESSUS, Benjamin (2011). La croissance verte, une illusion ? Energie et risque climatique : repenser nos modèles de développement. *Futuribles*. Paris, avril 2011, n°373, pp. 29-45.

DESSUS, Benjamin, LAPONCHE, Bernard (2011). *En finir avec le nucléaire : pourquoi et comment ?* Paris, éditions du seuil, octobre 2011, 176 p.

DEVEAUX, Pascal, COLLIGNON, Albert (2013). *Le livre blanc sur la sûreté nucléaire des installations civiles de la Manche « Post Fukushima »*. Inter-cli, commissions locales d'information de la Manche. Décembre 2013, 183 p., http://climanche.fr/newsletter/doc-201312/livre-blanc-sureté-installations-nucléaires-civiles-manche-post-fukushima.pdf (consulté le 18/12/2014).

DOLLFUS, Olivier (1999). Mondialisation et gaz à effet de serre. *Espace géographique*. Paris, tome 28, n°1, pp. 29-35. http://www.persee.fr/doc/spgeo_0046-2497_1999_num_28_1_1216 (consulté le 18/12/2014).

DOMERGUE, Lucas (2005). *La Chine, puissance nucléaire : stabilisation régionale ou prolifération?* L'Harmattan, 2005, 224 p.

DONG, Fengxia (2007). Food security and biofuels development: the case of China. *Briefing Paper 07-BP 52*. Center for Agricultural and Rural Development, Iowa State University, 18 p., http://absfafrica.org/downloads/Food%20Security%20and%20Biofuels%20Development%20The%20Case%20of%20China.pdf (consulté le 1/03/2015).

ETIENNE, Ludovic et al (2003). *Intelligence économique et stratégique : les systèmes d'information au cœur de la démarche.* Paris, Cigref, Rapport, mars 2003, 131 p.

Bibliographie

FAZILOV, Fakhmiddin, CHEN, Xiangming (2013). China and Central Asia : a significant new energy nexus. *The European Financial review*. China & the World Series, April-may 2013, pp. 37-43, http://www.europeanfinancialreview.com/?p=926#!prettyPhoto (consulté le 5/03/2015).

FELDMAN, Ronen, DAGAN, Ido (1995). Knowledge discovery from textual databases. *In Proceedings of the International Conference on Knowledge Discovery from DataBases*, KDD-95 Proceedings, pp. 112–117.

FILLIAS, Edouard (2005). Retour sur investissement d'un logiciel de veille stratégique : évaluer et calculer la valeur ajoutée de la mise en place d'une solution logicielle de veille stratégique. Digimind consulting, White paper, 42 p.

FILLIAS, Edouard (2007). Le WEB 2.0 pour la veille et la recherche d'information : explorer les ressources du Web social. Digimind consulting, White paper, 61 p.

FIRTH, John Rupert (1957). A Synopsis of Linguistic Theory 1930-1955. *In studies Linguistic Analysis, Oxford*. Philological Society, 18 p.

FLAMENT, Claude, MILLAND, Laurent (2005). Un effet Guttman en ACP. *Math. & Sci. hum. ~ Mathematics and Social Sciences*, 43ème année, n° 171, 2005, pp. 25-49.

FLEURY, Serge (2014). *Le Métier Textométrique : Le Trameur, Manuel d'utilisation*. CLA2T/SYLED, Université Sorbonne Nouvelle Paris 3, Centre de Textométrie, version 12.43, http://tal.univ-paris3.fr/trameur/ (consulté le 10/06/2015).

FRION, Pascal (2002). Entre veille et intelligence économique, il faut choisir ! *Technologies Internationales*. Strasbourg, Association pour la diffusion de l'information technologique, mai 2002, n°84, pp 37-40.

FRION, Pascal (2004). *Accompagnement à la Recherche d'Information Economique : l'Intelligence économique expliquée pour une PME-PMI*. Arn Editions, coll. Intelligence économique pas à pas, 255 p.

FRION, Pascal (2004). *Accompagnement au Traitement de l'Information Essentielle : La veille et la gestion de l'information pour une PME-PMI*. Arn Editions, coll. Intelligence économique pas à pas, 252 p.

FRION, Pascal (2012). *Généalogie de la faible percée du discours sur l'intelligence économique dans les TPE françaises : errements épistémologiques et propositions opérationnelles*. Thèse : Sciences de l'Information et de la Communication, Université de Poitiers Centre de Recherche en Gestion (Cerege), 7 décembre 2012, 464 p.

GALE, William A, CHURCH, Kenneth W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, vol. 19, number 1, pp. 75-102, http://www.aclweb.org/anthology/J93-1004 (consulté le 23/04/2015).

GILAD, Benjamin, GILAD, Tamar (1998). *The Business Intelligence System: A New Tool for Competitive Advantage*. New York, American Management Association, 242 p.

GOEURIOT, Lorraine (2009). *Découverte et caractérisation des corpus comparables spécialisés*. Thèse : Informatique, Université de Nantes, 30 janvier 2009, 150 p.

GOEURIOT, Lorraine, MORIN Emmanuel, DAILLE Béatrice (2009). Reconnaissance de critères de comparabilité dans un corpus multilingue spécialisé. Actes de la Sixième Conférence Francophone en Recherche d'Information et Applications (CORIA 2009), 5 - 7 mai 2009, pp. 33-47.

GOUYSSE, Vincent (2010). *Le réveil du dragon*. Lulu.com, 516 p.

GRESILLON, Gabriel (2012). La Chine accélère son programme de développement des gaz de schiste. *Les*

Echos, 26/10/2012, http://www.lesechos.fr/26/10/2012/LesEchos/21300-100-ECH_la-Chine-accelere-son-programme-de-developpement-des-gaz-de-schiste.htm (consulté le 4/03/2015).

GRISHMAN, Ralph, SUNDHEIM, Beth (1996). Message Understanding Conference- 6 : A Brief History. *In proceedings of the 16th International Conference on Computational Linguistics* (COLING), I. Kopenhagen, pp. 466–471.

GRISHMAN, Ralph (2003). Information Extraction. *In Mitkov*, R., The Oxford Handbook of Computational Linguistics, Oxford: Oxford University Press, 2003, pp. 545-559.

GUERMOND, Yves (2007). *La Chine*. Paris, Belin, coll. Memento Géographie, 175 p.

GUERMOND, Yves, MA, Kun (2011). La production d'énergie en Chine. *M@ppemonde*. Revue trimestrielle sur l'image géographique et les formes du territoire, Université d'Avignon et des pays du Vaucluse, 2011/1, n°101, 12 p, http://mappemonde.mgm.fr/num29/lieux/lieux11101.html (consulté le 12/04/2015).

GUIDERE, Mathieu (2008). La veille multilingue : défense et illustration de la traduction stratégique. Premier colloque international sur la veille stratégique multilingue. Université de Genève, ETI, Suisse, 28-29 mai 2008, 29 p.

GUIDERE, Mathieu (2008). *Traduction et Veille stratégique multilingue*. Paris, éditions Le Manuscrit, 275 p.

HABERT, Benoît, NAZARENKO, Adeline, SALEM, André (1997). *Les linguistiques de corpus*. Paris, Armand Colin/Masson, 254 p.

HABERT, Benoît, ZWEIGENBAUM, Pierre (2002). Problèmes épistémologiques : Régler les règles. *TAL*. Paris, Association pour le traitement automatique des langues, vol. 43, n°3, pp. 83-105.

HARARI, François, CHAUVIN, Carole (2012). Le réacteur AP1000. *Annales des Mines-Réalités industrielles*. Paris, Eska, 2012 /3, pp. 142-145.

HARBULOT, Christian, BAUMARD, Philippe (1997). Perspective historique de l'intelligence économique. *Revue Intelligence économique*, n°1, 17 p, http://mouradpreure.o.m.f.unblog.fr/files/2010/04/16perspectivehistorique.pdf (consulté le 20/12/2014).

HARBULOT, Christian (2000). Contre-Influence. *Doctrines*, Infoguerre.fr, http://www.infoguerre.fr/doctrines/contre-influence-par-christian-harbulot-112 (consulté le 13/02/2015).

HARBULOT, Christian (2001). De la guerre économique à la guerre de l'information. *Géostratégiques*. Paris, juin 2001, n°5, 7 p, http://www.strategicsinternational.com/f5harbulot.htm (consulté le 13/02/2015).

HARBULOT, Christian (2004). L'émergence de l'intelligence économique en France. *Constructif*. Paris, éditeur Fédération Française du Bâtiment, mai 2004, n°8, http://www.constructif.fr/bibliotheque/2004-5/l-emergence-de-l-intelligence-economique-en-france.html?item_id=2547 (consulté le 11/01/2015).

HARRIS, Brian (1988). Bi-text, a new concept in translation theory, http://mt-archive.info/LangMonthly-54-1988-Harris.pdf (consulté le 23/04/2015).

HASKI, Pierre (2013). Chine : une manif antinucléaire fait céder le gouvernement chinois. *L'OBS avec Rue89*, publié le 13/07/2013, http://rue89.nouvelobs.com/2013/07/13/Chine-manif-antinucléaire-provoque-lannulation-dun-projet-industriel-geant-244207 (consulté le 15/04/2015).

Bibliographie

HENRIET, Fanny, MAGGIAR, Nicolas (2012). Croissance verte et croissance économique. *Bulletin de la Banque de France*. Paris, 4ème trimestre 2012, n°190, pp. 143-152.

HERMEL, Laurent (2010). *Maîtriser et pratiquer... Veille stratégique et intelligence économique.* AFNOR éditions, 120 p.

HERZOG, Siegfried (1915). *Die Zukunft des deuschen Ausfuhrhandels Wegleitungen und praktische Winke zur sicherung und Förderung deutscher Ausfuhrtätigkeit auf technischem Gebiet nach Beendifung des Krieges, Verlag von Ferdinand Enke*. In Stuttgart, édité en France sous le titre « le plan de guerre commerciale de l'Allemagne », Payot, 1919, 249 p.

HILL, Joshua S (2014). IRENA Says China Can Nearly Quadruple Renewable Energy By 2030. *Clean Technica*. http://cleantechnica.com/2014/11/25/irena-says-china-can-nearly-quadruple-renewable-energy-2030/ (consulté le 28/02/2015).

JAKOBIAK, François (2004). Un atout supplémentaire pour les grandes entreprises. *Constructif*. Paris, éditeur Fédération Française du Bâtiment, mai 2004, n°8, http://www.constructif.fr/bibliotheque/2004-5/un-atout-supplementaire-pour-les-grandes-entreprises.html?item_id=2563 (consulté le 11/01/2015).

JAKOBIAK, François (2004). *L'Intelligence économique : la comprendre, l'utiliser, l'implanter*. Paris, éditions d'Organisation, 335 p.

JAKOBSON Roman, LEVI-STRAUSS Claude (1962). « Les Chats » de Charles Baudelaire. In: *L'Homme*, 1962, tome 2 n°1. pp. 5-21; http://www.persee.fr/doc/hom_0439-4216_1962_num_2_1_366446

JANCOVICI, Jean-Marc (2003). Le nucléaire civil, péché majeur du XXème siècle ? *Le Débat*. Gallimard, 1/2003, n° 123, pp. 175-192, http://www.manicore.com/documentation/articles/idee_nucléaire.html (consulté le 11/04/2015).

JAOUEN, Claude, BÉROUX, Pierre (2012). Généralités sur les réacteurs nucléaires. *Annales des Mines-Réalités industrielles*. Paris, Eska, 2012 /3, pp. 113-141.

JENSTER, Per V., SOLBERG SOILEN, Klaus (2009). *Market Intelligence*. Copenhagen Business School Press, 240 p.

JUILLET, Alain (2004). L'état relance la dynamique. *Constructif*. Paris, éditeur Fédération Française du Bâtiment, mai 2004, n°8, http://www.constructif.fr/bibliotheque/2004-5/l-etat-relance-la-dynamique.html?item_id=2551 (consulté le 11/01/2015).

KAN, Naoto (2013). Mon expérience de premier ministre durant l'accident nucléaire de Fukushima. Symposium de New York, les conséquences médicales et écologiques de l'accident nucléaire de Fukushima, 11-12 mars 2013, http://www.fukushima-blog.com/naoto-kan-mon-expérience-de-premier-ministre-durant-l-accident-nucléaire-de-fukushima (consulté le 17/02/2015).

KEPPLER, Jan H, MERITET, Sophie (2004). Les perspectives énergétiques de la Chine. *Revue de l'énergie*. Paris, Editions techniques et économiques, juin 2004, n°557, pp. 316-320.

KOEHN, Philipp (2004). *Europarl : A Parallel Corpus for Statistical Machine Translation.* http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.126.7716, (consulté le 23/04/2015).

KOENIG, Gérard (1990). *Management stratégique : vision, manœuvres et tactiques*. Paris, éditions Nathan, 399 p.

KOENIG, Gérard (1996). *Management stratégique : paradoxes, interactions et apprentissage*. Paris, éditions Nathan, 544 p.

KREFT, Heinrich (2006). La diplomatie chinoise de l'énergie. Institut français des relations internationales

(IFRI), *Politique étrangère*, 2006/2 (Eté), pp. 349-360.

KRIEG-PLANQUE, Alice (2009). *La notion de « formule » en analyse du discours. Cadre théorique et méthodologique*. Paris, Presses Universitaires de Franche-Comté, 146 p.

KÜBLER, Natalie (2011). Traduction pragmatique, linguistique de corpus, traducteur : un ménage à trois explosif ? Tralogy, (en ligne), Tralogy II, 3 et 4 mars 2011, Session 3, Machine and Human Translation : Finding the Fit? / TA et Biotraduction, http://lodel.irevues.inist.fr/tralogy/index.php?id=288&format=print (consulté le 20/04/2015).

LACROIX-LANOË, Cécile (2013). Stop ou encore ? L'opinion publique française face au nucléaire. *Délits d'opinion*. Publié le 21/10/2013, http://www.delitsdopinion.com/theme/societe/stop-ou-encore-lopinion-publique-francaise-face-au-nucléaire-12989/ (consulté le 13/03/2015).

LAFARGUE, François (2013). Les enjeux géopolitiques de la transition énergétiques. *Magazine Green Innovation*. juillet, août, septembre 2013, n°1, pp. 22-28, http://www.green-magazine.fr/?page_id=8902 (consulté le 10/04/2015).

LAFON, Pierre (1980). Sur la variabilité de la fréquence des formes dans un corpus. In : *Mots*, n°1, octobre 1980, pp. 127-165.

LAFON, Pierre (1981). Analyse lexicométrique et recherche des cooccurrences. In : *Mots*, octobre 1981, n°3, pp. 95-148.

LAMALLE, Cédric, SALEM, André (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. Morin, A. Sébillot (eds.), *Actes des 6èmes Journées d'analyse des données textuelles*, Saint-Malo.

LANDAIS, Patrick (2012). La gestion des Déchets radioactifs. SFEN, http://www.sfen.org/La-gestion-des-dechets-radioactifs,1867 (consulté le 17/02/2015).

LA POLLA, Randy-John (1990). *Grammatical Relations in Chinese* : Synchronic and Diachronic Considerations. Doctorat de l'université Berkeley, Californie.

LAPONCHE, Bernard (2015). ATMEA et ses concurrents. *Les Cahiers de Global Chance*, n°37, juin 2015, pp.73-79, http://www.global-chance.org/IMG/pdf/gc37.pdf (consulté le 6/09/2015).

LARAMEE de TANNENBERG, Valéry (2012). La Chine découvre le risque nucléaire. *Journal de l'environnement*, publié le 22/10/2012), http://www.journaldelenvironnement.net/article/la-Chine-decouvre-le-risque-nucléaire,31271 (consulté le 13/11/2014) et http://english.mep.gov.cn/News_service/news_release/201206/t20120621_231995.htm# (consulté le 13/11/2014).

LARIVET, Sophie (2001). Intelligence économique : acceptation française et multidimensionnalité. *Xème conférence de l'association internationale du management stratégique*, Québec, 13, 14, 15, juin 2001.

LAROCHE, Hervé, NIOCHE, Jean-Pierre (1994). L'approche cognitive de la stratégie d'entreprise. *Revue Française de Gestion*, n°99, juin, juillet, août, pp. 64-78.

LAURELUT, Jacques, ARLABOSSE, François, AZOULAY, Edouard et al (1998). Veille stratégique : organiser la veille sur les nouvelles techniques de l'information. CIGREF, 91 p, http://www.cigref.fr/cigref_publications/RapportsContainer/Parus1998/Veille_strategique_1998_web.pdf (consulté le 15/01/2015).

LAURENT, Eloi (2011). Quelle crédibilité économique et écologique pour la gauche en 2012 ? *Multitudes.* Paris, Assoc. Multitudes, 2011/3, n°46, pp. 110-121.

Bibliographie

LAURIER, Dominique, ROMMENS, Catherine, DROMBRY-RINGEARD, Caroline, MERLE-SZEREMETA, Aurélie, DEGRANGE, Jean-Pierre (2000). Evaluation du risque de leucémie radio-induite à proximité d'installations nucléaires : l'étude radio-écologique Nord-Cotentin. *Revue d'épidémiologie et de santé publique*. Congrès colloque « Epidémiologie, environnement et santé » Saint-Malo, Issy-les-Moulineaux, Elsevier Masson, vol.48, SUP2, pp. 2524-2536.

LAURIER, Dominique (2007). Risque de leucémie infantile autour des installations nucléaires. *Rapport scientifique et technique IRSN*, pp. 123-128, http://www.irsn.fr/FR/Larecherche/publications-documentation/aktis-lettre-dossiers-thematiques/RST/RST-2007/Documents/RST2007_Chap3_3-2_Risque-leucemie.pdf (consulté le 17/12/2014).

LEBART, Ludovic, SALEM, André (1994). *Statistique textuelle*. Paris, Dunod, 370 p. http://lexicometrica.univ-paris3.fr/livre/st94/st94-tdm.html (consulté le 12/01/2015).

LEBLANC, Jean-Marc, MARTINEZ, William (2006). L'analyse contrastive des réseaux de cooccurrence. Le monde dans les discours des présidents de la Cinquième République. *JADT 2006 : 8es Journées internationales d'Analyse statistique des Données Textuelles,* http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-054.pdf (consulté le 01/04/2014)

LE CORRE, Philippe (2006). La Chine. *Études.* 10/2006, Tome 405, pp. 307-318, www.cairn.info/revue-etudes-2006-10-page-307.htm (consulté le 01/04/2015).

LE LEUCH, Honoré (2010). Evolutions récentes de l'énergie aux Etats-Unis. *Géostratégiques* n°29, 4ème trimestre 2010, pp.225-239, http://www.strategicsinternational.com/29_16.pdf (consulté le 15/10/2014).

LENGLET, François (2010). *La guerre des empires. Chine contre Etats-Unis*. Paris, Fayard, 244 p.

LEPLÂTRE, Simon (2015). La Chine ferme les centrales à charbon situées à Pékin. *La Croix*, publié le 31/03/2015, http://www.la-croix.com/Actualite/Economie-Entreprises/Economie/La-Chine-ferme-les-centrales-a-charbon-situees-a-Pekin-2015-03-31-1297408 (consulté le 13/04/2015).

LESCA, Humbert (1986-1990). *Système d'information pour le management stratégique : l'entreprise intelligente*. Paris, éditions Mc Graw Hill, 146 p.

LESCA, Humbert (1989). *Information et adaptation de l'entreprise*. Paris, Masson, 220 p.

LESCA, Humbert (1992). Le problème crucial de la veille stratégique : la construction du puzzle. *Revue Annales des Mines*, avril, pp. 67-71.

LESCA, Humbert, RAYMOND, Louis (1993). Expérimentation d'un système expert pour l'évaluation de la veille stratégique dans les PME. *Revue internationale PME*, Québec, Canada, vol. 6, n°1, pp. 49-65.

LESCA, Humbert (1994). Veille stratégique pour le management stratégique : état de la question et axes de recherche. *Economie et sociétés*. Série Sciences de Gestion, n°20, vol 5, pp. 31-50.

LESCA, Humbert (1994). *Veille stratégique : L'intelligence de l'entreprise*. Editions Aster, Villeurbanne, 154 p.

LESCA, Humbert (1995). Comment ne pas être noyé sous les informations. *Actes du colloque VSST*, Toulouse.

LESCA, Humbert, CARON, Marie-Laurence (1995). Veille stratégique : créer une intelligence collective au sein de l'entreprise. *Revue Française de gestion*, sept-oct., pp. 58-68.

LESCA, Humbert (1996). Comment sélectionner les informations de veille stratégique. *Actes du colloque AIMS*, Lille.

Bibliographie

LESCA, Humbert (1997). Veille stratégique, concepts et démarche de mise en place dans l'entreprise. *Guides pour la pratique de l'information scientifique et technique*, Ministère de l'Education Nationale, de la Recherche et de la technologie, 27 p.

LESCA, Humbert, SCHULER, Maria (1998). Veille stratégique : comment ne pas être noyé sous les informations. *In Economies et sociétés*, Sciences de la gestion, Série S.G., n°2/1998, pp. 159-177.

LESCA, Humbert, CHOKRON, Michel (2000). Intelligence d'entreprise : retours d'expériences. *Actes du 5ème colloque de l'AIM*, Montpellier, 8-10 novembre, 19 p.

LESCA, Humbert (2001). Veille stratégique : passage de la notion de signal faible à la notion de signe d'alerte précoce. *Colloque VSST 2001*, Barcelone, Actes du colloque, Textes des communications, tome 1, pp. 270-277.

LESCA, Humbert (2008). Gouvernance d'une organisation : prévoir ou anticiper ? *Revue des sciences de gestion*. Direction et gestion (RGS), 2008/3-4, n° 231-232, pp 11-17, http://www.cairn.info/article.php?ID_ARTICLE=RSG_231_0011 (consulté le 11/01/2015).

LESCA Humbert, LESCA Nicolas (2011). *Les signaux faibles et la veille anticipative pour les décideurs : méthodes et applications*. Hermès - Lavoisier, Science Publications, coll. Business, économie et société, Paris, 248 p.

LESER, Eric, THOMPSON, Gordon (2003). Etats-Unis - Indian-Point : la centrale de tous les dangers. *Sortir du nucléaire*, n°21 avril 2003, article paru dans le journal Le Monde du 31/01/2003, http://www.sortirdunucléaire.org/Indian-Point-la-centrale-de-tous (consulté le 20/10/2014).

LESOURNE, Jacques (2008). *L'énergie nucléaire et les opinions publiques européennes*. IFRI, vol.2, Gouvernance européenne et géopolitique de l'énergie, 146 p.

LEVET, Jean-Louis (2008). *Les pratiques de l'Intelligence Economique : huit cas d'entreprises*. Economica, coll. l'Intelligence Economique, 159 p.

LI, Bo (2012). *Mesurer et améliorer la qualité des corpus comparables*. Thèse : Informatique, Université de Grenoble, juin 2012, 121 p.

LI, Charles N, THOMPSON, Sandra A (1974). An explanation of world order change : SVO > SOV. *Foundations of language*. 12, 2, pp. 201-214.

LIBAERT, Thierry, WESTPHALEN, Marie-Hélène (1999). *Communicator : Toute la communication d'entreprise*. Paris, Dunod, 6ème édition, 2012, coll. Livres en Or, 640 p.

LIEBERMANN, Alexandre (2014). La stratégie énergétique chinoise. http://les-yeux-du-monde.fr/actualite/asie-oceanie/18195-la-strategie-energetique-chinoise (consulté le 01/04/2015).

LIU, Bing (2007). *Web Data Mining : Exploring Hyperlinks, Contents and Usage Data*. Springer, 1ère édition, décembre 2006, 532 p., 2ème édition, juillet 2011, 622 p.

LIU, Y (1987). New advances in computers and natural language processing in China. *Information Science*. Vol.8, pp. 64-70.

LOCATELLI, Catherine, MARTIN-AMOUROUX, Jean-Marie (2005). L'intégration internationale des industries chinoises de l'énergie et ses conséquences géopolitiques. Laboratoire d'Economie de la Production et de l'Intégration Internationale. Département Energie et Politiques de l'Environnement, *5e colloque international sur l'économie chinoise*, Cerdi, Clermont-Ferrand, 20-21 octobre 2005.

MACHENAUD, Hervé (2005). La Chine bientôt le centre de gravité de l'industrie électrique mondiale. *Revue de l'énergie*. Paris, éditions techniques et économiques, n°563, pp. 25-28.

Bibliographie

MACMURRAY, Erin, SHEN, Liangcai (2010). Textual Statistics and Information Discovery : Using Co-occurrences to Detect Events. *In Proceedings Veille Stratégique Scientifique and Technologique*, Toulouse, France, octobre 2010.

MACMURRAY, Erin (2012). *Discours de presse et veille stratégique d'événements. Approche textométrique et extraction d'informations pour la fouille de textes*. Thèse : Sciences du langage, Université Sorbonne Nouvelle – Paris 3, juillet 2012, 432 p.

MADSLIEN, Jorn (2012). *China's car market matures after ultrafast growth.* BBC News, 22 avril 2012, http://www.bbc.co.uk/news/business-17786962 (consulté le 28/03/2015).

MAINGUENEAU, Dominique (1995). Présentation, In : *Langages*, 29ème année, n°117, Les analyses du discours en France, pp. 5-11.

MAINGUENEAU, Dominique (2012). Que cherchent les analystes du discours ?, *Argumentation et Analyse du Discours*, septembre 2012, mis en ligne le 15 octobre 2012,. URL : http://aad.revues.org/1354 (consulté le 22 avril 2015)

MALRIEU, Denise, RASTIER, François (2001). Genres et variations morphosyntaxiques. *Traitement Automatique des Langues (TAL)*, vol. 42, n°2, pp. 548–577, http://www.revue-texto.net/Inedits/Malrieu_Rastier/Malrieu-Rastier_Genres1.html (consulté le 26/04/2015).

MARCON, Christian, MOINET, Nicolas (2000). *La stratégie réseau : Essai de stratégie*. 00H00 Editions, coll. Stratégie, 235 p.

MARCON, Christian, MOINET, Nicolas (2006). *L'intelligence économique*. Dunod, 124 p.

MARGOT, Jean-Claude, MOUNIN, Georges (1990). *Traduire sans trahir : la théorie de la traduction et son application aux textes bibliques.* Lausanne, éditeur l'âge d'Homme, 388 p.

MARMUSE, Christian (1992). *Politique générale : langages, intelligence, méthode et choix stratégiques*. Ed. Economica, coll. Gestion, 2ème édition 1996, 646 p.

MARTIN-AMOUROUX, Jean-Marie (2004). Perspectives énergétiques mondiales. *Techniques de l'ingénieur. Génie énergétique*. Paris, vol.B3, n°BE8515.

MARTINET, Alain-Charles, PETIT, Georges (1982). *L'entreprise dans un monde en changement*. Ed. Seuil, 160 p.

MARTINET, Bruno, MARTY, Yves-Michel (1995). *L'intelligence économique : les yeux et les oreilles de l'entreprise*. Paris, éditions de l'organisation, Paris, 244 p.

MARTINEZ, William (2000). Mise en évidence de rapports synonymiques par la méthode des cooccurrences. *Actes des 5èmes Journées d'Analyse Statistique des Données textuelles*, Ecole Polytechnique de Lausanne, mars.

MARTINEZ, William (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*. Thèse pour le doctorat en Sciences du Langage, Université de la Sorbonne nouvelle - Paris 3, http://williammartinez.fr/coocs/page.php?P=1&L=1 (consulté le 13/02/2015).

MARTINEZ, William (2012). Au-delà de la cooccurrence binaire… Poly-cooccurrences et trames de cooccurrence, Corpus [En ligne], 11 | 2012, mis en ligne le 18 juin 2013, http://corpus.revues.org/2262 (consulté le 25/02/2016).

MARTINOT, Eric, LI, Junfeng, (2007). Powering China's Development : the Role of Renewable Energy. *Worldwatch Special Report*, Worldwatch Institute, novembre 2007, 50 p.

MARTRE, Henri (1994). *Intelligence économique et stratégie des entreprises*. AFNOR, Documentation Française, 213 p, http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/074000410/0000.pdf (consulté le 10/01/2015).

McENERY, Anthony, WILSON, Andrew (1996/2004). *Corpus Linguistics*. Edinburgh University Press, Series Edinburgh Textbooks in Empirical Linguistics, 2 nd edition, 235 p.

McENERY, Anthony, XIAO, Zhonghua (2007). Parallel and comparable corpora: What is happening? *Incorporating Corpora : The Linguist and the translator*. Edited by Gunilla Anderman and Margaret Rogers, Multilingual Matters, Chapter 2, pp. 18-31.

McENERY, Anthony, XIAO, Zhonghua (2007). Parallel and comparable corpora: The state of play. *Corpus-Based Perspectives in Linguistics*, Yuji Kawaguchi, Toshihiro Takagaki, Nobuo Tomimori and Yoichiro Tsuruga editors, Amsterdam, John Benjamins Publishing Company. pp. 131–145.

MERCIER, Arnaud, CHARON, Jean-Marie (2004). *Armes de communication massive : informations de guerre en Irak, 1991-2003*. Paris, CNRS éditions, coll. CNRS Communication, 216 p.

MERENNE-SHOUMAKER, Bernadette (2007). *Géographie de l'énergie : acteurs, lieux et enjeux*. Paris, Belin Sup Géographie, 272 p.

MERITET, Sophie (2009). Aujourd'hui, que peut-on attendre des Etats-Unis dans les discussions énergie-environnement ? *Economie et Société*, ISMEA, Economie et Climat, 8 p, http://www.meritet.net/uploaded/1319295698.pdf (consulté le 09/04/2015).

MERLAND, Jean-Pierre, BINOT, Christophe, CANSELL, Patrick et al (2005). *L'Intelligence Economique Appliquée à la Direction des Systèmes d'Information : Démarche et Fiches Pratiques*. Mastère spécialisé en Intelligence Economique et Stratégique. EISTI : Ecole Internationale des Sciences et du Traitement de l'information, 63 p, http://www.cigref.fr/cigref_publications/RapportsContainer/Parus2005/2005_-_Intelligence_Economique_appliquee_a_la_Direction_des_Systemes_d_Information_web.pdf (consulté le 15/01/2015).

MIAO, Jun, SALEM, André (2008). Comparaisons textométriques de traductions franco-chinoises. In *Explorations textométriques*, 20 p, http://lexicometrica.univparis3.fr/numspeciaux/special8/Mult2.pdf (consulté le 13/02/2015).

MIAO, Jun (2012). *Approches textométriques de la notion de style du traducteur. Analyses d'un corpus parallèle français-chinois : Jean-Christophe de Romain Rolland et ses trois traductions chinoises*. Thèse de doctorat : Traductologie, Université Sorbonne Nouvelle – Paris 3, 20 avril 2012, 513 p.

MINISTERE DE L'ECONOMIE DES FINANCES ET DE L'INDUSTRIE, MINISTERE DE L'ECOLOGIE ET DU DEVELOPPEMENT DURABLE (2006). *Division par quatre des émissions de gaz à effet de serre de la France à l'horizon 2050*. Rapport du gouvernement sur les EGES, Observatoire de l'AIE, rapport du groupe de travail sous la présidence de Christian de Boissieu, août 2006, http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/38/027/38027675.pdf (consulté le 17/12/2014).

MOIRAND, Sophie (2007). *Les discours de la presse quotidienne. Observer, analyser, comprendre.* Paris, Presses Universitaires de France, 179 p.

MORIN, Emmanuel, DAILLE, Béatrice (2006). Comparabilité de corpus et fouille terminologique multilingue. *Traitement automatique des langues, TAL*, vol.47, n°1, pp. 113-136.

MOUSNIER, Jean-Philippe (2005). Le modèle de management par l'intelligence économique, premier

support structuré de veille stratégique pour l'entreprise et le territoire. *Congrès Veille stratégique, scientifique et technologique*, Université Paul Sabatier, Toulouse, 546 p, http://www.xploorew.com/VSST/Colloque/04-Toulouse/Salle-B/B-08-MOUSNIER.pdf (consulté le 15/01/2015).

MULLER, Charles (1967). *Étude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*. Paris, Larousse, réimpression 2000, Slatkine, coll. Tavaux de linguistique quantitative, 379 p.

MULLER, Charles (1968). *Initiation à la statistique linguistique*. Paris, Larousse, 247 p.

MULLER, Charles (1973). *Initiation aux méthodes de la statistique linguistique*. Paris, Hachette, réimpression 1992, Honoré Champion, coll. Unichamp.

MULLER, Charles (1975). Fréquence et probabilité d'emploi. Travaux de linguistique et de littérature, tome 13, 1, Strasbourg: Université de Strasbourg, pp. 219-225.

MULLER, Charles (1977). *Principes et méthodes de statistique lexicale*. Paris, Hachette, réimpression 1992, Paris, Honoré Champion-Slatkine, 210 p.

MULTON, Bernard (1998). L'énergie sur la terre : analyse des ressources et de la consommation. La place de l'énergie électrique. *Revue 3EI*, septembre 1998, pp. 29-38.

MULTON, Bernard (2003). Production d'énergie électrique par sources renouvelables. *Techniques de l'ingénieur, Génie électrique*. Paris, Techniques de l'ingénieur, vol. D8, n°D4005, pp. 1-11.

MULTON, Bernard, ROBOAM, Xavier, DAKYO, Brayima, NICHITA, Christian, GERGAUD, Olivier, BEN AHMED, Hamid (2004). Aérogénérateurs électriques. *Techniques de l'ingénieur. Génie électrique*. Paris, Techniques de l'ingénieur, vol D7, n°D3960.

MUNTEANU, Dragos Stefan, MARCU, Daniel (2002). Processing Comparable Corpora With Bilingual Suffix Trees. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP-2002), Philadelphia, PA, July 6-7.

NAÏM, Patrick, BAZSALICZA, Mylène (2001). *Data mining pour le Web. Profiling-Filtrage collaboratif Personnalisation client*. Eyrolles, coll. Solutions d'entreprise, 296p.

NAKAMURA-DELLOYE, Yayoi (2007). *Alignement automatique de textes parallèles français-japonais.* Thèse de doctorat en linguistique, Université Paris 7, 580 p.

NEE, Emilie (2008). Insécurité et élections présidentielles dans le journal Le Monde. *Lexicometrica*, numéro thématique « Explorations Textuelles », vol. 1 « Corpus et problèmes », (dir) S. Fleury & A. Salem, http://lexicometrica.univ-paris3.fr/numspeciaux/special8/Volume1.pdf (consulté le 13/02/2015).

NEE, Emilie (2009). *Sûreté, sécurité, insécurité. D'une description lexicologique à une étude du discours de presse : la campagne électorale 2001-2002 dans le quotidien Le Monde*. Thèse de doctorat : Sciences du Langage, Université Sorbonne Nouvelle – Paris 3, 30 novembre 2009, 437 p.

NODE-LANGLOIS, Fabrice (2015). La Chine devient le plus gros investisseur dans le solaire et l'éolien. Le Figaro.fr Economie, publié le 13/01/2015, http://www.lefigaro.fr/conjoncture/2015/01/10/20002-20150110ARTFIG00021-la-Chine-devient-le-plus-gros-investisseur-dans-le-solaire-et-l-eolien.php (consulté le 21/02/2015).

OCDE (2013). Etudes économiques de l'OCDE : Chine – 2013.Editions OCDE, mars 2013, vol ; 2013/4, 180 p, https://books.google.fr/books?id=PZ8zAAAAQBAJ&pg=PA166&lpg=PA166&dq=consommation+pollution+climatiseurs+Chine&source=bl&ots=8DAZF7SeTJ&sig=pygsqlmHCq4IbqUm_jW7zx-0RXI&hl=fr&sa=X&ei=YuwWVar3JcHTUcjJgsgH&ved=0CDoQ6AEwBA#v=onepage&q=consom

mation%20pollution%20climatiseurs%20Chine&f=false (consulté le 25/03/2015).

ORIOL, Louise, MEINZEL, Thomas, PESCIA, Dimitri, LEHMANN, Frédéric (2013). Comparaison des prix de l'électricité en France et en Allemagne. *Les cahiers de la DG Trésor*, n°2013-5, nov. 2013, 23 p.

OUSTINOFF, Michaël (2013). La diversité linguistique, enjeu central de la mondialisation. *Revue Française des Sciences de l'information et de la communication*, 2/2013 : communication et diversité culturelle.

PARISSE, Christophe (2009), La morphosyntaxe : Qu'est-ce qu'est ? - Application au cas de la langue française ? *Rééducation Orthophonique*, 2009, 47 (238), pp.7-20.

PEAN, Pierre, COHEN, Philippe (2003). *La face cachée du Monde : du contre-pouvoir aux abus de pouvoir*. Fayard, éditeur Mille et Une Nuits, 631 p.

PELLATON, Michel, DELOBBE, Georges (2006). *Histoire de la presse écrite*. Pemf, coll. un œil sur l'histoire, 103 p.

PICONE, Michael D. (1992). Le Français face à l'anglais : aspects linguistiques. In: *Cahiers de l'Association internationale des études francaises*, 1992, N°44. pp. 9-23.

PINCEMIN, Bénédicte (2011). Sémantique interprétative et textométrie – Version abrégée. *Corpus* 10/2011, pp. 259-269, http://corpus.revues.org/2121 (consulté le 13/01/2015).

PINEIRA, Carmen, TOURNIER, Maurice (2009). Ségolène Royal entre François Bayrou et Nicola Sarkozy. Approche lexicométrique, *Mots*, n° 89, Les langages du politique : 2007 : débats pour l'Elysée, avril 2009, ENS Editions, pp. 83-104.

POIBEAU Thierry (2003). *Extraction automatique d'information. Du texte brut au web sémantiqu*e. Paris, Hermès Sciences, p. 239.

POIBEAU, Thierry (2005). Sur le statut référentiel des entités nommées. *Proceedings Traitement Automatique des Langues Naturelles*, Dourdan, France.

PRICE, Lynn, LEVINE, Mark, ZHOU, Nan, FRIDLEY, David et al (2011). Assessment of China's energy-saving and emission-reduction accomplishments and opportunities during the 11th Five Year Plan. *Energy Policy* 39, pp. 2165-2178.

QUACH, Alina (2011). L'opinion publique n'ébranle pas la politique nucléaire du pays. Journal dirigeant.fr, un autre regard sur l'entreprise, le journal en ligne des entrepreneurs, publié le 3/05/2011, http://www.jeune-dirigeant.fr/011-294-Fukushima-vu-de-Chine-du-Chili-et-d-Espagne.html (consulté le 15/04/2015).

RAPP, Reinhard, ZWEIGENBAUM, Pierre, SHAROFF, Serge (2016). *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora*. ELDA, Portorož, Slovenia, mai 2016. https://comparable.limsi.fr/bucc2016/pdf/BUCC.pdf (consulté le 15/06/2016).

RAPP, Reinhard, SHAROFF, Serge, ZWEIGENBAUM, Pierre (2016). *Recent advances in machine translation using comparable corpora*. Natural Language Engineering, vol. 22, chap. 4, pp. 501-516, juillet 2016. http://dx.doi.org/10.1017/S1351324916000115 (consulté le 01/07/2016).

RASTIER, François (1987). *Sémantique interprétative*. Paris, Presses universitaires de France, 284 p.

RASTIER, François, PINCEMIN, Bénédicte (1999). Des genres à l'intertexte. *Cahiers de praxématique 33*, Montpellier, Presses de l'université Montpellier 3, pp. 83-111.

RASTIER, François (2011). *La mesure er le grain. Sémantique de corpus*. Paris, Honoré Champion, Collection Lettres numériques, 280 p.

Bibliographie

RAVIGNAN de, Antoine (2008). Energies renouvelables : la Chine carbure plus vert. Worldwatch Institute, *Alternatives Economiques*, janvier 2008, n°265.

ROBERT, Magali (2010). Puissance Chine, la stratégie d'affirmation internationale chinoise. *Fiche de l'IRSEM*, mars 2010, 14 p.

RÜDINGER, Andreas (2013). Le tournant énergétique allemand : État des lieux et idées pour le débat français. *Les cahiers de Global Chance*, mars 2013, n°33, 11 p.

SALEM, André (1987). *Pratique de segments répétés*. Publication de l'INaLF, coll. St.Cloud, Paris, Klinksieck, 333p.

SALEM, André (1988). Approches du temps lexical. Statistique textuelle et séries chronologiques. *Mots* n°17, octobre 1988, pp. 105-143.

SALEM, André (1991). Les séries textuelles chronologiques. *Histoire & Mesure*, VI-1/2, pp. 149-175.

SALEM, André (1993). *Méthodes de la statistique textuelle*. Thèse pour le doctorat d'Etat ès lettres et sciences humaines, Université Sorbonne Nouvelle – Paris 3, mars 1993, 3 vol, 998 p.

SALEM, André (1994). La lexicométrie chronologique. *Actes du colloque de lexicologie politique « Langages de la Révolution »*, coll. St. Cloud, Paris, Klincksieck.

SALEM, André (2006). Proximités Segmentales. *Actes 8èmes Journées Internationales d'Analyse Statistique des Données Textuelles, JADT06*, Université Franche-Comté, Vol II, pp. 839-849, http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-075.pdf (consulté le 15/01/2015).

SALEM, André, LAMALLE, Cédric (2009). Lexico 3 : outils de statistiques textuels, manuel d'utilisation, version 3.6. Université Sorbonne Nouvelle Paris 3.

SANSONETTI, Luiggi (2010). *Apports de la textométrie pour l'analyse de corpus d'interactions verbales entre adulte et enfant au cours de l'acquisition du langage*. Thèse : Sciences du langage, Université Sorbonne Nouvelle – Paris 3, 4 décembre 2010, 332 p.

SAWIN, Janet L, BHATTACHARYA, Sribas Chandra, MARTINOT, Eric et al (2012). Rapport mondial 2012 sur les énergies renouvelables. *REN21*. Paris, 20 p, www.ren21.net/gsr (consulté le 1/03/2015).

SCHNEIDER, Mycle (2000). Changement climatique et énergie nucléaire. Rapport commandité par le WWF (World Wide Fund for Nature), WISE Paris, édition française août 2000, 22 p, http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/34/066/34066680.pdf (consulté le 25/03/2015).

SCHNEIDER, Mycle (2001). Nucléaire *plus* effet de serre. Wise-Paris, avril 2001, 14 p, http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/37/066/37066512.pdf (consulté le 17/12/2014).

SCHNEIDER, Mycle (2008). Changement climatique et énergie nucléaire. Wise-Paris, 21 p, http://www.wise-paris.org/francais/rapports/NucléaireClimatWISEParis.pdf (consulté le 17/12/2014).

SCHNEIDER, Mycle (2008). Le nucléaire en France : au-delà du mythe. Rapport commandité par le Groupe des Verts/Alliance Libre Européenne au Parlement Européen, Bruxelles, décembre 2008, 43 p.

SCHNEIDER, Mycle, FROGGATT, Antony (2014). 2012-2013 world nuclear industry status report. *Bulletin of the Atomic Scientists*, vol.70, issue 1, janvier-février 2014, pp. 70-84.

SERRES, Michel (1996), *Atlas*, Paris, Flammarion, collection Champs, 279 p.

SHAROFF, Serge, BABYCH, Bogdan, HARTLEY, Anthony (2006). Using comparable corpora to solve

problems difficult for human translators. *Joint COLING-ACL*. http://mt-archive.info/Coling-ACL-2006-Sharoff.pdf (consulté le 1/07/2016)

SHAROFF, Serge, RAPP, Reinhard, ZWEIGENBAUM, Pierre, FUNG, Pascale (2013). *Building and Using Comparable Corpora*. Springer, 347p.

SHEN, Liangcai (2009). *Veille stratégique et analyses textuelles*, mémoire de Master II sous la direction d'André Salem, Université Paris 3.

SHEN, Liangcai, SALEM, André (2009). Qu'en pensent les Chinois ? Essai d'exploration de l'opinion publique chinoise à travers des documents disponibles sur la toile. [Bad karma], *Lexicometrica,* numéro special, http://lexicometrica.univ-paris3.fr/numspeciaux/special8.htm

SLODZIAN, Monique (2000). L'émergence d'une terminologie textuelle et le retour du sens. *Le sens en terminologie*, 2000, p. 61-85.

SRIVASTAVA, Jaideep, COOLEY, Robert Walker, DESHPANDE, Mukind, TAN, Pang-Ning (2000). Web Mining : Discovery of Interesting usage patterns from Web data. University of Minnesota 200 Union St SE, Minneapolis, In *ACM SIGKDD Exploration newsletter*, vol.1, Issue 2, pp. 12-23.

SUN, Maosong, SHEN, Dayang, TSOU, Benjiamin K (1998). Chinese word segmentation without using lexicon and hand-crafted training data. *In proceedings of the 36th ACL*, pp. 1265-1271.

SUN, Maosong, ZHOU, Jiayan (2001). Commentaires sur la recherche de la segmentation automatique du chinois" (汉语自动分词评述). *Revue linguistique contemporaine* (当代语音学), n°1, pp. 22-32.

SWAMINATHAN, Monkombu Sambavisan, RAHMANIAN, Maryam, BERTINI, Catherine et al. (2013). Groupe d'experts de haut niveau, 2013. Agrocarburants et sécurité alimentaire. Rapport du Groupe d'experts de haut niveau sur la sécurité alimentaire et la nutrition, Comité de la sécurité alimentaire mondiale, Rome, juin 2013, 156 p, http://www.fao.org/3/a-i2952f.pdf (consulté le 1/03/2015).

TAN, Xiaomei, ZHAO, Yingzhen, POLYCARP, Clifford, BAI, Jianwen (2013). China's overseas investments in the wind and solar industries : trends and drivers. World resources Institute, April 2013, 24 p.

TELLIER, Isabelle. Introduction à la fouille de textes. Université de Paris 3 – Sorbonne Nouvelle, 71 p. URL : http://www.lattice.cnrs.fr/sites/itellier/poly_fouille_textes/fouille-textes.pdf

TEUBERT, Wolfgang (1996). Comparable or parallel corpora? *International Journal of lexicography*, vol. 9, n° 3, pp. 238-264.

TEUBERT, Wolfgang (2009). Corpus Linguistics : an alternative. *Semen*. Editions électronique scientifique, Fédération de revues en sciences humaines, n°27, pp. 185-211, http://semen.revues.org/8914 (consulté le 12/01/2015).

TLFi (1960). Trésor de la langue française informatisé, http://www.cnrtl.fr/definition/corpus (consulté le 23/04/2015).

TONNAC de, Alain, PERVES, Jean-Pierre (2012). Le programme nucléaire chinois. SFEN, GR21, groupe de réflexion sur l'énergie et l'environnement au 21ème siècle, octobre 2012, 6 p, http://www.sfen.org/IMG/pdf/programme_nucléaire_chinois_octo2013.pdf (consulté le 18/02/2015).

TUFFERY, Stephane (2012). *Data mining et statistique décisionnelle : l'intelligence des données*. Paris, éditions Technip, 4ème éd., 826 p.

TURENNE, Nicolas (2000). *Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles*. Thèse mention : Sciences, spécialité : informatique,

Université Louis Pasteur Strasbourg, 24 novembre 2000, 198 p., https://tel.archives-ouvertes.fr/tel-00006210/document (consulté le 15/01/2015).

VALETTE, Mathieu (2004). Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet, approches sémantiques du document numérique. P.Enjalbert et M.Gaio, (eds), *Actes du 7ème colloque international sur le document électronique*, 22-25 juin 2004, pp. 215-230.

VALETTE, Mathieu (2008). Textes, documents numériques, corpus. Pour une science des textes instrumentée. *Syntaxe et sémantique*, n°9/2008, pp. 9-14.

VALETTE, Mathieu, ESTACIO-MORENO, Alexander, PETITJEAN, Etienne, JACQUEY, Evelyne (2006). Eléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens, Verbum ex machina. *Actes de la 13ème conférence sur le traitement automatique des langues naturelles* (TALN 06), P.Mertens, C.Fairon, A.Dister, P.Watrin (eds), cahiers du Central, 2.1, UCL, presses universitaires de Louvain, vol.1, pp. 357-366.

VALETTE, Mathieu, RASTIER, François (2006). Prévenir le racisme et la xénophobie, propositions de linguistes. *Les langues modernes*, 2/2006, Frath P. (ed), Enseignez le mal, pp. 68-77.

VALETTE, Mathieu, SLODZIAN, Monique (2008). Sémantique des textes et recherche d'information. Condamines A., Poibeau, T. (ed), Extraction d'information : l'apport de la linguistique, *revue française de linguistique appliquée*, vol. XIII-1, juin 2008, pp. 119-133.

VENIARD, Marie (2007). *La nomination d'un événement dans la presse quotidienne nationale. Une étude sémantique et discursive : la guerre en Afghanistan et le conflit des intermittents dans le Monde et le Figaro.* Thèse : Sciences du langage, Université Sorbonne Nouvelle Paris 3, 1er décembre 2007, 488 p.

VERNIER, Jacques (2014). *Les énergies renouvelables*. Presses Universitaires de France Paris, 7ème éd., coll. Que sais-je, 128 p.

VERONIS, Jean (2000). *Parallel Text Processing: Alignment and Use of Translation Corpo*ra. Springer Science & Business Media, 30 sept. 2000, 402 p.

VERONIS, Jean (2000). Alignement de corpus multilingues. In *J.-M. Pierrel* (Ed.), Ingénierie des langues, Paris, Editions Hermès, chapitre 6, pp. 151-171, http://sites.univ-provence.fr/veronis/pdf/2000hermes6.pdf (consulté le 27/04/2015).

VIEILLEFOSSE, Aurélie (2009). Le changement climatique : quelles solutions? La documentation française, *Les études*, n°5290-91, 184p.

VOÏTA, Thierry (2010). Le géant noir. Politique économique du charbon en Chine. Sciences Po, septembre 2010, 8 p, http://www.ceri-sciences-po.org (consulté le 1/03/2015).

WALD, Matthew (2010). Vermont Nuclear Plant up for sale. Le New York Times, Business Day, Energy & Environment, publié le 4/11/2010, http://www.nytimes.com/2010/11/05/business/energy-environment/05nuke.html?_r=4&src=busln& (consulté le 20/10/2014).

WANG, Shujie, YUAN, Peng, LI, Dong, JIAO, Yuhe (2011). An overview of ocean renewable energy in China. *Renewable and Sustainable Energy Reviews 15*, septembre 2011, pp. 91-111.

WANG, Xuan, RAES, Vincent (2012). Le nucléaire chinois après Fukushima. China Institute, Economie, avril 2012, 15 p, http://www.china-institute.org/articles/Le_nucléaire_chinois_apres_Fukushima.pdf (consulté le 20/12/2014).

WANG, Yong, HODGES, Julia, TANG, Bo (2003). Classification of Web documents using a naive Bayes

method. In ICTAI 03 : *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, IEEE Computer Society, pp. 560-564.

WILENSKY, Harold (1967). Organizational Intelligence : Knowledge and Policy in Government and Industry. *American Sociological Review*, Vol. 33, Issue 1, feb 1968, pp. 131-132, http://www.gwu.edu/~ccps/etzioni/documents/D22_000.pdf (consulté le 12/01/2015).

WONG, Kam-Fai, LI, Wenjie, XU, Ruifeng, ZHANG, Zheng-Sheng (2010). *Introduction to Chinese Natural Language Processing*. Toronto: Morgan & Claypool Publishers.

WRIGHT, Keith (2006). Using Open Source Common Sense Reasoning Tools in Text Mining Research. *The international Journal of Applied Management and Technology*, vol.4, n°2, pp.349-387, http://www.swdsi.org/.../SWDSI_submission_usingOpenToolsinTextMining%20(T4D2).pdf

WU, WeiGuang, HUANG, Jikun, DENG, XianZheng (2009). « Potential land for the planting of *Jatropha curcas* as feedstock for biodiesel in China ». *Sciences China, Earth Sciences,* vol.53, n°1, pp. 120-127p, http://link.springer.com/article/10.1007/s11430-009-0204-y#page-1 (consulté le 1/03/2015).

YE, Liming, YANG, Jun, VERDOODT, Ann, MOUSSADEK, Rachid, VAN RANST, Eric (2010). China's food security threatened by soil degradation and biofuels production. *Proceedings of the 19th World Congress of Soil Science, Soil Solutions for a Changing World*.

ZAKARIA, Fareed (2008). *The Post-American World*. W.W.Norton & Company, 304 p.

ZANETTIN, Federico (1998). Bilingual comparable corpora and the training of translators. In *Meta* vol. 43, n° 4, pp. 613-630.

ZHANG Hua-Ping, LIU Qun, CHENG Xue-Qi, ZHANG Hao, YU Hong-Kui (2003). Chinese Lexical Analysis Using Hierarchical Hidden Markov Model, Second SIGHAN workshop affiliated with 41th ACL; Sapporo Japan, July, 2003, pp. 63-70.

ZHANG, Q (1989). Shanggu hanyu de SOV zixu ji dingyu houzhi. Yuyan jiaoxue yu yanjiu, 1, pp. 101-110.

ZHANG, Yan (2011). *La stratégie de développement de la Chine vue par le reste du monde.* http://fr.cntv.cn/program/journal/20110307/103445.shtml (consulté le 26/02/2015).

ZHOU, Di, DELBOSC, Anaïs (2013). Les outils économiques de la politique énergie-climat chinoises à l'heure du 12ème plan quinquennal. Etude climat, la recherche en économie du changement climatique, n°38, janvier 2013, 35 p, URL : http://www.cdcclimat.com (consulté le 27/02/2015).

ZHOU, Nan, FRIDLEY, David, McNEIL, Michael, ZHENG, Nina, KE, Jing, LEVINE, Mark (2011). China's Energy and Carbon Emissions Outlook to 2050. Berkely Lab, Ernest Orlando Lawrence Berkely National Laboratory, China Sustainable Energy Program of the Energy Foundation through the U.S. Department of Energy, avril 2011, 83 p.

ZHU, Xianli, PAN, Jiahua (2007). Le développement de la petite hydraulique en Chine. Energies renouvelables, développement et environnement : discours, réalités et perspectives. *Les Cahiers de Global Chance*, avril 2007, Traduction de Jean-Luc Thierry, pp.55-59, http://www.global-chance.org/IMG/pdf/GC23p55-59.pdf (consulté le 28/02/2015).

ZIMINA, Maria (2004). *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*. Thèse de doctorat : Sciences du Langage, Université Sorbonne Nouvelle – Paris 3, 26 novembre 2004, 328 p.

ZIMINA, Maria (2005). Topographie bi-textuelle et approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles. *Actes des 7èmes Journées scientifiques du Réseau de chercheurs « Lexicologie, Terminologie, Traduction »*, Institut supérieur de traducteurs et interprètes

(ISTI), Bruxelles (Belgique), 8-10 septembre 2005, pp. 175-186.

ZIPF, George Kingsley (1932). *Selected Studies of the Principle of Relative Frequency in Language.* Cambridge, Havard University Press, 51 p.

ZIPF, George Kingsley (1935). *The Psycho-Biology of Language*. Boston, Houghton Miffin company, annotated, reprint 1999, London : routledge, series: International Library of Psychology, 348 p.

ZONABEND, Françoise (2014). *La presqu'île au nucléaire : Three Mile Island, Tchernobyl, Fukushima… et après ?* Paris, éditions Odile Jacob, mai 2014 (1ère parution 1989), 242 p.

Index des termes

Index des termes

Index des auteurs

Index des auteurs

Index des auteurs

Figures et Tableaux

Liste des figures

Figures et Tableaux

Liste des tableaux

Annexe A : la veille et l'intelligence économique

A.1 La veille

Les différents types de veille

Les typologies caractérisant la veille sont multiples et varient en fonction des besoins des organismes et des sociétés. Les informations peuvent être regroupées soit selon leurs origines (confidentielles, semi-publiques, publiques), soit sont ponctuelles, périodiques, occasionnelles, stratégiques.
Nous retiendrons un exemple de classement réalisé par le CIGREF[227] et présenté ci-après :

- *« la veille ponctuelle est une démarche volontariste et ciblée pour combler rapidement un manque. Elle correspond à un état de l'art ou à une analyse de l'existant à un moment donné et dans un contexte donné. Elle se rapproche un peu d'une étude de marché,*
- *la veille occasionnelle est une surveillance organisée sur des thèmes cibles. Elle découle surtout d'une prise de conscience par l'entreprise de l'importance de la cible à surveiller mais, contrairement à la veille ponctuelle, les informations que l'on obtient sur ladite cible ne sont pas fixées en aval comme des objectifs à atteindre. La découverte des informations décisives sur la cible est due principalement au hasard et à l'intelligence et est par nature inattendue. Dans le but d'optimiser et d'augmenter le nombre de ces découvertes, l'entreprise tient donc à ce que la cible soit surveillée de manière permanente, ce qui distingue ce type de veille, de la veille ponctuelle qui est caractérisée par sa courte durée dans le temps,*
- *la veille périodique est une surveillance régulière de la cible, le rythme de récolte et de traitement des informations étant fortement lié à la périodicité de parution ou de manifestation des différentes sources qui les fournissent. Elle s'apparente aux bilans de sociétés, rapports d'études, articles de magazine, banques de données (BDD) mais aussi revues de presse ».*

Quant à la veille stratégique, elle a fait l'objet de nombreuses définitions, nous n'en proposerons que trois, dont celles de :

- Laroche et Nioche : *« Outil d'aide au processus de décision stratégique* » (Laroche et Nioche, 1994),
- D.Coudol et S.Gros[228] : *« Système d'aide à la décision qui observe et analyse l'environnement scientifique, technique, technologique et les impacts économiques présents et futurs pour en déduire les menaces et les opportunités de développement. Elle s'appuie essentiellement sur les informations ayant un caractère stratégique ou décisions importantes lui associant le terme de veille stratégique »*,
- Office québécois de la langue française[229] : *« Consiste en une activité de surveillance permanente de l'environnement interne ou externe d'une organisation qui doit permettre un repérage de signes ou d'indices révélateurs de changements importants ».*

[227] CIGREF : Club Informatique des Grandes Entreprises Françaises
[228] http://www.agentintelligent.com/veille/veille_strategique.html (consulté le 3/12/2009)
[229] http://gdt.oqlf.gouv.qc.ca/ficheOqlf.aspx?Id_Fiche=8869964 (consulté le 7/04/2016)

Annexe A

Le processus de veille

Selon la norme AFNOR XP X5-0-053 (avril 1998), le processus de veille peut se décomposer en cinq grandes étapes :

- expression des besoins, une étape qui relève plus particulièrement de la direction générale d'une entreprise (Marcon et Moinet, 2000 ; 2006),
- recherche des sources,
- collecte des données et surveillance,
- traitement et analyse,
- diffusion de l'information stratégique.

De plus, ce processus nécessite de communiquer les informations et d'effectuer de nombreux retours en arrière afin d'éviter principalement toute erreur d'incompréhension.

Le processus de veille ainsi que la description détaillée de ses principales étapes sont disponibles dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-A-Veille-IE.

La veille est désormais une activité à part entière de la stratégie des entreprises et présente de multiples aspects (Laurelut, Arlabosse, Azoulay et al, 1998). Les caractéristiques de la veille stratégique sont également disponibles dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-A-Veille-IE.

A.2 Panorama de l'intelligence économique en France : dates et définitions

Dans les années 1980-1990, la France a posé les grands jalons de la politique de l'intelligence économique. Toute une série de publications, de circulaires, de décrets ainsi que la mise en place d'une organisation gouvernementale ont suivi ces prises de décision politique. Des informations complémentaires sont disponibles dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-A-Veille-IE.

A.3 Sources formelles et informelles

Nous avons vu que les informations recueillies au cours de la veille n'ont pas toute la même provenance, ni le même degré de confidentialité. Le tableau A.2 expose les avantages et les inconvénients de ces sources formelles et informelles et le tableau A.3 présente les 3 grandes catégories d'information : information blanche, information grise et information noire. Ces 2 tableaux sont disponibles dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-A-Veille-IE.

(Lien du serveur dédié :
https://drive.google.com/folderview?id=0B8XHfHwNzWAAeDN6UEc1b2dGa1U&usp=sharing)

Annexe B : dépouillement général du corpus comparable

Dénomination générique du corpus comparable

Pour faciliter la lecture des tableaux et des figures, la dénomination du corpus comparable contenant les trois sous-corpus, français, américain et chinois, devient la suivante :

- Nom du corpus comparable : **ENRG** pour énergies et environnement.
- Nom des trois sous-corpus : **ENRG_FR**, **ENRG_US**, **ENRG_CN**.

B.1 Sous-corpus français, journal « Le Monde »

Le sous-corpus français est constitué à partir d'articles recensés dans la rubrique « Planète » du Monde, datés du 24-09-1999 au 17-04-2012, la taille du fichier en .txt est de 16,8 Mo.

Pourquoi cette période ?
La date de début correspond au premier article disponible de cette rubrique du journal et la date de fin est celle du lancement de la requête informatique.

Principales caractéristiques textométriques du sous-corpus français

Tableau B.1
ENRG_FR : principales caractéristiques textométriques par année

Nombre d'occurrences: 2783702 Nombre de formes: 76848
Nombre d'hapax: 32151 Fréquence maximale: 153265

	Num	Partie	Occurences	Formes	Hapax	Fmax	Forme
✓	1	1999	4838	1691	1150	306	de
✓	2	2000	13553	3738	2389	759	de
✓	3	2001	8727	2673	1781	486	de
✓	4	2002	23728	5161	2982	1297	de
✓	5	2003	38134	7393	4254	2166	de
✓	6	2004	23926	4764	2673	1099	de
✓	7	2005	142020	14731	6805	8021	de
✓	8	2006	205758	18422	8243	11683	de
✓	9	2007	216616	19434	9073	11755	de
✓	10	2008	157850	17185	8437	8557	de
✓	11	2009	324697	24177	11066	17605	de
✓	12	2010	657393	35906	15790	35907	de
✓	13	2011	660245	35108	15239	36656	de
✓	14	2012	306217	25539	12210	16968	de

Le tableau B.1 ci-dessus montre que l'ensemble du sous-corpus compte 2 783 702 occurrences et 76 848 formes. Ce sous-corpus se divise en 43 333 paragraphes. La forme la plus fréquente est l'article partitif « de ». Les années 1999 et 2012 sont des périodes incomplètes. L'année civile complète où il y a le plus de mots est l'année 2011 avec 660 245 occurrences et celle qui en comporte le moins est l'année 2001 avec 8 727.

Annexe B

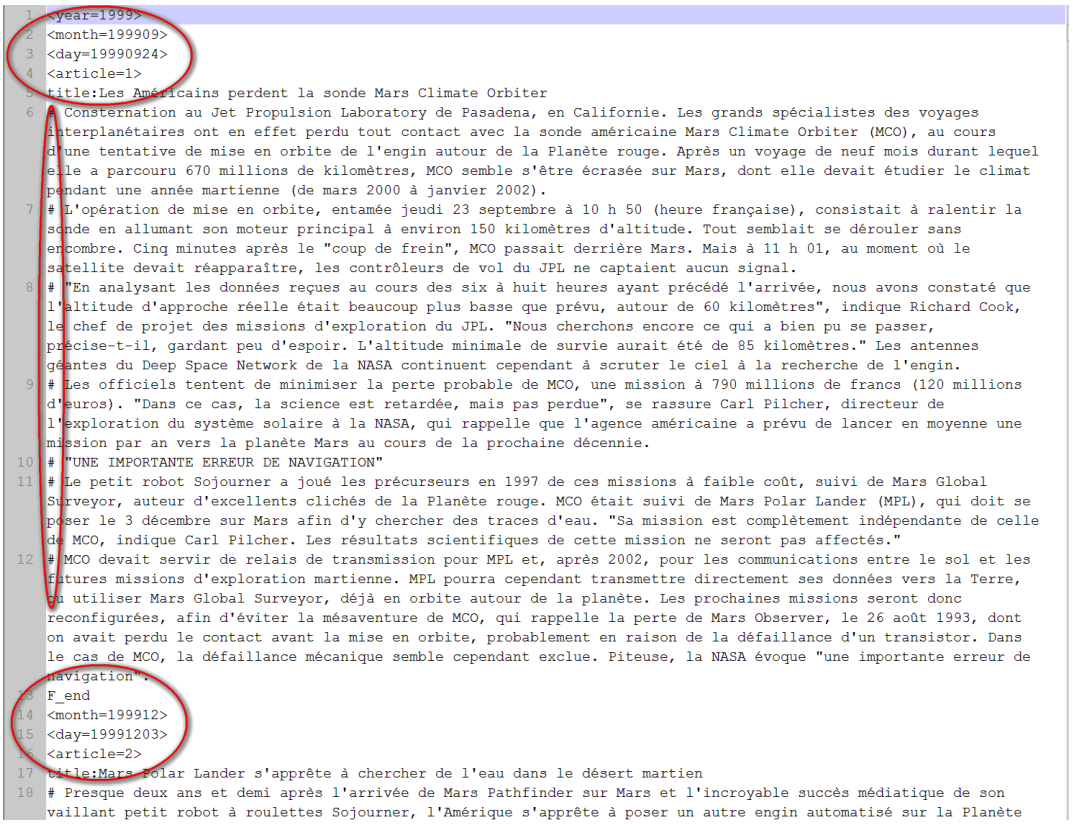


Figure B.1
ENRG_FR : structure du sous-corpus

Le sous-corpus se décompose en différents empans textuels suivant une chronologie. Les articles sont triés par ordre croissant sur leurs dates de publication (aaaammjj), regroupés par jour, par mois, par année, à l'aide de programmes informatiques spécifiques. Chaque balise désigne le début d'un empan textuel (figure B.1) :

- <year=aaaa> : balise indiquant l'année des articles. Le sous-corpus français s'étale sur 14 années, de 1999 à 2012.
- <month=aaaamm> : balise indiquant le mois des articles.
- <day=aaaammjj> : balise indiquant le jour des articles, sachant que dans une journée, plusieurs articles sont susceptibles d'être regroupés sous cette balise. L'ensemble du sous-corpus contient 1 788 jours.
- <article=nnnn> : balise indiquant le numéro de chaque article.
- Le titre de chaque article est introduit par la chaîne de caractères « title :».
- Le signe « # » marque le début de chaque paragraphe de chaque article. Les textes rassemblés dans le sous-corpus français sont au nombre de 4 817 articles.
- La fin de chaque article est marquée par la chaîne de caractères « F_end », par la suite cette chaîne de caractères est remplacée par le signe « § » afin de faciliter les traitements de Lexico.

L'évolution du nombre d'articles du sous-corpus français

La figure B.2 ci-dessous montre l'évolution du nombre d'articles extraits du journal en ligne constituant le sous-corpus français du corpus comparable.

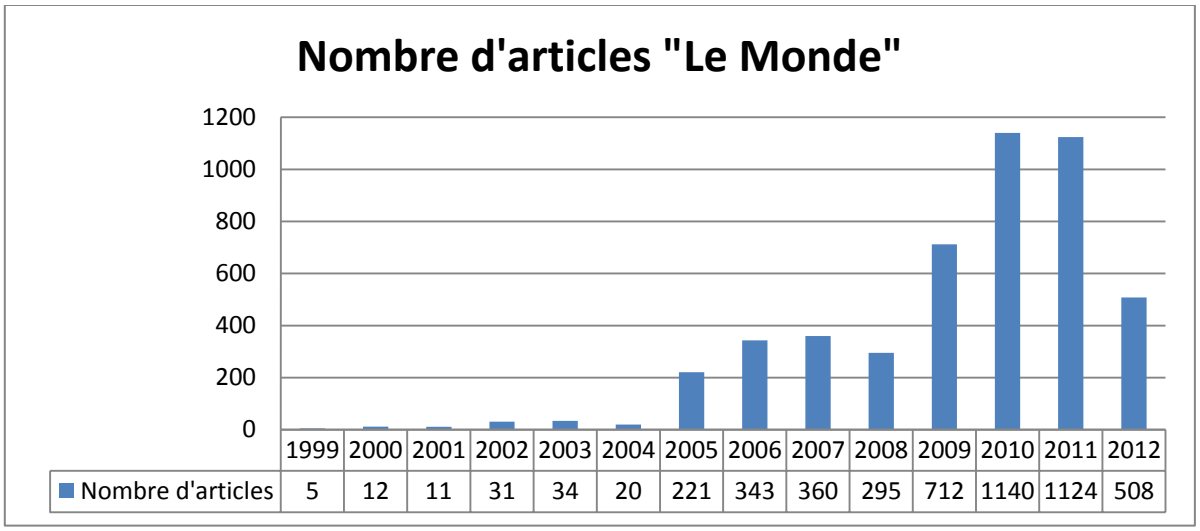


Figure B.2
ENRG_FR : évolution du nombre d'articles
du mois de septembre 1999 au mois d'avril 2012

Les années 2010 et 2011 se détachent des autres années par **leurs nombres d'articles**. Le nombre d'occurrences concernant ces deux années est également en tête du palmarès. L'année 2011 détient le plus d'occurrences (660 245 contre 657 393 en 2010). Cet écart s'explique par la répétition intensive de formes liées à l'événement de 2011. Mais, l'année 2010 a « détrôné » l'année 2011 par son nombre de formes (35 906), le plus élevé du sous-corpus, grâce à la richesse du vocabulaire lié aux divers événements de 2010.

Les deux « pics » constatés dans la figure B.2 correspondent à des événements majeurs survenus :

- L'année 2010 compte 1 140 articles : ceci se traduit entre autres par l'éruption volcanique de l'Eyjafjöll en Islande et l'explosion de la plate-forme « Deepwater Horizon» de la société BP dans le golfe du Mexique,
- L'année 2011 compte 1 124 articles : cela s'explique principalement par le séisme du 11 mars 2011 au large des côtes japonaises suivi d'un tsunami qui a provoqué l'accident nucléaire à Fukushima et par toutes les réactions sur le nucléaire à l'échelle mondiale.

Dans ce cas, ces constats nous amènent à la déduction suivante : le nombre élevé d'articles est en rapport direct avec la richesse de vocabulaire (nombre de formes).

Spécificités de la textométrie

Le calcul de spécificités est une méthode statistique permettant de mettre en évidence l'utilisation « atypique » d'une forme (ou plusieurs) dans une unité ou quantité textuelle donnée, par rapport à sa fréquence totale. Dans notre cas, l'empan est la période chronologique. Cette utilisation atypique d'une forme (ou plusieurs) se traduit par le « suremploi » quand la répétition s'intensifie localement, ou le « sous-emploi » lorsque cette forme est particulièrement moins utilisée de manière locale (Lafon, 1980, 1981 ; Lebart & Salem, 1994).

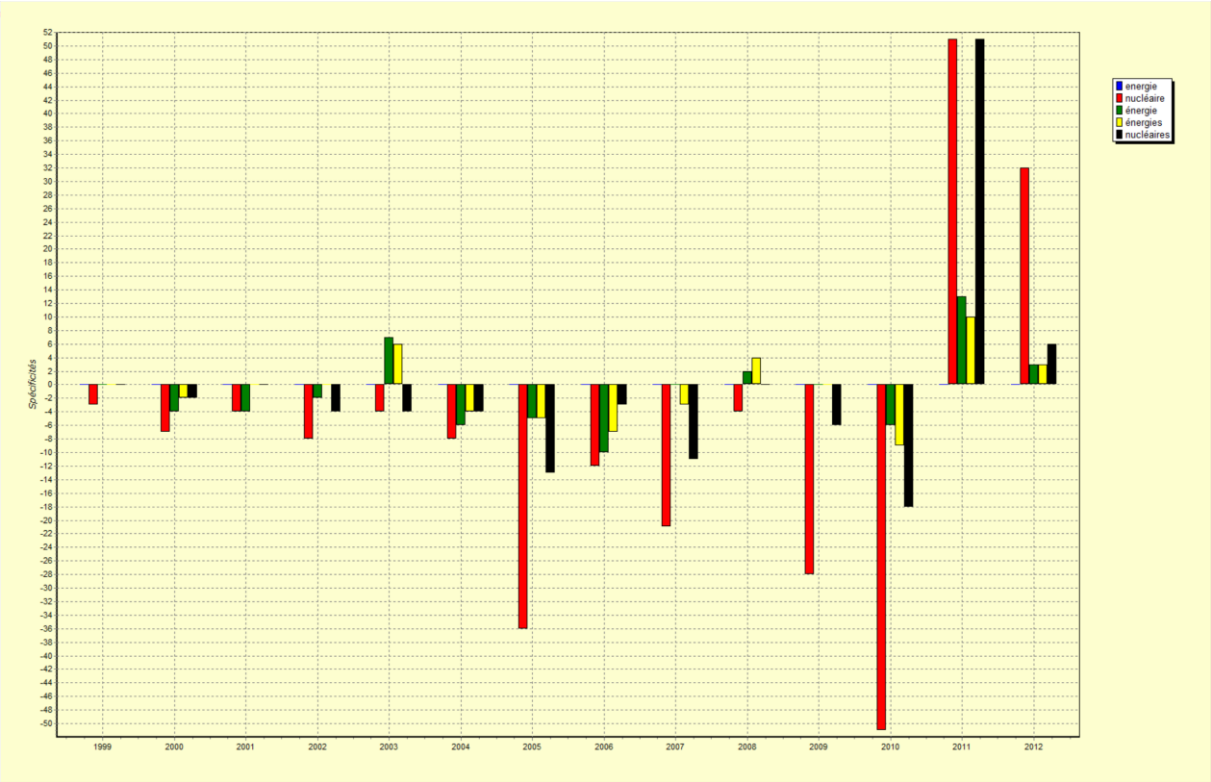


Figure B.3
ENRG_FR : exemples du calcul et de la ventilation de spécificités des formes *energie*, *nucléaire*, *énergie*, *énergies* et *nucléaires*

Dans la figure B.3, les formes «énergie » (vert), « énergies » (jaune) sont sur-employées en 2003, 2008, 2011 et 2012 ; « nucléaire » (rouge) et « nucléaires » (noir) le sont également en 2011 et 2012, tandis que « nucléaire » (rouge) est sous-employé entre 1999 et 2010 par rapport à sa fréquence totale.

Nous faisons appel au calcul de spécificités avec le seuil égal à 5 et la fréquence minimum égale à 10.

Dans les premières lignes du tableau B.2 ci-dessous, nous constatons que pour la période choisie (entre 2010 et 2011), les mots tels que Fukushima, nucléaire, BP, marée, noire, catastrophe, golfe, pétrole, Cancun etc. (colorés en vert) sont amplement répétés. Leurs fréquences présentes dans la période, occupent la quasi-totalité de leurs apparitions dans le sous-corpus. « Fukushima » est utilisé 526 fois dans cette période sur un total de 672 pour le sous-corpus total, et l'entité nommée « BP » a été répétée 835 pour un total de 867. Cela explique scientifiquement leurs spécificités d'utilisation par rapport au reste du texte.

Tableau B.2
ENRG_FR : Sélection des spécificités positives des années 2010 et 2011

Forme	Frq. Tot.	Fréquence	Coeff.
Fukushima	672	526	***
nucléaire	2672	1655	***
BP	867	835	***
marée	570	507	***
noire	679	570	***
catastrophe	797	590	***
golfe	391	341	***
pétrole	1187	828	***
Cancun	192	186	***
cendres	174	168	*47*

puits	525	405	*45*
aérien	299	247	*37*
Louisiane	165	154	*37*
Japon	816	561	*36*
Tepco	192	170	*34*
séisme	598	421	*31*
Deepwater	132	121	*28*
volcan	158	140	*28*
Horizon	133	121	*27*
radioactivité	233	180	*21*
bloqués	103	92	*20*
islandais	79	75	*20*
nuage	198	154	*19*
Islande	108	95	*19*
Oil	84	76	*17*
explosion	383	259	*16*
Eyjafjöll	47	47	*16*
éruption	161	124	*15*
(…)	(…)	(…)	(…)
pétrolière	229	164	*14*

Le tableau B.2 ci-dessus montrant le vocabulaire spécifique des années 2010 et 2011 témoigne d'un grand nombre de formes se rapportant à trois événements majeurs :

1. L'éruption du volcan islandais *Eyjafjöll* a débuté le 20 mars 2010, deux phases d'activités se sont succédées dont la deuxième plus explosive a projeté un nuage de cendres dans l'atmosphère bloquant les trafics aériens (*cendre*, *aérien*, *bloqués*, *islandais*, *nuage*, *Islande*, *Eyjafjöll*, *éruption*, etc. colorés en turquoise),
2. L'explosion de la plate-forme pétrolière «Deepwater Horizon» a provoqué une marée noire dans le golfe du Mexique (*BP*, *marée*, *noire*, *catastrophe*, *golfe*, *pétrole*, *Cancun* colorés en vert, *puits*, *Louisiane*, *Oil*, *explosion*, *pétrolière*, etc. colorés en jaune),
3. Le séisme sous la mer au large des côtes japonaises suivi d'un tsunami dévastateur ont été à l'origine de la catastrophe nucléaire de Fukushima (*Fukushima*, *nucléaire* colorés en vert, *Japon*, *Tepco*, *séisme*, *radioactivité*, etc. colorés en gris).

Typologie d'ensemble et études de la répartition des années du sous-corpus

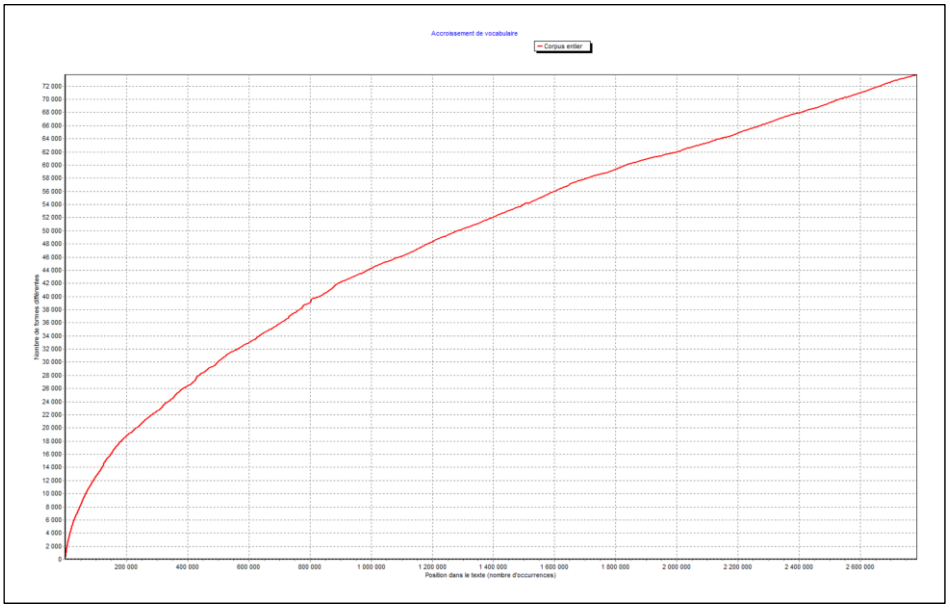


Figure B.4
ENRG_FR : accroissement de vocabulaire sur l'ensemble de la période (du mois de septembre 1999 au mois d'avril 2012)

Le diagramme d'accroissement de vocabulaire (figure B.4) permet d'observer l'apparition de nouvelles formes au fur et à mesure de l'avancement dans le sous-corpus français. L'ensemble du sous-corpus compte 2 783 702 occurrences et 76 848 de formes (tableau B.1 ci-dessus), le renouvellement de formes se stabilise après 1 000 000 d'occurrences. Par la suite, pour chaque intervalle de 200 000 occurrences supplémentaires, le nombre de formes augmente de 4 000 environ.

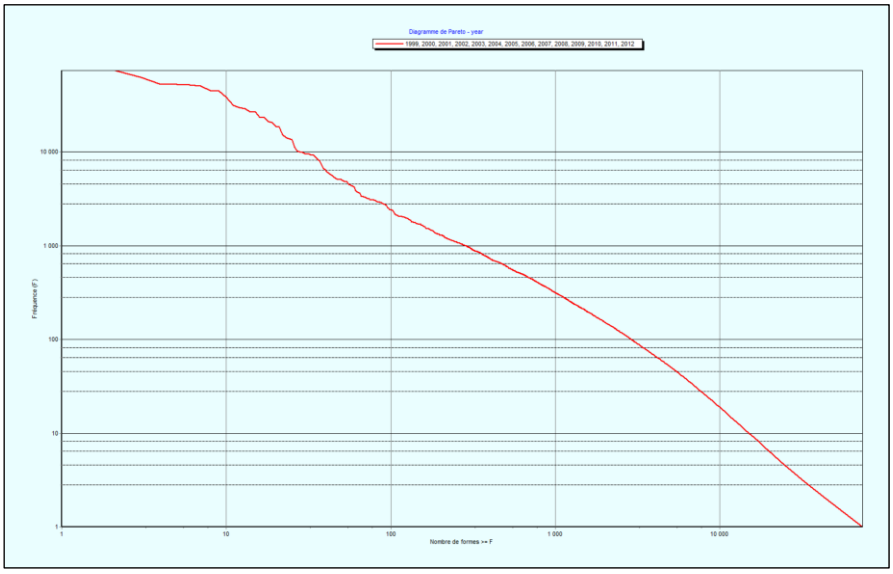


Figure B.5
ENRG_FR : diagramme de Pareto sur l'ensemble de la période

Le diagramme de Pareto (figure B.5) permet de visualiser la répartition du vocabulaire à la loi de Zipf. Ce diagramme est représenté approximativement sous la forme d'une droite, appelée Droite de Zipf (Zipf, 1932, 1935) :

- l'axe vertical représente la fréquence F des formes du sous-corpus français de 1999 à 2012, laquelle varie de 1 à Fmax, fréquence maximale calculée pour le texte T.
- l'axe horizontal porte le nombre de formes du sous-corpus dont la fréquence est supérieure à F.
- avant d'établir le diagramme, chacune de ces quantités est transformée en son logarithme décimal.

Cette loi n'est valable que pour des sous-corpus volumineux et est parfois traduite sous la forme simplifiée : Rang x fréquence = Constante.[230]
Ce diagramme prouve que la structure de la gamme (degré) des fréquences du vocabulaire du sous-corpus reste relativement stable.

[230] André Salem : « Tutoriels pour l'analyse textométrique », http://lexicometrica.univ-paris3.fr/numspeciaux/special8/tutoriel1.pdf (consulté le 01/03/2014)

Annexe B

Tableau B.3
ENRG_FR : sélection des spécificités chronologiques (1999-2012)

Formes/SR	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	
A	7	20	2	-3 15	-2 40	15	-2 100	-5 129	-9 169	-4 116	-4 713	**** 632	536	-8 259	-3
si	10	+3 18	6	35	+3 29	28	114	-3 180	-2 232	180	+3 345	617	661	297	
Etats	1	11	15	+3 48	+6 41	15	-2 201	+8 159	-4 254	+4 130	-3 459	+17 606	533	-8 235	-5
nucléaire	0	0	1	-4 3	-8 18	-4 3	-8 21	-36 112	-12 93	-21 116	-4 149	-28 175	**** 1484	**** 501	+33
avoir	1	13	1	-4 15	38	26	140	172	193	138	305	680	+4 602	267	
C	3	14	3	-2 34	+3 45	32	+3 135	144	-4 175	-2 135	317	661	+4 560	-3 284	
encore	3	8	7	27	41	17	92	-4 196	178	129	269	633	+3 623	272	
1	4	9	1	-3 16	29	13	-2 103	-3 160	217	+3 123	279	590	627	+3 278	
personnes	1	8	7	13	-2 31	17	123	185	112	-11 105	-4 395	+13 564	642	+4 215	-4
sans	8	8	18	+4 22	42	45	+7 129	178	200	142	297	531	-2 500	-5 292	+2
Pour	3	14	7	29	+2 25	17	141	+2 172	162	-3 142	286	582	615	+3 207	-5
effet	7	15	6	18	47	+3 17	130	187	320	+22 156	+3 299	427	-12 501	-4 232	-2
bien	4	16	9	28	35	22	110	159	187	153	+2 265	523	572	276	
leurs	1	16	7	22	25	15	114	122	-5 224	+5 156	+3 264	576	+2 481	-4 257	
déjà	2	4	-3 4	24	27	15	137	+3 174	192	128	248	525	546	220	-2
Ce	3	10	10	18	35	26	131	+3 144	184	107	265	526	473	-3 228	
rapport	0	4	-3 14	+3 5	-5 26	13	93	133	-3 296	+22 77	-6 230	501	519	241	
2	7	7	1	-3 16	34	19	86	-3 153	218	+6 81	-5 222	-2 475	564	+4 244	
Une	3	6	2	-2 17	35	14	96	148	134	-3 109	265	512	515	242	
avant	2	12	9	22	31	12	94	157	155	124	231	554	+4 457	-3 226	
Un	4	11	2	-2 13	24	15	70	-5 123	-3 136	-3 106	256	507	542	+3 265	+3
soit	0	10	5	19	21	17	98	149	156	112	241	492	499	231	
monde	3	12	4	15	37	20	116	150	179	118	261	442	-3 474	216	
donc	3	11	3	23	24	16	77	-4 131	-2 139	119	224	461	587	+8 223	
gaz	4	5	2	-2 0	20	0	124	+3 92	-8 284	+22 97	-2 199	-3 227	-48 666	+21 320	+11
Dans	2	3	-3 10	16	29	15	84	-3 134	133	-3 115	230	496	520	+3 245	
énergie	3	2	-4 0	10	-2 57	+7 2	-6 66	-5 86	-10 141	135	+2 250	396	-6 622	+13 262	+3
trois	3	10	7	12	32	6	-4 104	207	+7 141	96	-2 250	465	515	+2 175	-5
santé	1	9	5	13	33	38	+6 144	+5 179	+3 135	-2 118	490	**** 353	-12 348	-13 149	-9
années	5	20	+4 7	20	31	23	110	138	164	105	176	-6 437	-3 504	272	+5
fois	6	12	5	12	33	19	105	146	144	104	216	486	507	217	
environnement	1	3	-3 3	17	9	-5 5	-4 112	112	-4 236	+11 223	+22 192	-4 472	384	-7 225	
Unis	1	11	11	38	+6 37	+2 11	159	+8 106	-5 202	+5 100	328	+11 409	-4 396	-6 184	-3
euros	3	3	-3 0	22	16	-3 5	-4 53	-8 110	-4 135	107	204	-2 539	+5 473	295	+8

Le calcul des spécificités chronologiques met en évidence les faits les plus saillants et montre l'évolution de la répartition des formes retenues telles que, nucléaire (2010, 2011), énergie (2011) et environnement (2010, après la conférence de Copenhague) sur l'ensemble du sous-corpus de 1999 à 2012.

Pour réaliser une étude de « mapping » du sous-corpus français, nous allons recourir à la méthode « Analyse Factorielle des Correspondances ».

Annexe B

Typologie du sous-corpus français par année

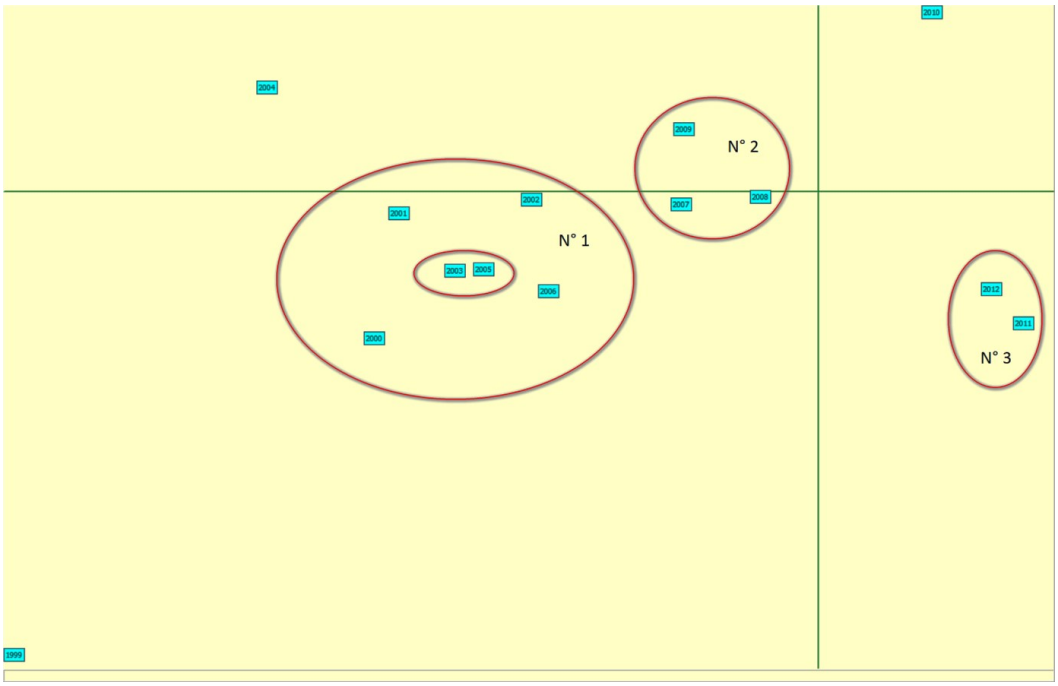


Figure B.6
ENRG_FR : analyse factorielle des correspondances sur l'ensemble de la période

Le graphique de la figure B.6 donne la représentation des proximités de chacune des années dans le plan factoriel du sous-corpus français, entre 1999 et 2012 de la rubrique « Planète » du journal le Monde.

D'une part, il y a une nette séparation entre les années 1999-2009, à gauche de l'axe vertical et les années 2010-2012, à droite de ce même axe.

D'autre part, nous voyons apparaitre approximativement trois groupements d'années et trois années extrêmes en rupture avec ces trois groupements. Cela signifie que chacun des trois groupements partage respectivement des formes communes dans l'utilisation de leur vocabulaire. Les périodes extrêmes (1999, 2004 et 2012) sont écartées.

Après l'observation de cette Analyse Factorielle des Correspondances, plusieurs constats en résultent :

3. Trois années extrêmes : 1999, 2004 et 2010 se détachent, s'écartent des groupements et deviennent par leurs vocabulaires spécifiques, les perturbateurs de la chronologie.
4. Des chronologies ont été plus ou moins conservées dans chacun de ces trois groupements, à savoir,
   - Groupement N° 1 : 2000, 2001, 2002, 2003, 2005 et 2006
     Un constat immédiat s'impose avec une proximité plus importante entre 2003 et 2005, mais 2004 est absente et devient le perturbateur de la chronologie.
   - Groupement N° 2 : 2007, 2008 et 2009
     L'année 2009 est en rupture avec les deux autres années, séparée par l'axe horizontal.
   - Groupement N° 3 : 2011 et 2012

Ce groupement reste dans une dimension individuelle à cause de leurs vocabulaires spécifiques.

Annexe B

Les périodes extrêmes : pourquoi cet éloignement ?

En 1999

Tableau B.4
ENRG_FR : sélection de spécificités de l'année 1999

Forme	Frq. Tot.	Fréquence	Coeff.
Lander	32	22	***
Mars	410	59	***
Polar	28	20	*50*
sonde	230	23	*33*
Orbiter	25	8	*17*
Propulsion	23	6	*13*
Surveyor	23	6	*13*
NASA	567	14	*12*
orbite	310	11	*12*
Lancement	14	5	*11*
martiennes	10	4	*10*
mission	555	12	*10*
(…)	(…)	(…)	(…)
Climate	40	5	*9*
martien	39	5	*9*
spatiale	482	10	*9*
Global	71	6	*9*
engin	79	6	*9*
missions	138	6	*8*
Planète	51	5	*8*
sondes	54	4	*7*
américaine	783	10	*7*
Planète	989	11	*7*

Comme nous le montrent les résultats du calcul de spécificités dans le tableau B.4, les formes colorées en vert, liées notamment aux domaines de l'aérospatiale, sont immédiatement mises en valeur. Au vu de ces premiers résultats, il est difficile de connaître les événements liés à ces formes. Pour déceler des renseignements et des informations, nous avons dû recourir au module « Concordance » de l'outil Lexico 3 qui visualise toutes les occurrences d'une forme en contexte. La concordance permet un « va-et-vient » systématique entre le texte et l'environnement immédiat de la forme.

La figure B.7, concordance du mot « Mars », répété 59 fois en 1999, montre que la fin de cette année 1999 est marquée par les événements aérospatiaux américains, en particulier, la perte de contact de la sonde « Mars Polar Lander » et la fin de mission de la sonde « Mars Climate Orbiter » orchestrées par le programme « Mars Surveyor 98 ». Ce dernier comprend deux satellites lancés séparément, le « Mars Climate Orbiter » et le « Mars Polar Lander ». Les deux missions avaient pour but d'étudier la météorologie, la climatologie, le cycle hydrologique et le dioxyde de carbone (CO2) de la Planète Mars, et entre autres de rechercher les changements climatiques épisodiques.

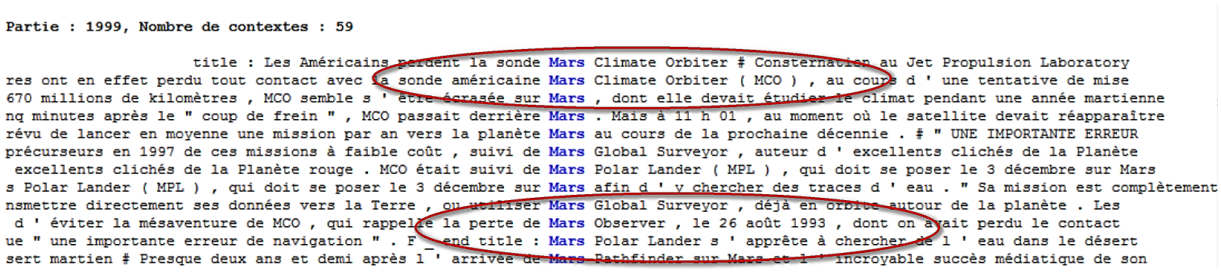


Figure B.7
ENRG_FR : concordance de la forme *Mars* en 1999

En 2004

Tableau B.5
ENRG_FR : spécificités de l'année 2004

Forme	Frq. Tot.	Fréquence	Coeff.
nosocomiales	75	58	***
infections	235	102	***
hôpital	305	41	*36*
nosocomiale	21	15	*27*
(…)	(…)	(…)	(…)
infection	221	29	*25*
hôpitaux	113	22	*24*
(…)	(…)	(…)	(…)

Les « infections nosocomiales » ont fortement marqué l'année 2004. En effet, le tableau B.5 met en évidence les événements en rapport avec les infections dans les hôpitaux, repérés par les formes, nosocomiales, infections, hôpital, nosocomiale, infection, hôpitaux (en vert). La figure B.8 ci-dessous, montre la concordance de ces deux termes les plus spécifiques qui ont été répétés 58 fois et en janvier uniquement, ce qui explique l'éloignement de cette année avec le reste des groupes.

Figure B.8
ENRG_FR : concordance de la forme *nosocomiales* pour le mois de janvier 2004

En 2010

Tableau B.6
ENRG_FR : spécificités de l'année 2010

Forme	Frq. Tot.	Fréquence	Coeff.
aérien	299	202	***
plate	303	201	***
golfe	391	288	***
aéroports	236	171	***
trafic	278	183	***
thon	293	216	***
vols	407	255	***
noire	680	493	***
BP	872	774	***
pétrole	1192	528	***
puits	525	338	***
marée	570	439	***
Mexique	673	345	***
Cancun	194	170	***
Louisiane	165	137	***
cendres	174	143	***
volcan	158	117	*41*
(…)	(…)	(…)	(…)
Deepwater	132	100	*37*
Horizon	133	100	*36*

Les deux événements révélés par le tableau B.6, qui nous semblent les plus marquants de 2010 concernent l'éruption du volcan islandais qui a projeté des cendres dans l'air, bloquant les trafics aériens et l'explosion de la plate-forme « Deepwater Horizon » dans le golfe du *Mexique*, comme déjà remarqué dans le tableau B.2 « Sélection des spécificités positives des années 2010 et 2011» (tableau plus haut). Deux groupes de formes ont été ainsi constitués avec les formes (colorées en vert) les plus saillantes suivantes par leurs fréquences dans cette période : *volcan*, *aérien*, *aéroports*, *trafic*, *vols*, *cendres*, appelé « Volcan 2010 » (figure B.9 ci-dessous) et *BP*, *noire*, *pétrole*, *puits*, *marée*, *Mexique*, *Cancun*, *Louisiane*, *Deepwater*, *Horizon*, nommé « Deepwater pétrole » (figure B.10 ci-dessous). Nous obtenons par la suite deux figures de concordances de ces deux groupes dans le sous-corpus, c'est-à-dire l'apparition simultanée de ces mots (figures B.11 et B.12).

Forme	Fréquence
volcan	158
aérien	299
aéroports	236
trafic	278
vols	407
cendres	174

Figure B.9
ENRG_FR : groupe de formes « Volcan 2010 » en 2010

Annexe B

Forme	Fréquence
BP	872
noire	680
pétrole	1192
puits	525
marée	570
Mexique	673
Cancun	194
Lousiane	4
Deepwater	132
Horizon	133

Figure B.10
ENRG_FR : groupe de formes « Deepwater pétrole » en 2010

```
Partie : 2010, Nombre de contextes : 1071

iques , cimenteries . . . - sont concernées . Le cas du transport aérien domestique est aussi posé . # L ' opération doit se faire sans alourdir
ons de la bande passante et qui nous permet de gérer les aléas du trafic , arrive en tête . " Le fait d ' externaliser est déjà un très bon
s en propre quelques serveurs ( 46 au total ) , le reste de notre trafic est géré par une société américaine , Akamaï , un " content delivery
 réseau mondial de serveurs qui nous permet de gérer les aléas du trafic , notamment lors de gros pics de connexions comme lors d ? un 11
ui on achète de la bande passante pour pouvoir gérer les aléas du trafic , qui arrive en tête . # Vient ensuite le matériel bureautique et
e , à Vitry - sur - Seine . Ils servent principalement à gérer le trafic abonnés , les blogs , les newsletters et les forums . Le reste de
s , les blogs , les newsletters et les forums . Le reste de notre trafic est supporté par une société américaine , Akamaï , un " content delivery
 réseau mondial de serveurs qui nous permet de gérer les aléas du trafic , notamment lors de gros pics de connexions comme lors d ' un 11
ction . En leur fournissant la répartition mois par mois de notre trafic par pays d ' origine pour 2009 , ils devraient être en mesure de
es Américains sur la Lune à l ' horizon 2020 et , au - delà , des vols habités vers Mars . # RÉDUIRE LA DÉPENDANCE AUX VAISSEAUX RUSSES
de 2009 en offrant un menu de plusieurs options pour le futur des vols spatiaux habités . L ' approche de l ' administration Obama viserait
is navettes de la flotte , en principe en septembre 2010 . # Cinq vols sont encore prévus , tous vers l ' avant - poste orbital , le prochain
, sur les 18 milliards de son budget annuel , à ses programmes de vols habités , à savoir actuellement la navette et le développement de
ter des astronautes , dit Daniel Sacotte , ancien responsable des vols habités à l ' Agence spatiale européenne . Seuls les gros du secteur
serve la maîtrise intégrale des missions les plus délicates , les vols inauguraux vers Mars ou ailleurs , l ' agence ne pourra plus justifier
isiter les classiques des missions Apollo et les fondamentaux des vols soviétiques . Avec cependant une manière bien à eux de ne pas afficher
la fin des années 1990 , ils disaient avancer prudemment vers les vols habités , ils montraient quelques vagues études pour un avenir lointain
de découvrir tout cela . Le programme Shenzhou , au bout de trois vols habités , lui a permis de réussir la première sortie d ' un cosmonaute
nouveau . Son agence spatiale ( ISRO ) ne maîtrise pas encore les vols habités , et elle n ' a , pour l ' heure , envoyé qu ' une sonde
rs de particules . Les particules fines , émises notamment par le trafic automobile , pénètrent profondément dans les poumons et sont particulièrement
GAC ) a demandé aux compagnies aériennes de réduire de 20 % leurs vols à l ' aéroport de Roissy - Charles de Gaulle mercredi entre 11 heures
ion européenne prépare l ' élargissement du dispositif au secteur aérien pour 2012 , et s ' apprête à franchir une étape décisive en 2013
 la Guadeloupe resteront fermés vendredi , à cause de la pluie de cendres qui a commencé à s ' y abattre jeudi soir après l ' explosion du
 oculaire et des voies respiratoires que pourraient provoquer les cendres volcaniques . Il a également appelé les automobilistes à la prudence
a prudence , les routes étant rendues glissantes par la couche de cendres qui s ' y dépose . L ' île de Montserrat , voisine de la Guadeloupe
aient recouvertes jeudi soir par une mince couche gris - blanc de cendres volcaniques . Les rares voitures qui circulent soulevaient un nuage
jeudi , à titre préventif , à la suite des premières retombées de cendres provoquées par l ' explosion , plus tôt dans l ' après - midi , du
du dôme volcanique de La Soufrière de Montserrat . La plupart des vols à destination de Pointe - à - Pitre doivent être déroutés sur Fort
 de la Guadeloupe . # Essentiellement constituées de silice , les cendres , émises suite à l ' explosion du dôme volcanique de La Soufrière
rbourg où l ' attendait le même navire russe . F _ end title : Le trafic de déchets toxiques gagne toute l ' Italie # A ce prix - là - 150
nts . # Les entreprises concernées nient avoir eu connaissance du trafic . Les enquêteurs s ' étonnent , eux , qu ' elles aient accepté sans
e haut sur une base de 3 hectares . Le chiffre d ' affaires de ce trafic largement géré par la Mafia atteindrait 7 milliards d ' euros . Limité
irectement sur la route de l ' aéroport . Il est fermé . Tous les vols sont annulés pour au moins 24 heures . Plus tard , j ' apprendrai
( 12 h 25 en France ) , a indiqué le centre russe de contrôle des vols ( Tsoup ) dans un communiqué . # Toutes les opérations se sont déroulées
us allez en mer , vous trouvez du thon . Si vous faites du survol aérien , vous trouvez des quantités plus grandes qu ' il y a dix ans . Mais
servateurs , le commerce international est le principal moteur du trafic . " A partir du moment où il existe un marché , on encourage le braconnage
protection des animaux ( IFAW ) , inquiète de la recrudescence du trafic d ' ivoire : 6 , 2 tonnes saisies au Vietnam en mars 2009 , 3 , 3
Seattle , qui incrimine l ' implication du crime organisé dans le trafic d ' ivoire . " Les grosses saisies d ' ivoire sont de plus en plus
 médecine nucléaire 1 , 6 % . F _ end title : L ' éruption d ' un volcan islandais s ' intensifie # L ' éruption volcanique dans le sud de
es inondations consécutives à la fonte de la glace qui entoure ce volcan . # " La police a accru la surveillance dans toute la zone autour
é plusieurs routes et averti du danger de circuler à proximité du volcan , où de nouveaux petits séismes ont été mesurés lundi matin . # La
0 kilomètres de la capitale Reykjavik et qui surmonte le puissant volcan Katla , rapporte également la télévision publique RUV . La dernière
e , à Vitry - sur - Seine . Ils servent principalement à gérer le trafic abonnés , les blogs , les newsletters et les forums . Le reste de
s , les blogs , les newsletters et les forums . Le reste de notre trafic est supporté par une société américaine , Akamaï , un " content delivery
 réseau mondial de serveurs qui nous permet de gérer les aléas du trafic , notamment lors de gros pics de connexions comme lors d ' un 11
ction . En leur fournissant la répartition mois par mois de notre trafic par pays d ' origine pour 2009 , ils devraient être en mesure de
tants ont même trouvé des poissons dans leurs maisons . Plusieurs vols entre Rio et Sao Paulo ont été annulés , l ' aéroport intérieur de
s . Mais d ' ici là l ' avion devra encore effectuer une série de vols d ' essais et notamment , vers l ' été , un périple de trente - six
es habitants évacués d ' une zone menacée par l ' éruption d ' un volcan en Islande # Plusieurs centaines de personnes ont été évacuées ,
```

Figure B.11
ENRG_FR : concordance du groupe « Volcan 2010 » en 2010

Partie : 2010, Nombre de contextes : 3290

ons sont en cours avec notamment l ' Ukraine et le **Mexique** , mais Roselyne Bachelot a dit douter que les ventes
(Cambodge) , le lac Baïkal (Russie) , Oaxaca (**Mexique**) , le Sud de l ' Afrique , le district des lacs (
lié selon l ' Agence à l ' augmentation du prix du **pétrole** et à l ' effet Grenelle , est d ' emblée mis en avant
pèdes non aviaires étaient probablement de couleur **noire** et brun roux , selon l ' étude . Les zones où aucun
boom " est lié à une forte augmentation du prix du **pétrole** en 2008 tout comme à un effet Grenelle . Entretien
la même région , la population de gibbons à crête **noire** (Nomascus nasutus) est limitée à environ 110 individus
ne sont pas égaux : " Le Brésil , la Colombie , le **Mexique** , le Maroc et l ' Afrique du Sud ont le plus grand
andes commandes de vaccins . F _ end title : Marée **noire** en Italie après le sabotage d ' une raffinerie # Un
raîné mardi 23 février le déversement de tonnes de **pétrole** dans le Lambro , un affluent du Pô , provoquant une
ans le Lambro , un affluent du Pô , provoquant une **marée** noire qui a atteint en milieu de journée le plus grand
Lambro , un affluent du Pô , provoquant une marée **noire** qui a atteint en milieu de journée le plus grand fleuve
iseaux , sont déjà morts , englués par la nappe de **pétrole** à l ' odeur nauséabonde , longue de plusieurs kilomètres
er d ' utiliser l ' eau courante . # " La nappe de **pétrole** est partie de Monza , a traversé la région de Milan
ment (ARPA) , Monia Maccarini . " La quantité de **pétrole** déversée s ' élève à au moins 1 000 m3 , mais elle
. La coordination des opérations pour endiguer la **marée** noire a été confiée à la protection civile et une
coordination des opérations pour endiguer la marée **noire** a été confiée à la protection civile et une unité
pour élever des digues et entraver l ' avancée du **pétrole** ont déjà été effectuées , mais jusqu ' à présent sans
éside dans sa conjonction à un fort coefficient de **marée** sur la côte atlantique , au moment même de la marée
arée sur la côte atlantique , au moment même de la **marée** haute . Ce sont ces trois phénomènes naturels réunis
trière , c ' est précisément à cause des effets de **marée** et de houle longue qui se sont conjugués aux hautes
ée et 160 km / h en Charente - Maritime , avec une **marée** qui a connu une surcote importante , a fait céder
ur chiffrer ces travaux . F _ end title : " Raz de **marée** meurtrier " , " le pire bilan depuis 1999 " selon
ffet combiné du vent et d ' un fort coefficient de **marée** (102) , ont prévenu mercredi les autorités préfectorales
t de nord - est de 60 à 80 km / h , conjugué à une **marée** élevée . Le risque a été identifié sur les côtes ouest
' ici à la prochaine conférence sur le climat , à **Cancun** (Mexique) , en fin d ' année . # Le chef de l '
la prochaine conférence sur le climat , à Cancun (**Mexique**) , en fin d ' année . # Le chef de l ' Etat a par
ution et de rotation sont égales . # Les effets de **marée** , à cause de la non uniformité de la force gravitationnelle
ite et lui donnent une forme aplatie . La force de **marée** agit sur le bourrelet formé et applique un couple
e la préservation économique de deux villages , le **Mexique** a empêché des mesures imposant le quota zéro pêche
a contamination des maïs des Indiens zapotèques du **Mexique** par des organismes génétiquement modifiés (OGM)
Ce texte a été conçu pour que les responsables de **marée** noire n ' aient jamais rien à payer , sauf s ' ils
xte a été conçu pour que les responsables de marée **noire** n ' aient jamais rien à payer , sauf s ' ils ont intentionnellement
ermettre de savoir de quoi est composée la matière **noire** qui représente 23 % de notre Univers , contre 4 %
les planètes . Les 73 % restants sont l ' énergie **noire** , ou force d ' expansion de l ' Univers . # Pour en
. # De quoi sont faites , par exemple , la matière **noire** et l ' énergie sombre qui forment 96 % de l ' Univers
10 km / h , associés à un important coefficient de **marée** (110 à 112) , pourraient engendrer sur le littoral
nn , avec , en ligne de mire , la rendez - vous de **Cancun** , au Mexique , en fin d ' année . " En fait , les
en ligne de mire , la rendez - vous de Cancun , au **Mexique** , en fin d ' année . " En fait , les négociations
te maisons ont été gravement touchées par la folle **marée** de la nuit du 27 au 28 février . Leur sort va dépendre
er les gens aujourd ' hui . " Redoutée , la grande **marée** du mardi 30 mars n ' a pas causé plus de dégâts .
fendre . F _ end title : (?) , combien compte le **pétrole** . | Oil Man # Is " Big Oil " also " too big to fail
qu ' il est possible que la production mondiale de **pétrole** et de ses substituts décline à partir de l ' an prochain
é technologique ne sait pas se passer , c ' est le **pétrole** . # D ' après l ' enquête publiée ici le 23 mars [
s émissions de CO2 induites par notre addiction au **pétrole** , ndlr) . # J ' ai dit six . . . Ah , oui ! Et bien
on esprit : 1 - la production mondiale maximale de **pétrole** aurait - elle pu être atteinte par mégarde , M . Sweetnam
isoire d ' investissements . D ' autres experts du **pétrole** (diable , ce qu ' ils sont nombreux !) clament que
' un demi - siècle . # Ceci revient à dire que le **pétrole** risque quoi qu ' il arrive de devenir plus rare et
ut peut - être pas la chandelle . . . # Combien le **pétrole** compte Bien sûr , le pétrole est la ressource nodale

Figure B.12
ENRG_FR : concordance du groupe « Deepwater pétrole » en 2010

Selon le tableau B.6 et les figures B.11 et B.12, 2010 est une année marquée par des catastrophes. Nous en avons retenu deux et défini deux groupes de formes « Volcan 2010» et «Deepwater pétrole». Les nombres de contextes de ces deux groupes sont respectivement 1 071 et 3 290, en 2010 uniquement, alors que la figure B.13 ci-dessous, nous retrace leurs fréquences absolues dans le sous-corpus entier et prouve leurs distributions saillantes par rapport au reste du sous-corpus.

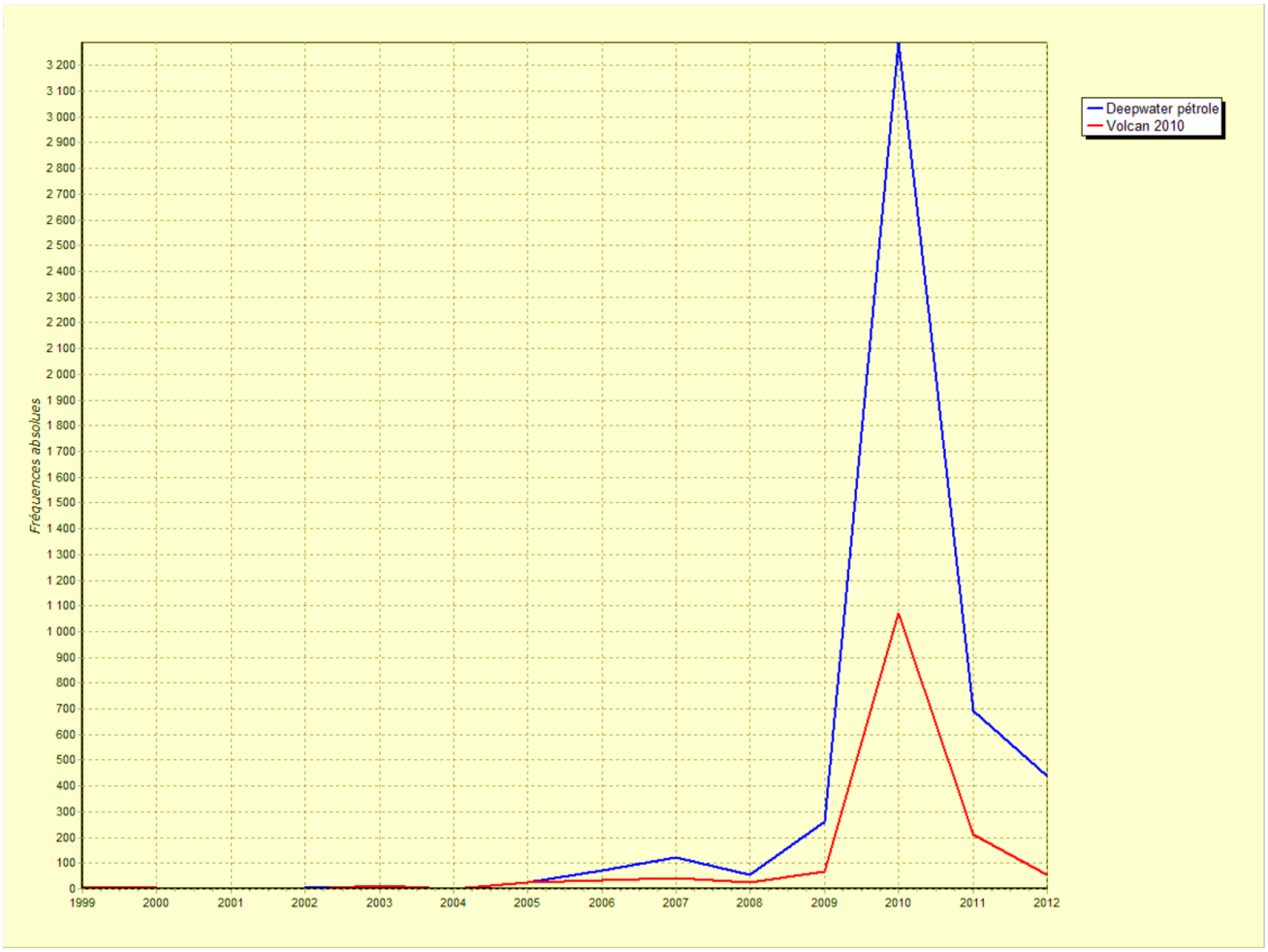


Figure B.13
ENRG_FR : ventilation des fréquences absolues des groupes « Deepwater pétrole » et « Volcan 2010 » en 2010

Les trois groupements d'années

Groupement N° 1 : 2000, 2001, 2002, 2003, 2005 et 2006 avec une proximité importante entre 2003 et 2005.

Tableau B.7
ENRG_FR : sélection de spécificités des années 2000, 2001, 2002, 2003, 2005 et 2006

Forme	Frq. Tot.	Fréquence	Coeff.
H5N1	323	237	***
SRAS	92	81	***
volailles	297	236	***
grippe	1643	513	***
aviaire	577	438	***
oiseaux	444	189	*42*
migrateurs	58	53	*38*
(…)	(…)	(…)	(…)

L'analyse du tableau B.7 nous relève la spécificité de ce groupement qui se traduit par de fortes fréquences de formes, par exemple, *aviaire*, *grippe*, *SRAS*, *volailles*, *H5N1*, *migrateurs*, *oiseaux*, etc. (en vert), formes relatives à l'épidémie de grippe aviaire, qui a sévi dans ces années.

Annexe B

Groupement N° 2 : 2007, 2008 et 2009

Tableau B.8
ENRG_FR : spécificités des années 2007, 2008 et 2009

Forme	Frq. Tot.	Fréquence	Coeff.
H1N1	477	411	***
réchauffement	1113	523	***
porcine	241	236	***
grippe	1643	981	***
climatique	1780	765	***
émissions	1787	794	***
climat	1083	459	*36*
serre	1052	444	*34*
Grippe	113	89	*33*
Copenhague	441	212	*26*
biocarburants	158	97	*22*
épidémie	619	264	*22*
CO2	1139	428	*21*
pandémie	368	176	*21*
(…)	(…)	(…)	*(...)*

Dans la chronologie de 2007, 2008 et 2009, le tableau B.8 présente les mots marquants relatifs aux événements tels que la nouvelle pandémie de grippe porcine (*H1N1*, *porcine*, *grippe*, *Grippe*, *épidémie* et *pandémie* colorés en vert) ainsi que les sujets autour de la Conférence de Copenhague sur le climat (*réchauffement*, *climatique*, *émission*, *climat*, *serre*, *Copenhague*, *biocarburant* et *CO2* colorés en turquoise). Le vocabulaire spécifique de ce dernier tableau explique bel et bien la raison pour laquelle l'année 2009 est en rupture avec 2007 et 2008.

Groupement N° 3 : 2011 et 2012

Tableau B.9
ENRG_FR : spécificités des années 2011 et 2012

Forme	Frq. Tot.	Fréquence	Coeff.
centrale	1138	877	***
séisme	598	406	***
réacteurs	609	520	***
réacteur	619	471	***
Tepco	192	192	***
sûreté	494	387	***
ASN	354	275	***
tsunami	311	248	***
nucléaire	2676	1985	***
centrales	760	558	***
Japon	817	506	***
Allemagne	840	509	***
Fukushima	672	668	***
nucléaires	683	474	***
radioactivité	233	190	*49*
Fessenheim	161	144	*48*
EELV	70	68	*29*
Hollande	95	85	*29*

tremblement	162	124	*28*
Tchernobyl	254	170	*26*
(…)	(…)	(…)	*(...)*

D'après le tableau B.9, 2011 et 2012 forment une suite de deux années marquées par le *nucléaire*, à cause d'un enchaînement de catastrophes dont le point de départ a été un *séisme* au large des côtes du *Japon*, suivi d'un *tsunami* provoquant la destruction de la centrale nucléaire à *Fukushima* (*centrale*, *séisme*, *réacteurs*, *réacteur*, *Tepco*, *tsunami*, *Japon*, *Fukushima*, *radioactivité*, *tremblement*, etc. colorés en vert). A ces événements et ces moments de panique qui saisissent toute la Planète, s'ensuit notamment une prise de conscience de l'ensemble de la classe politique et des citoyens sur le nucléaire et sa sûreté à l'échelle mondiale (*sûreté*, *ASN*, *nucléaire*, *centrale*, *Allemagne*, *nucléaires*, *centrales*, *Fessenheim*, *EELV*, *Hollande*, *Tchernobyl*, etc. colorés en turquoise).

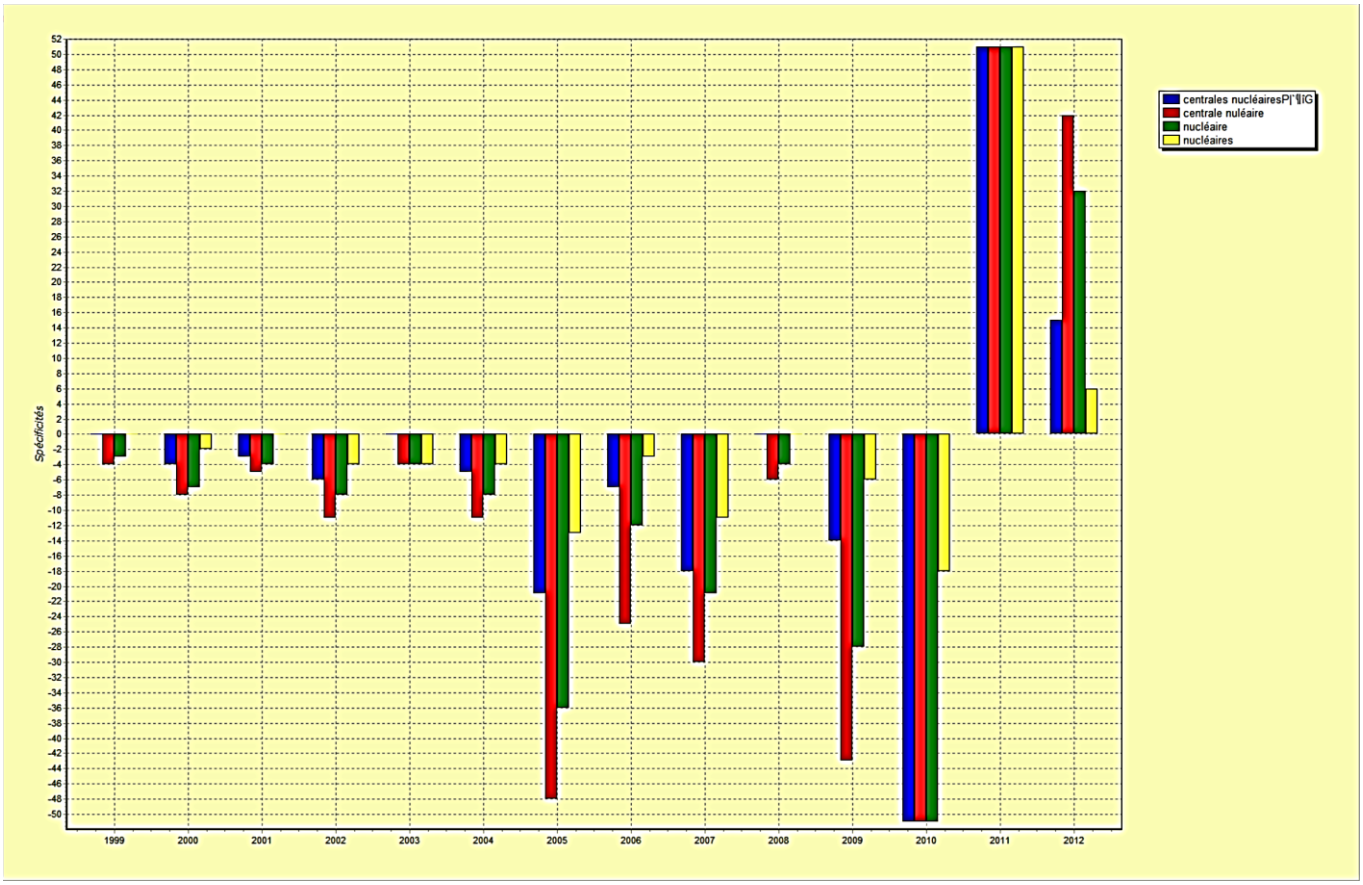


Figure B.14
ENRG_FR : ventilation des spécificités
des groupes «centrale, nucléaire », «centrales, nucléaires» et «nucléaire(s)»

La figure B.14 ci-dessus basée sur la ventilation de deux groupes de formes, *centrale*, *nucléaire* et *centrales*, *nucléaires* ainsi que les deux formes du mot *nucléaire* (au singulier ou au pluriel), nous laisse indiscutablement des preuves tangibles sur la spécificité des années 2011 et 2012.

Similitudes et différences des trois groupements
Similitudes
Les événements autour des grippes animales ont conduit le rapprochement relatif des Groupements N° 1 et N° 2 avec l'année 2009 en rupture avec le reste des groupements.
Différences
Les événements provoqués par l'accident nucléaire de Fukushima ont totalement basculé le Groupement N° 3 (2011 et 2012) dans une dimension individuelle.

Constat général du sous-corpus français

Nous venons de présenter un aperçu général du sous-corpus français du corpus comparable ENRG. Les analyses de spécificités au cours de cette présentation nous révèlent que les événements internationaux dépassent largement les événements nationaux, mis à part quelques faits saillants de la société française (figure B.14) :

1. Les infections nosocomiales dans les hôpitaux français en 2004,
2. L'affaire de l'achat de vaccins contre la grippe H1N1 opérée sous la direction de Madame Roselyne Bachelot en 2009,
3. La doyenne des centrales françaises « Fessenheim » provoque de vives réactions publiques et politiques en France suite à la catastrophe de Fukushima.

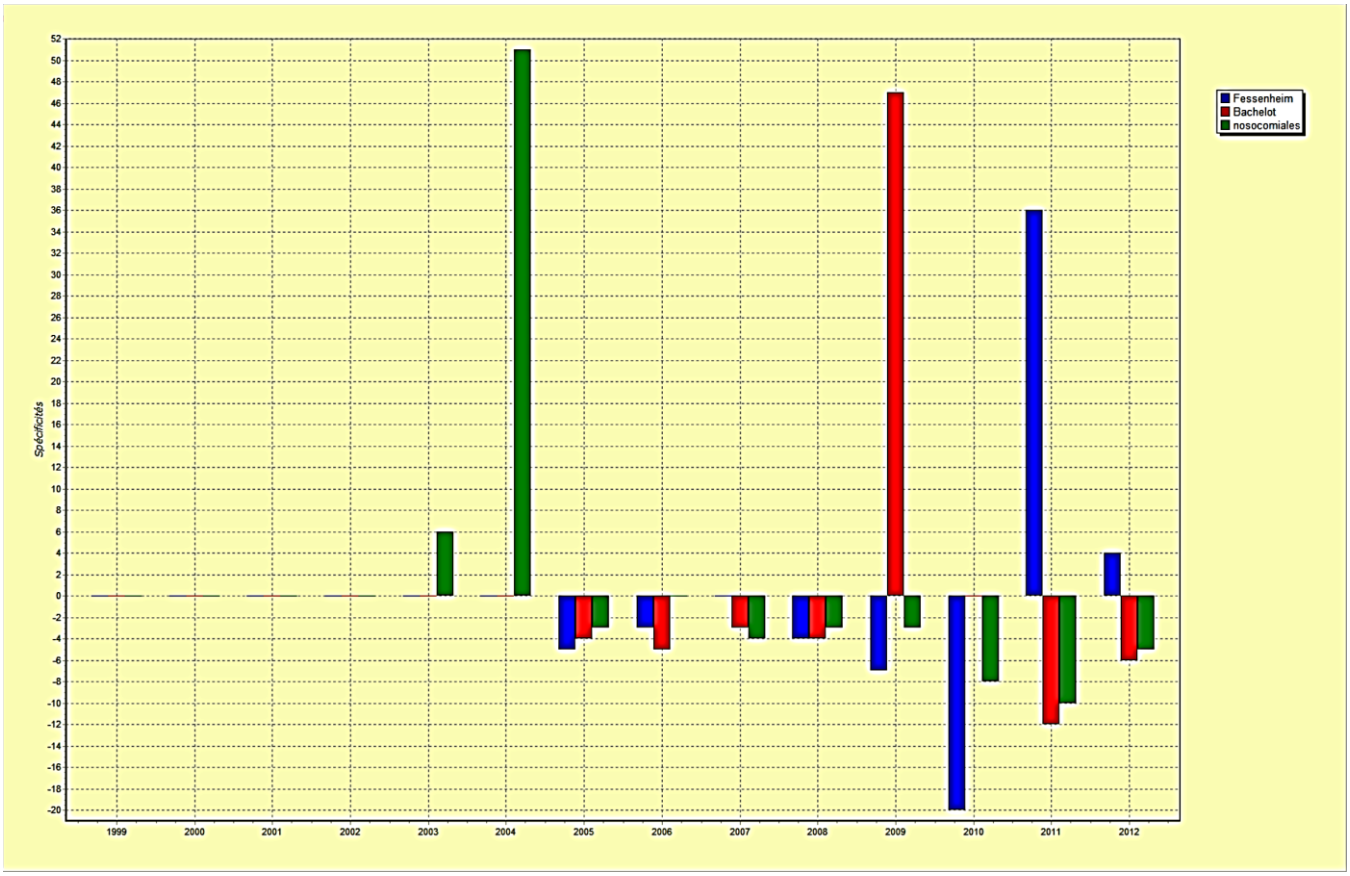


Figure B.15
ENRG_FR : ventilation des spécificités
des formes *Fessenheim*, *nosocomiales* et *Bachelot*

Nous sommes dans la rubrique « Planète », c'est pourquoi les articles relatifs aux questions internationales restent dominants.

B.2 Sous-corpus américain, journal *New York Times*

Le sous-corpus anglais américain (*American English*) est constitué à partir d'articles agrégés par la rubrique « *Environment* » du New York Times, datés du 26-01-2005 au 18-04-2012, la taille du fichier en .txt est de 15,8 Mo.

Pourquoi cette période ?
La date de début correspond au premier article disponible de cette rubrique du journal en ligne et la date de fin est celle du dernier lancement de la requête informatique.

Principales caractéristiques textométriques du sous-corpus américain

Tableau B.10
ENRG_US : principales caractéristiques textométriques par année

Nombre d'occurrences:	2735535	Nombre de formes:	66966
Nombre d'hapax:	25021	Fréquence maximale:	153517

	Num	Partie	Occurenc	Formes	Hapax	Fmax	Forme
✓	1	2005	325464	23582	10159	19075	the
✓	2	2006	396373	26923	11458	22119	the
✓	3	2007	563921	32618	13756	31034	the
✓	4	2008	493213	30458	12992	27236	the
✓	5	2009	345016	24447	10573	18788	the
✓	6	2010	296343	21900	9564	17469	the
✓	7	2011	271929	22020	9864	15344	the
✓	8	2012	43276	7988	4218	2452	the

Le tableau B.10 présente les principales caractéristiques textométriques de l'ensemble du sous-corpus : **2 735 535 occurrences** et 1**89 936 formes**. Ce sous-corpus se divise en **59 452 paragraphes**. La forme la plus fréquente est l'article «the». Seule l'année **2012 est une période incomplète**. L'année civile complète où il y a le plus de mots est l'année **2007 avec 563 921 occurrences** et celle qui en comporte le moins est l'année **2011 avec 271 929**.

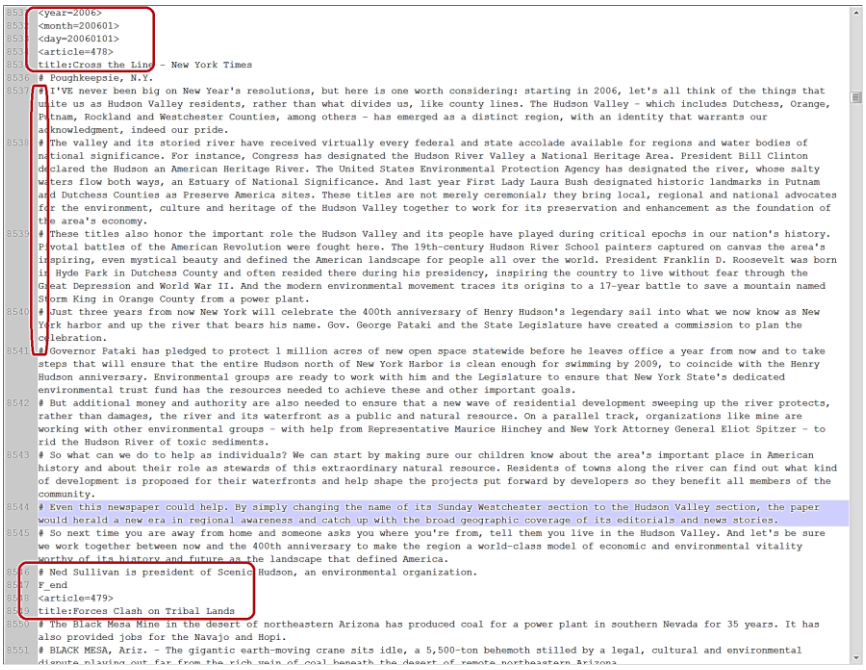


Figure B.16
ENRG_US : structure du sous-corpus

Ce sous-corpus se divise en différents empans textuels suivant une chronologie. Les articles sont triés par ordre croissant sur leurs dates de publication (aaaammjj), regroupés par jour, par mois, par année, à l'aide de programmes informatiques spécifiques. Chaque balise désigne le début d'un empan textuel (figure B.16) :

- <year=aaaa> : balise indiquant l'année des articles. Le sous-corpus américain s'étale sur 8 années, de 2005 à 2012.
- <month=aaaamm> : balise indiquant le mois des articles.
- <day=aaaammjj> : balise indiquant le jour des articles, sachant que dans une journée, plusieurs articles sont susceptibles d'être regroupés sous cette balise. L'ensemble du sous-corpus contient 1 908 jours.
- <article=nnnn> : balise indiquant le numéro de chaque article.
- Le titre de chaque article est introduit par la chaîne de caractères « title :».
- Le signe « # » marque le début de chaque paragraphe de chaque article. Les textes rassemblés sont au nombre de 3 993 articles.
- La fin de chaque article est marquée par la chaîne de caractères « F_end ».

L'évolution du nombre d'articles du sous-corpus américain

La figure B.17 ci-dessous montre l'évolution du nombre d'articles extraits du journal en ligne constituant le sous-corpus français du corpus comparable.

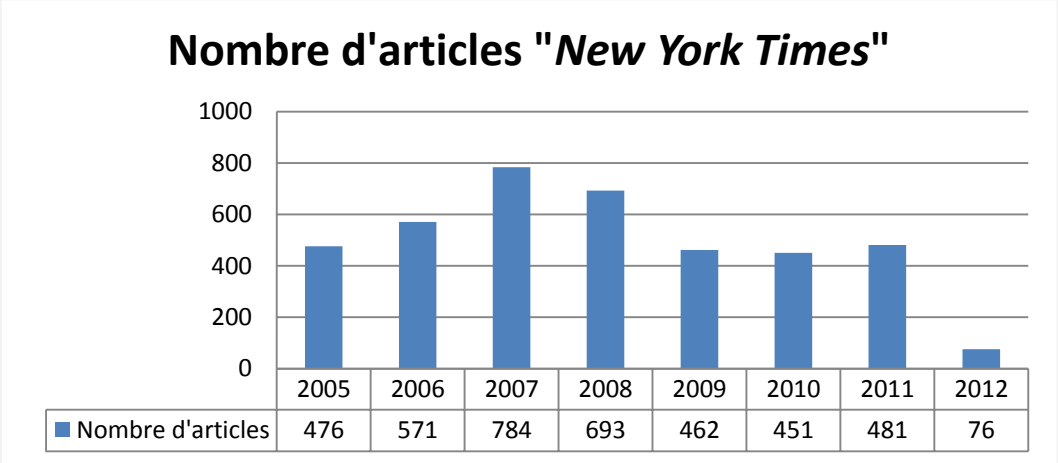


Figure B.17
ENRG_US : évolution du nombre d'articles pour la rubrique *Environment* du mois de janvier 2005 au mois d'avril 2012

Au vu de ces données (figure B.17 ci-dessus) et celles du sous-corpus français (tableau B.1 plus haut), une première remarque : « *New York Times* » (ci-après NYT) consacre moins d'articles à l'environnement que « Le Monde » sur le nombre total d'articles de la période comparable. Ceci est dû probablement au fait que les deux rubriques ne couvrent pas le même périmètre et que le thème « *Environment* » est une sous-catégorie de la rubrique « Science » dans le NYT.

Les nombres d'occurrences et de formes sont également inférieurs au sous-corpus français.

D'après la figure B.17, le nombre d'articles sur la période de 2005 à 2011 (l'année 2012 étant incomplète) témoigne d'une certaine homogénéité. Les années 2007 (784 articles) et 2008 (693

articles) se démarquent des autres années par un nombre d'articles supérieur d'environ 42% (pour 2007) et 34% (pour 2008) par rapport à celui de l'année 2010 (année où il y a le moins d'articles).

L'année 2007 occupe la première place que ce soit par son nombre d'articles ou par son nombre d'occurrences ou par son nombre de formes. Ceci corrobore la même déduction au sujet du rapport entre le nombre d'articles et le nombre de formes lors des analyses du sous-corpus français, *i.e.* le nombre d'articles est en rapport direct avec le nombre de formes.
Dans ce sous-corpus, l'interprétation des résultats est plus complexe que pour le sous-corpus français. Pour ce faire, nous allons recourir à d'autres modules de la textométrie.

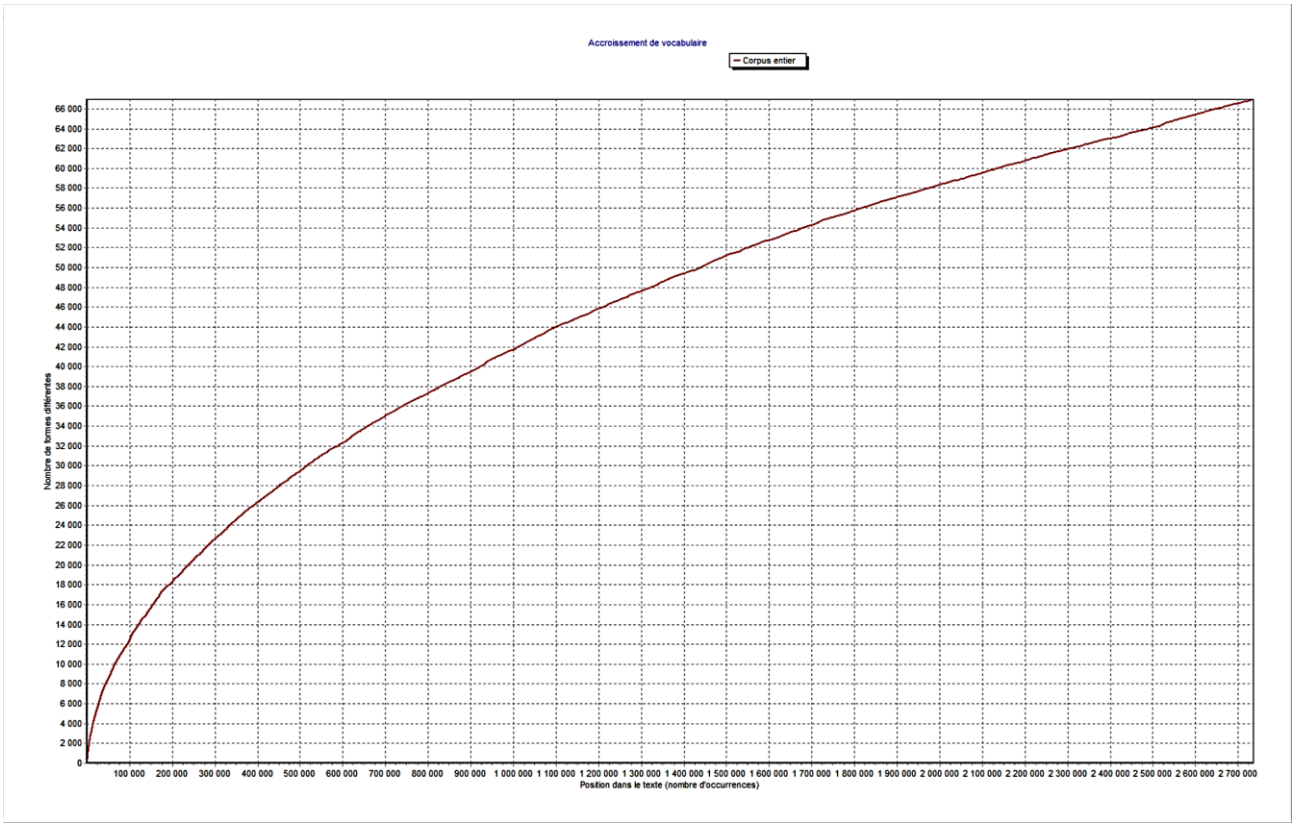


Figure B.18
ENRG_US : accroissement de vocabulaire de l'ensemble de la période (du mois de janvier 2005 au mois d'avril 2012)

Le diagramme d'accroissement de vocabulaire (figure B.18) ci-dessus, montre, comme le sous-corpus français, l'apparition de nouvelles formes au fur et à mesure de l'avancement dans le temps. Rappelons que l'ensemble du sous-corpus compte 2 735 535 occurrences et 189 936 de formes (tableau B.10), le renouvellement de formes se stabilise après 1 000 000 occurrences. Par la suite, pour chaque tranche de 100 000 occurrences supplémentaires, le nombre de formes augmente de 2 000 environ.

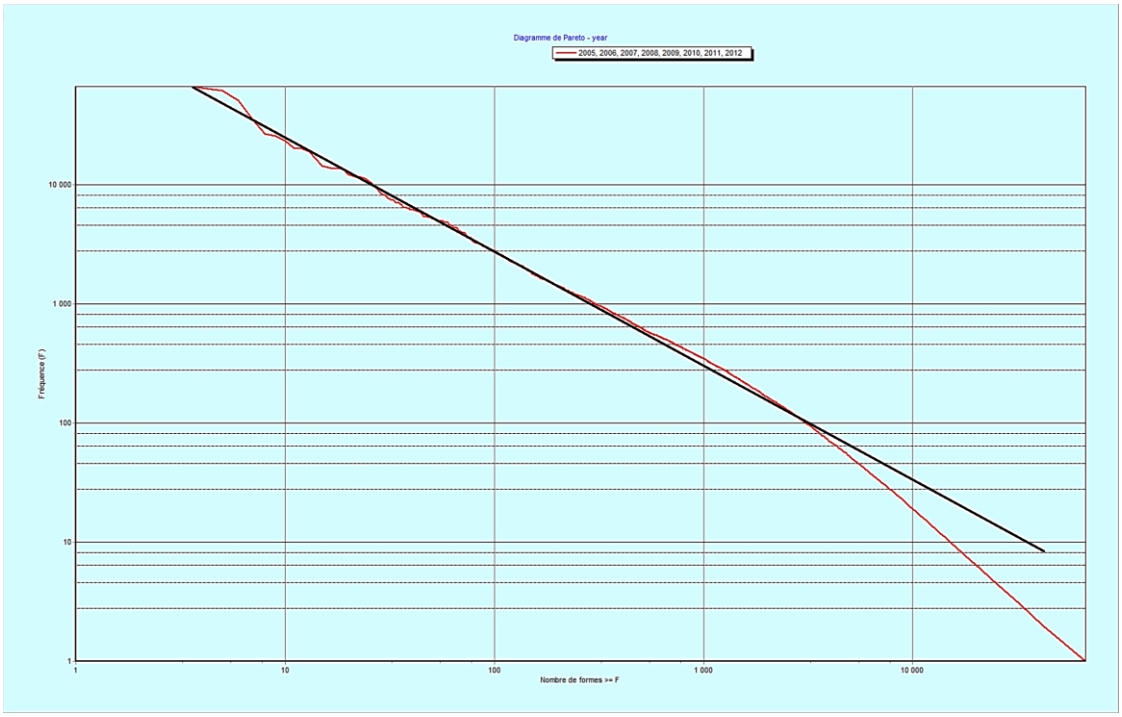


Figure B.19
ENRG_US : diagramme de Pareto de l'ensemble de la période

Le diagramme de Pareto ci-dessus (figure B.19) prouve que la structure de la gamme (degré) des fréquences du vocabulaire du sous-corpus reste relativement stable. C'est-à-dire, le renouvellement du vocabulaire dans le NYT est assez stable dans l'avancement du sous-corpus et le sous-corpus américain renouvelle plus de formes que le sous-corpus français au fil du temps.

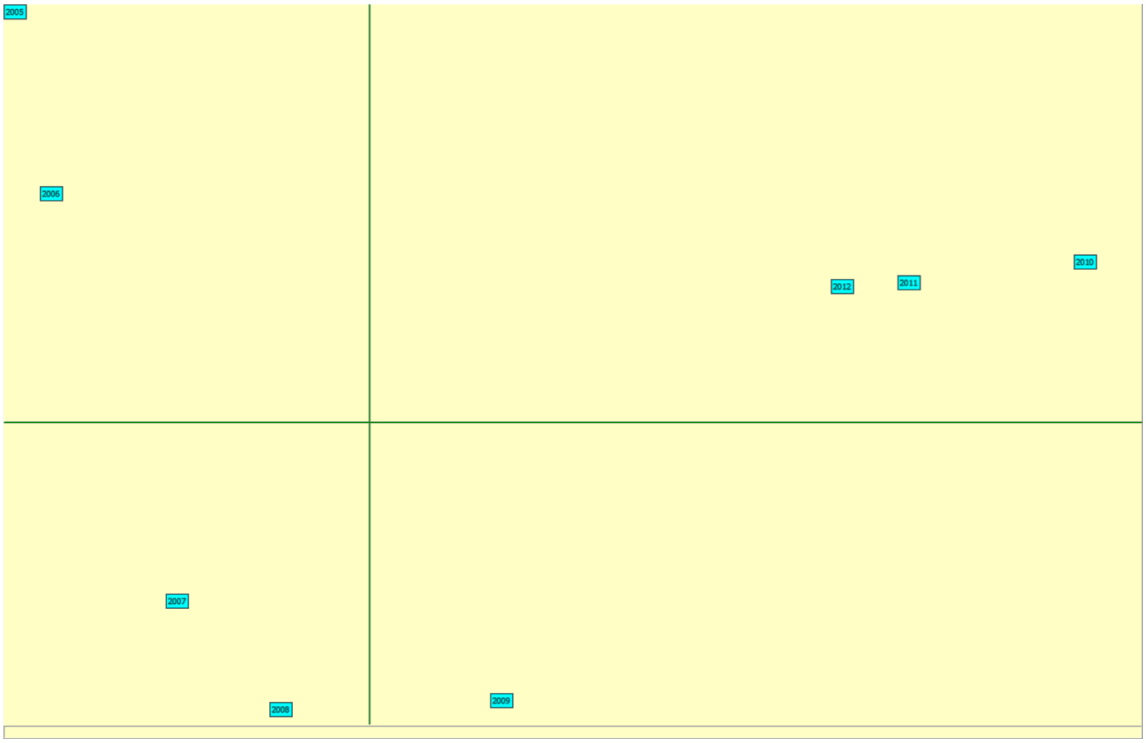


Figure B.20
ENRG_US : analyse factorielle des correspondances sur l'ensemble des années

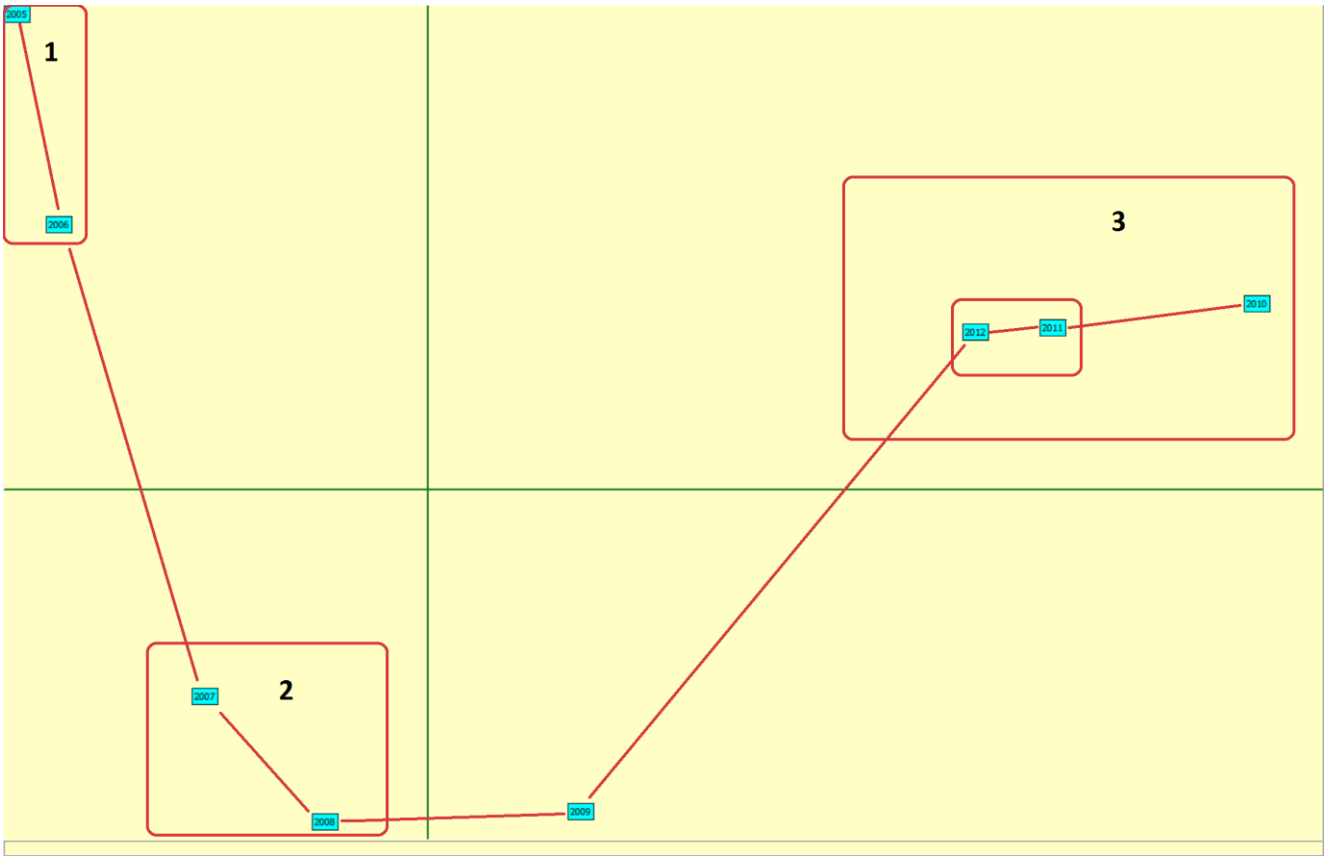


Figure B.21 (figure B.20 annotée)
ENRG_US : analyse factorielle des correspondances sur l'ensemble des années

Le graphique de la figure B.21 met clairement en évidence la proximité relative de chacune des années sur le plan factoriel, entre 2005 et 2012. Dans l'objectif de faciliter l'analyse, nous avons annoté la figure B.21 par leurs groupements textuels et par la série chronologique.

D'une part, il y a une nette séparation entre les années 2005, 2006, 2007 et 2008, à gauche de l'axe vertical et les années 2009, 2010, 2011 et 2012, à droite de ce même axe. Les années sont relativement éloignées les unes des autres, mises à part les années 2011 et 2012.
D'autre part, comme le sous-corpus français, nous voyons apparaitre approximativement, 3 groupements d'années et 1 année extrême en rupture avec ces 3 groupements (cf. figure B.21). Cela signifie que chacun des 3 groupements partage, *grosso modo*, respectivement des formes communes dans l'utilisation de leur vocabulaire. La période extrême (2009) écartée, se singularise dans une dimension individuelle, en rupture avec les 3 groupements.
Après l'observation de cette Analyse Factorielle des Correspondances, plusieurs constats en résultent :

4. Une année extrême : 2009 se singularise par son vocabulaire spécifique (tableau B.12) et devient l'un des perturbateurs de la série chronologique.
5. Une série chronologique apparait avec 2010 se trouvant à la fin de la courbe.
6. Les 3 groupements
   - Groupement 1 : 2005 et 2006

   L'écart est relativement important entre ces deux années.
   - Groupement 2 : 2007 et 2008

   L'écart est relativement important entre ces deux années.
   - Groupement 3 : 2010, 2011 et 2012

Ce dernier présente la particularité d'une série chronologique inversée, l'année 2010 étant à la fin de la série chronologique et éloignée du reste de ce groupement. De ce fait, 2010 est également un perturbateur de la série chronologique.

Afin de comprendre la typologie de ces 3 groupements et des 2 perturbateurs (2009 et 2010), nous faisons appel au calcul de spécificités avec le seuil égal à 5 et la Fréquence minimum égale à 10.

Les 2 perturbateurs (2009 et 2010)

En 2009

Tableau B.11
ENRG_US : sélection des spécificités positives de l'année 2009

Forme	Frq. Tot.	Fréquence	Coeff.
Obama	788	266	***
Copenhagen	152	105	***
stimulus	114	85	***
climate	2672	543	*30*
carbon	2781	528	*22*
energy	4842	820	*19*
emissions	3026	520	*14*

Selon le tableau B.11, l'année 2009 est marquée par les événements suivants tels que la Conférence de Copenhague *(Copenhagen, stimulus, climate, carbon, energy, emissions, colorés en vert*) et l'entrée en fonction à la Présidence des États-Unis de Barack Obama (*Obama coloré en vert*).

En 2010

Tableau B.12
ENRG_US : sélection des spécificités positives de l'année 2010

Forme	Frq. Tot.	Fréquence	Coeff.
BP	514	320	***
Gulf	321	150	***
gulf	260	201	***
oil	3614	1072	***
Deepwater	76	62	*47*
Horizon	72	58	*43*
Louisiana	208	83	*28*
Oil	307	104	*27*
deepwater	48	37	*27*
Obama	788	186	*25*
Mexico	431	123	*24*
offshore	358	108	*24*
disaster	256	79	*19*

D'après le tableau B.12, l'année 2010 est l'année de la marée noire dans le golfe du Mexique (*BP*, *Gulf*, *gulf*, *oil*, *drilling*, *Deepwater*, *Horizon*, … *disaster*, etc. colorés en vert, etc.), mais cet événement touche directement les États-Unis. Dans les spécificités saillantes, nous ne retrouvons pas de formes liées à l'éruption du volcan islandais, événement touchant principalement l'Europe. L'information n'est pas symétrique au sous-corpus français.

Annexe B

Groupement 1 : 2005 et 2006

Tableau B.13
ENRG_US : sélection des spécificités positives des années 2005 et 2006

Forme	Frq. Tot.	Fréquence	Coeff.
Pataki	143	123	***
Campbell	76	69	*33*
MTBE	56	56	*33*
mercury	393	216	*33*
Skies	77	69	*32*
Clear	87	75	*32*
McWane	43	43	*26*
Bush	1324	514	*24*
Boehlert	34	33	*19*
Freeport	39	36	*18*
Barton	43	38	*18*
(…)	(…)	(…)	*(...)*

Nous remarquons dans les spécificités positives des termes liés à la grippe aviaire (*bird,birds, flu*), que ces formes ne sont pas classées en tête du tableau avec la spécificité sur les deux années : *flu* fréquence totale de 24, fréquence de 15 pour la période considérée, coefficient 5, un phénomène qui avait pourtant marqué une bonne partie du monde en 2005 et 2006, notamment en France, selon la figure B.22 ci-dessous, calcul de ventilation des formes du sous-corpus français. La spécificité locale de cette forme a été submergée par les autres événements. Pour ce faire, nous avons recouru au module « concordance » (figure B.23) et au module ventilation (figure B.24) pour tenter d'illustrer le contexte.

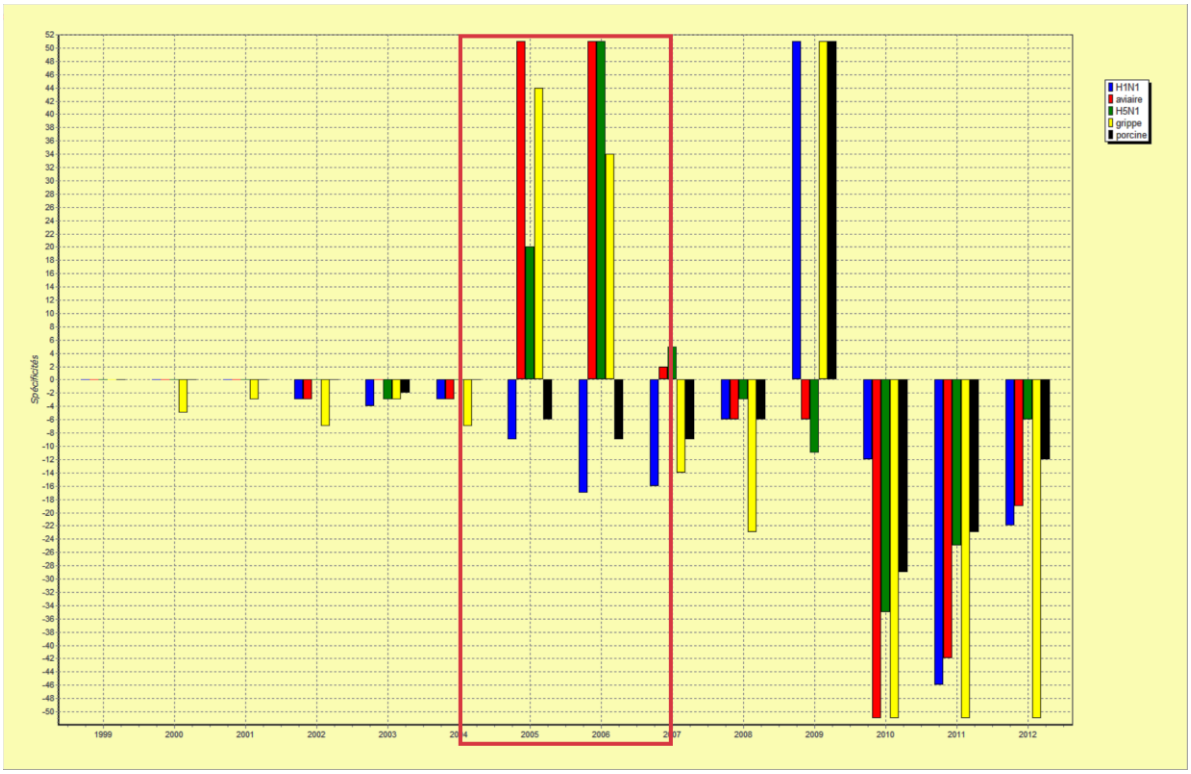
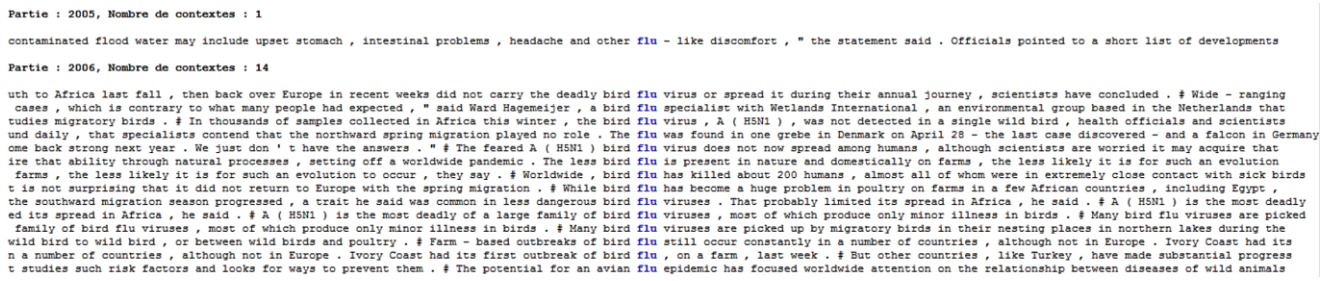


Figure B.22
ENRG_FR : ventilation des formes, *H1N1* en bleu, *aviaire* en rouge, *grippe* en jaune, *H5N1* en vert, *porcine* en noir, en 2005 et 2006



```
Partie : 2005, Nombre de contextes : 1

contaminated flood water may include upset stomach , intestinal problems , headache and other flu - like discomfort , " the statement said . Officials pointed to a short list of developments

Partie : 2006, Nombre de contextes : 14

uth to Africa last fall , then back over Europe in recent weeks did not carry the deadly bird flu virus or spread it during their annual journey , scientists have concluded . # Wide - ranging
 cases , which is contrary to what many people had expected , " said Ward Hagemeijer , a bird flu specialist with Wetlands International , an environmental group based in the Netherlands that
tudies migratory birds . # In thousands of samples collected in Africa this winter , the bird flu virus , A ( H5N1 ) , was not detected in a single wild bird , health officials and scientists
und daily , that specialists contend that the northward spring migration played no role . The flu was found in one grebe in Denmark on April 28 - the last case discovered - and a falcon in Germany
ome back strong next year . We just don ' t have the answers . " # The feared A ( H5N1 ) bird flu virus does not now spread among humans , although scientists are worried it may acquire that
ire that ability through natural processes , setting off a worldwide pandemic . The less bird flu is present in nature and domestically on farms , the less likely it is for such an evolution
 farms , the less likely it is for such an evolution to occur , they say . # Worldwide , bird flu has killed about 200 humans , almost all of whom were in extremely close contact with sick birds
t is not surprising that it did not return to Europe with the spring migration . # While bird flu has become a huge problem in poultry on farms in a few African countries , including Egypt ,
the southward migration season progressed , a trait he said was common in less dangerous bird flu viruses . That probably limited its spread in Africa , he said . # A ( H5N1 ) is the most deadly
ed its spread in Africa , he said . # A ( H5N1 ) is the most deadly of a large family of bird flu viruses , most of which produce only minor illness in birds . # Many bird flu viruses are picked
 family of bird flu viruses , most of which produce only minor illness in birds . # Many bird flu viruses are picked up by migratory birds in their nesting places in northern lakes during the
wild bird to wild bird , or between wild birds and poultry . # Farm - based outbreaks of bird flu still occur constantly in a number of countries , although not in Europe . Ivory Coast had its
n a number of countries , although not in Europe . Ivory Coast had its first outbreak of bird flu , on a farm , last week . # But other countries , like Turkey , have made substantial progress
t studies such risk factors and looks for ways to prevent them . # The potential for an avian flu epidemic has focused worldwide attention on the relationship between diseases of wild animals
```

Figure B.23
ENRG_US : concordance du mot *flu* (grippe) en 2005 et 2006

Selon les concordances (figure B.23 ci-dessus), l'apparition des formes liées à la grippe aviaire reste très faible. Au vu de ces constats, il semblerait qu'il y ait une légère ignorance de cet événement en 2005. Mais la ventilation (figure B.24) montre le contraire en 2006.

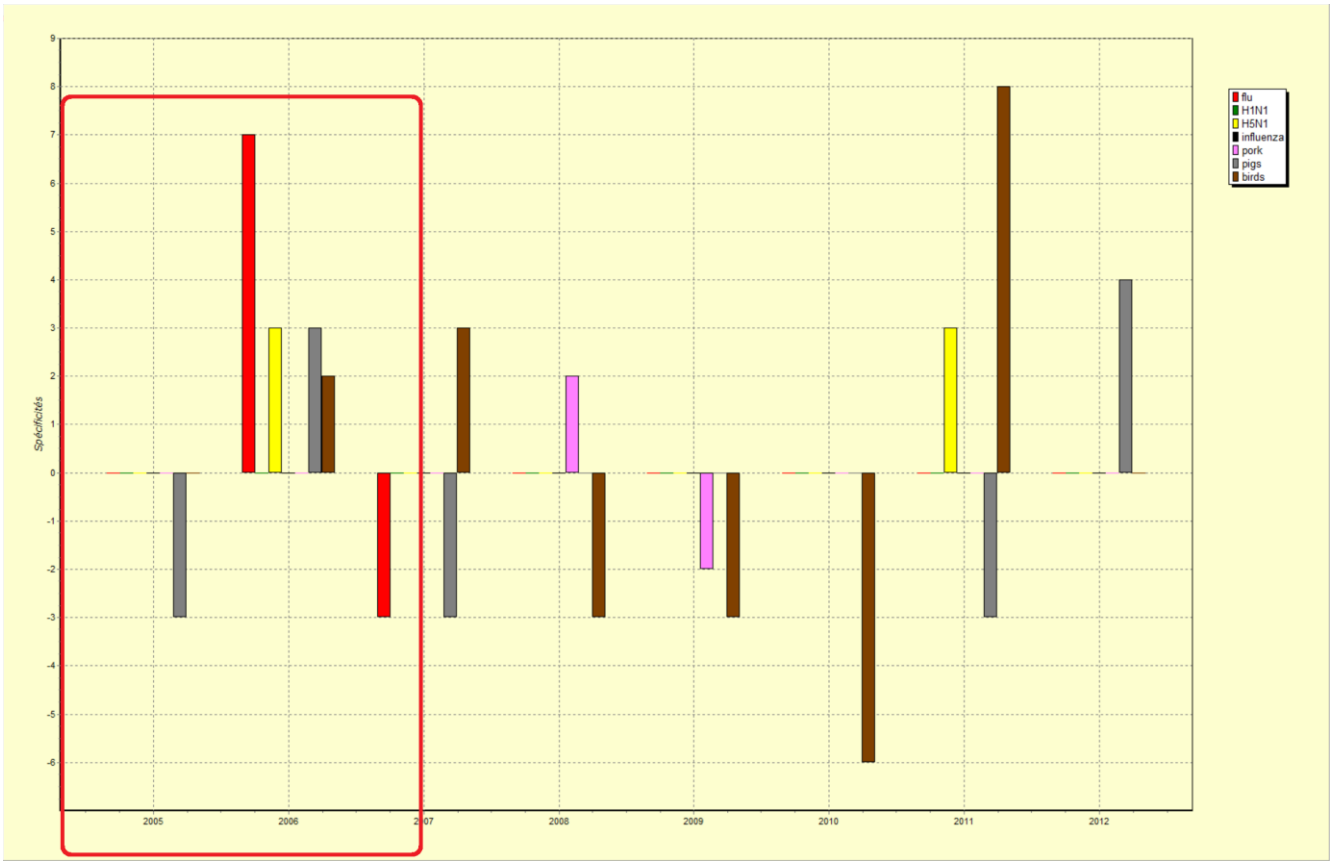


Figure B.24
ENRG_US : ventilation des formes *flu* (grippe) en bleu, *H1N1* en rouge, *H5N1* en vert, *birds* en jaune en 2005 et 2006

Cette différence de perception peut être due à la classification différente des événements entre la rubrique « Planète » du Monde et la sous-rubrique *Environment* du NYT.

Alors, quels sont les événements qui apparaissent primordiaux pour NYT ?

Dans cette période, NYT relate très largement les événements politico-environnementaux des États-Unis. Nous constatons, dans le tableau B.15 ci-dessous, une forte présence de formes liées aux noms d'hommes politiques américains (*Pataki, Campbell, Bush, Boehlert, Barton* en jaune,), à une loi fédérale (*Clear Skies* en jaune également) sur la protection de l'environnement et à des sociétés et des composants polluants (*McWane*, *Freeport*, *MTBE*[231], *mercury* en gris). Trois groupes de formes ont été constitués (figure B.25).

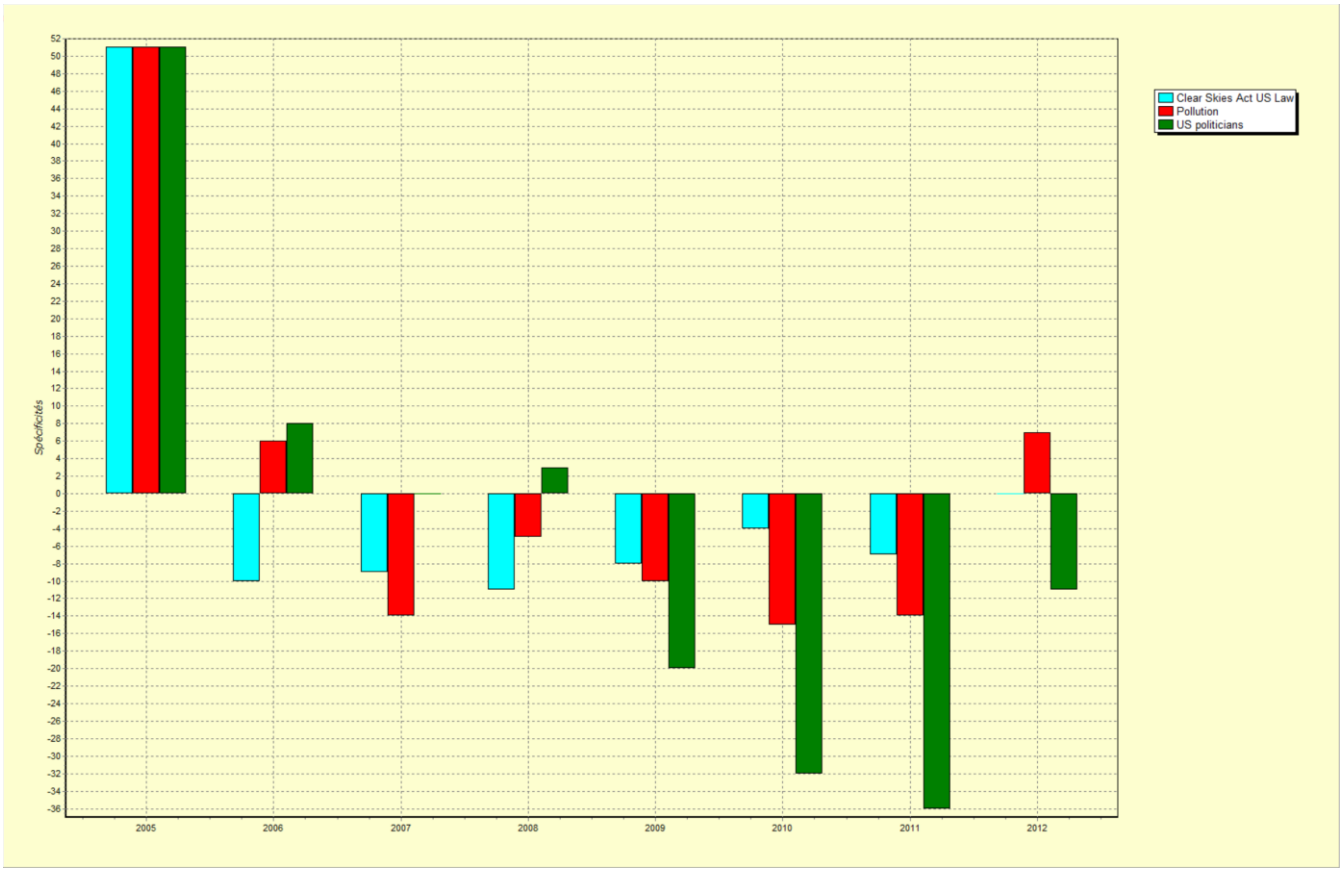


Figure B.25
ENRG_US : ventilation des 3 groupes de formes *Clear Skies Act US Law* en turquoise, *Pollution* en rouge, *US Politicians* en vert en 2005 et 2006

[231] Le Méthyl Tert-Butyl Ether ou MTBE est un composé organique de formule CH3OC(CH3)3. C'est un éther liquide, incolore, volatil et inflammable qui est non-miscible dans l'eau. Le MTBE a une odeur vaguement évocatrice du diéthyl éther et donne un goût désagréable et une odeur à l'eau. Le MTBE est un additif oxygéné de l'essence automobile, son utilisation a décliné en particulier aux États-Unis en réponse aux problématiques d'environnement et de santé. http://www.inrs.fr/publications/bdd/doc/fichetox.html?refINRS=FT%20242 (consulté le 26/06/2015)

Groupement 2 : 2007 et 2008

Tableau B.14
ENRG_US : sélection des spécificités positives des années 2007 et 2008

Forme	Frq. Tot.	Fréquence	Coeff.
McCain	208	194	***
bulbs	224	170	*30*
bags	488	302	*26*
TXU	101	90	*26*
Beijing	313	207	*23*
Games	62	55	*16*
Gore	323	197	*16*
carbon	2781	1273	*15*
green	2084	959	*13*
fluorescent	110	80	*13*
bikes	72	57	*13*
Bush	1324	634	*13*
(…)	(…)	(…)	*(...)*
Barack	86	64	*12*

Le tableau B.14 montre que les événements marquants sont :

1. En 2007, l'attribution du prix Nobel de la paix à Al Gore, défenseur de la cause environnementaliste, et l'achat de la société TXU par le plus grand emprunt de l'histoire,
2. En 2008, l'élection présidentielle (*McCain*, *Bush*, *Barack* formes colorées en vert), les Jeux Olympiques de Beijing et les thèmes liés à l'environnement en corrélation avec ces deux derniers (*bulbs*, *bags*, *green*, *fluorescent*, *bikes* formes colorées en vert).

Groupement 3 : 2010, 2011 et 2012

Nous avons considéré que l'année 2010 était un élément perturbateur de la série chronologique, année déjà traitée précédemment.

En 2011 et 2012

Tableau B.15
ENRG_US : sélection des spécificités positives des années 2011 et 2012

Forme	Frq. Tot.	Fréquence	Coeff.
fracking[232]	97	76	***
pipeline	289	142	***
Keystone	68	54	*38*
TransCanada	49	43	*35*
XL	51	42	*31*
hydrofracking	41	35	*28*
fracturing	125	61	*25*
gas	2616	478	*25*
drilling	959	219	*23*
Texas	467	127	*21*

[232] Terme technique désigne en français la fracturation hydraulique, le « fracking » est une fissuration massive d'une roche au moyen d'une injection d'un liquide sous pression. Cette technique permet de récupérer du pétrole ou du gaz dans des substrats trop denses, où un puits classique ne produirait rien ou presque. http://www.futura-sciences.com/magazines/terre/infos/dico/d/geologie-fracturation-hydraulique-9048/ (consulté le 30-06-2015)

hydraulic	136	58	*20*
sands	129	51	*16*
(…)	(…)	(…)	*(...)*
Fukushima	14	14	*14*

Les spécificités du tableau B.15 témoignent, encore une fois, d'une prédominance des événements nationaux (*fracking*, *pipeline*, *Keystone*, *TransCanada*, *XL*, *hydrofracking*, *fracturing*, *gas*, *drilling*, *hydraulic*, *sands* etc. colorés en vert), rapportant entre autres l'événement du 6 novembre 2011, 12 000 manifestants se rendent à la Maison-Blanche pour protester contre le projet d'extension de ce pipeline. Le 10 novembre 2011, le Département d'État des États-Unis ordonne une révision immédiate de l'itinéraire de Keystone XL afin que celui-ci évite la zone sensible de Sandhills dans le Nebraska, une zone humide fragile qui pourrait être menacée par un déversement de pétrole.
Cependant, par rapport à l'international (*Fukushima* coloré en rouge), qui a été une catastrophe aux répercussions internationales selon certains (classée niveau 7 sur l'échelle INES) n'apparait que dans le second plan.

Analyses transversales des sous-corpus français et américain

Bi-résonance, symétrie, discursivité entre les sous-corpus français et américain

A travers les résultats des analyses empiriques, que nous avons vus précédemment sur l'aperçu général des deux sous-corpus, nous appliquons la méthode analytique pour la textométrie multilingue « détections et analyses *critériées* (réparties par critères) par croisement des disciplines », nous constatons :

1. La perception des informations, relayant les événements dans le monde francophone français et anglophone américain, est très différente, que ce soit sur le plan de la classification des thèmes des articles, que sur l'accentuation et l'acharnement des répétitions des faits. Mais une certaine bi-résonance textuelle synchronique des formes s'illustre lors de grands événements internationaux.
2. Une certaine symétrie dans la classification des événements par l'AFC et des séries chronologiques a été respectée au sens très large dans le regroupement des spécificités des années (3 groupements d'années symétriques et des perturbateurs dans les deux sous-corpus).
3. La discursivité textuelle est différente dans les deux sous-corpus. D'une part, le sous-corpus français, comme dans la tradition de la langue de Molière, défend un système discursif solide dans la narration des faits par un foisonnement de vocabulaire stable dans le temps témoignant d'une richesse syntaxique, que ce soit en diachronie ou en synchronie. D'autre part, le NYT privilégie les événements informationnels avant tout nationaux et ponctuels, qu'internationaux, se manifestant par un système d'élocutions récurrentes, moins riche et moins diversifié et un renouvellement relativement lent du vocabulaire (figures des accroissements), (Bell, 1991 ; MacMurray, 2012).

B.3 Sous-corpus chinois issu de divers supports

Figure B.26
ENRG_CN : structure du sous-corpus

Ce sous-corpus chinois se divise en différents empans textuels suivant une chronologie. Les articles sont triés par ordre croissant sur leurs dates de publication (aaaammjj), regroupés par jour, par mois, par année, à l'aide de programmes informatiques spécifiques. Chaque balise désigne le début d'un empan textuel, comme illustré dans la figure B.26 :

- <year=aaaa> : balise indiquant l'année des articles. Le sous-corpus chinois s'étale sur 6 années, de 2008 à 2013.
- <month=aaaamm> : balise indiquant le mois des articles.
- <day=aaaammjj> : balise indiquant le jour des articles, sachant que dans une journée, plusieurs articles sont susceptibles d'être regroupés sous cette balise. L'ensemble du sous-corpus contient 698 jours.
- <media=nomdumédia> : balise indiquant le nom du media de cet article.
- <article=nnnn> : balise indiquant le numéro de chaque article. Les textes rassemblés sont au nombre de 14 514 articles.
- Le signe « # » marque le début de chaque paragraphe de chaque article.
- La fin de chaque article est marquée par la chaîne de caractères « $Fend »

Annexe B

Tableau B.16
ENRG_CN : sélection des spécificités positives du mois de juillet 2010 et leurs équivalences en français

Equivalent en français des formes	Forme	Frq. Tot.	Fréquence	Coeff.
Pavillon (Exposition Universelle)	馆	860	325	***
Profiter de l'air frais (se réfugier de la chaleur)	纳凉	76	72	***
Shanghang (nom du siège de comté)	上杭	201	147	***
Dalin	大连	895	246	***
Cuivre	铜	728	268	***
Exposition Universelle	世博会	731	324	***
Océan	海洋	3764	543	***
Industries minières	矿业	2189	968	***
Poisson(s)	鱼	2596	456	***
Accident(s)	事故	4759	900	***
Eaux polluées	污水	5759	840	***
Parc de l'expo	世博园	217	121	***
Shanghai	上海	4622	672	***
Fuite	泄漏	1393	304	***
Clair, éclairer, nettoyer,	清	1720	340	***
Zijin (nom d'une société)	紫金	1814	996	***
Explosion	爆炸	535	161	***
Tingjiang (ville)	汀江	540	403	***
Shanghang (siège du comté dans le Fujian)	上杭县	423	299	***
Zijinshang (nom d'un lieu ou d'un société)	紫金山	420	286	***
Mine de cuivre	铜矿	490	288	***
Province du Fujian	福建省	476	159	***
Exposition Universelle	世	595	215	***
Fuite (par pénétration)	渗漏	473	235	***
Huile lourde (de la pollution)	油污	624	193	***
Province du Fujian	福建	753	187	*50*

Annexe B

Tableau B.17
ENRG_CN : sélection des spécificités positives de l'année 2010 et leurs traductions

Equivalent en français des formes	Forme	Frq. Tot.	Fréquence	Coeff.
Diminuer ou moins	减	13614	6878	***
Mexique	墨西哥	1016	895	***
Conférence	会议	7234	4122	***
Economie d'énergie	节能	13829	6583	***
Parc de l'Exposition Universelle	世博园	217	198	***
Ecologie	生态	13295	5592	***
Pavillon	馆	860	601	***
Température élevée	高温	824	503	***
Climat	气候	14816	8620	***
Emission	排	15684	7627	***
Climatisation	空调	1151	646	***
Vert	绿色	14094	6148	***
Shanghang	上杭	201	190	***
Copenhague	哥本哈根	1446	1163	***
Convention	公约	1428	791	***
La rivière Songhua	松花江	373	265	***
Dingjiang (ville)	汀江	540	513	***
Fuite	渗漏	473	318	***
Le comté de Shanghang	上杭县	423	387	***
Cuivre	铜矿	490	372	***
Changement	变化	10427	5606	***
Zijin shan (nom de société)	紫金山	420	365	***
Exposition Universelle	世博会	731	594	***
Protocoles	议定书	3267	1854	***
Limiter	限	3237	1609	***
Protection environnementale	环保	28774	10823	***
Taxes, impôt	税	3316	1626	***
Capacité de production	产能	2798	1440	***
Cancun (Kan)	昆	3601	3245	***
Cancun (En)	坎	3561	3226	***
Comté Anping	安平县	134	132	***
Exploitation minière	矿业	2189	1699	***
Zijin (nom de société)	紫金	1814	1634	***
Faible	低	16181	7805	***
Onzième quinquennat	十一五	1942	1012	***
Dioxyde de carbone	二氧化碳	2628	1294	***
Carbone	碳	22186	10901	***
Kyoto	京都	2609	1393	***
Consommation d'énergie	能耗	2645	1345	***

Annexe B

En 2011

Tableau B.18
ENRG_CN : sélection des spécificités positives de l'année 2011 et leurs traductions

Equivalent en français des formes	Forme	Frq. Tot.	Fréquence	Coeff.
Radioactivité ou radiation	辐射	971	705	***
Arômes	香精	324	291	***
Incinération	焚烧	2937	1911	***
Taïwan	台湾	995	707	***
Lait	乳	987	769	***
Apple ou pomme	苹果	917	724	***
Sécurité	安全	10718	5968	***
Champ électrique	电场	1007	724	***
Electrique	电	15939	8557	***
Lait en poudre	奶粉	1147	832	***
Conservateur	防腐剂	330	297	***
Additifs ou Ajouter à	添加	1070	758	***
Aliments	食品	10433	7235	***
Usine	总厂	223	215	***
Frelater ou Mélange	勾兑	283	256	***
Viande maigre	瘦肉	642	537	***
Huile	油	8060	5305	***
Brioches ou gâteaux cuits à la vapeur	馒头	288	261	***
Lait au soja	豆浆	498	454	***
Animal	动物	6512	3981	***
Plomb-Acide	铅酸	499	445	***
Ramen (nouilles)	拉面	191	189	***
Vent	风	10246	6735	***
Porc	猪	708	598	***
ConocoPhillips[233] (nom d'une société américaine)	康菲	683	544	***
CNOOC	中海	749	563	***
Porc (viande de porc)	猪肉	725	603	***
Organique	有机	2211	1415	***
Repas	餐	1376	968	***
Bœuf	牛肉	381	349	***
Supermarchés	超市	1704	1139	***
Ordures	垃圾	18160	10376	***
Caniveaux ou égouts	地沟	2037	1647	***
Penglai (ville au Shangdong)	蓬莱	362	307	***
Scories (résidus)	渣	1418	977	***
Cuisine	厨	2009	1416	***
Ester (composé produit par la réaction entre un acide et un alcool)	酯	365	311	***
Lait	奶	1351	1062	***
Additif	添加剂	1578	1277	***
Lait de vache	牛奶	692	518	*50*
Apple (Groupe)	苹果公司	346	295	*50*

[233] ConocoPhillips (NYSE : COP) est une entreprise américaine spécialisée dans l'extraction, le transport et la transformation du pétrole. Elle exploite aussi des réseaux de stations-service dans différents pays.

Annexe B

En 2012

Tableau B.19
ENRG_CN : sélection des spécificités positives de l'année 2012 et leurs équivalences en français

Equivalent en français des formes	Forme	Frq. Tot.	Fréquence	Coeff.
Medecine chinoise traditionnelle	中药	503	401	***
Hechi (ville dans le Guangxi)	河池市	247	238	***
Œufs	鸡蛋	513	364	***
Concentration ou densite	浓度	2035	911	***
Wangjiang (siège d'un comte)	望江县	94	94	***
Qualité	质量	7916	2590	***
Poudre	粉	988	474	***
Prélèvement et surveillance	监测	7763	2889	***
Le lac Poyang	鄱阳湖	977	478	***
Source de la pollution	污染源	951	421	***
Médicinal	药用	251	193	***
Contenu	含量	2351	800	***
Vésicule biliaire ou bile	胆	2308	1820	***
Rentrer ou retour (nom de marque)	归	926	556	***
Air	空气	6987	2167	***
Sécurité	安全	10718	2771	***
Lac	湖泊	1103	450	***
Membre du comité	委员	1064	491	***
Phénol	苯酚	101	92	***
Médicaments	药品	1114	644	***
Centrale nucléaire	核电厂	206	144	***
Pengze (nom d'une centrale nucléaire au Jiangxi)	彭泽	205	193	***
Centrale Hydroélectrique	水电站	1056	436	***
Pharmaceutique	药业	311	263	***
Longjiang (nom d'une rivière)	龙江	493	355	***
Zhen (marque pharmaceutique)	真	1894	763	***
Guangxi (province)	广西	1036	513	***
Puissance nucléaire	核电	3302	1052	***
Prix	水价	691	323	***
Gélatine	明胶	606	603	***
Mètre cube	立方米	2843	1037	***
Temple Guizhen (marque pharmaceutique chinoise)	归真堂	597	504	***
Liuzhou (Ville)	柳州市	273	269	***
Sol	地板	388	303	***
Chenzhou (ville)	郴州	163	127	***
Dépasser normes autorisée	超标	3854	1354	***
Temple (marque pharmaceutique)	堂	663	454	***
Qualité de l'eau	水质	3442	1133	***
PM 2.5 (Les particules en suspension)	PM2	4111	2073	***
Hechi (ville au Guanxi)	河池	164	162	***
Ours	熊	3413	2538	***
Rivière(s)	江河	787	454	***
Aviation	航空	2487	831	***
Détecter	检测	5348	1693	***
Détournement d'une rivière (rediriger)	引流	254	212	***

Liuzhou (ville dans le Guangxi)	柳州	372	341	***
Capsule	胶囊	553	529	***
Gouverner	治理	4860	1529	***
Plomb	铅	2750	963	***
Sources d'eau	水源	2726	1161	***
Métaux lourds	重金属	2694	1061	***
Ours noir	黑熊	754	482	***
Atmosphère	大气	2635	859	***
Les terres rares[234]	稀土	2678	1028	***
Compléments alimentaires pour la santé	保健食品	172	146	***
La ville de Chenzhou	郴州市	110	96	***
Eau potable	饮用水	1511	701	***
Cadmium	镉	1514	1048	***
Précipitation (terme chimique) ou Subsidence (géologie)[235] ou descendre	沉降	461	253	***
Pharmaceutique	制药	437	246	***
Zhenjiang (ville dans le Jiangsu)	镇江	230	175	***
Liujiang (la rivière de la ville de Liuzhou)	柳江	230	229	***
Pollution	污染	22616	6364	***
Certifié conforme	合格	1601	580	***
Bile (un liquide organique)	胆汁	346	264	***
Standard ou normes	标准	14369	3581	***
Eau	水	17531	4816	***
Approvisionnement en eau	供水	1449	572	***
Marsouin	江豚	1209	511	***
Eau courante	自来水	1368	640	***
Crevette	虾	418	222	***
Urgence	应急	1505	575	***
Vie	活	1391	676	***
Médicament	药	1173	597	***
Dragon (ici, nom d'une rivière au Guangxi)	龙	1179	476	***
Filtre	净化器	85	80	*50*
Santé	保健品	161	119	*50*
Zhenjiang (ville)	镇江市	114	96	*50*
Prix	价格	6373	1740	*49*
Union Européen	欧盟	4382	1265	*48*

[234] Les terres rares sont entre autres composées de dysprosium, gadolinium, samarium, cérium, métaux aux propriétés particulières qui leur confèrent un magnétisme et une luminescence hors du commun et qui sont utilisés dans d'innombrables objets de technologie (smartphones, télévisions, avions, voitures, radars, missiles, etc.). 98% de la production mondiale de terres rares provient de Chine, soit une situation de quasi-monopole.

[235] La subsidence en géologie est un lent affaissement de la lithosphère entraînant un dépôt progressif de sédiments sous une profondeur d'eau constante.

Annexe C : la politique nucléaire mondiale

C.1 Quelques événements et accidents/incidents nucléaires marquants

Au début du développement de l'énergie nucléaire, les mesures concernant la sécurité et la sûreté des installations n'étaient pas une préoccupation majeure des gouvernements. Cette nouvelle énergie était mal connue et les conséquences d'incidents sur l'environnement n'étaient pas une source d'inquiétude majeure. Avant toute implantation de centrales, des études géologiques étaient menées afin de s'assurer de la stabilité des sols, des normes et règles de sécurité étaient également appliquées, mais les risques naturels, telle que la catastrophe de Fukushima, n'avaient pas été envisagés.

Des événements, incidents ou accidents ont émaillé toute l'activité nucléaire en France et dans le monde. Les informations relatives à ces événements proviennent principalement des ONG (par exemple Greenpeace) et d'associations libres (comme Sortir du nucléaire).

Le tableau C.1 ci-dessous, présente quelques événements marquants internationaux se rapportant d'une part au nucléaire, d'autre part à des catastrophes naturelles et humaines, qui ont engendré de gigantesques problèmes sur l'environnement.

Tableau C.1
Quelques événements et accidents/incidents marquants en Europe, U.S.A, Chine et Japon de 1995 à nos jours[236]

Evénements internationaux (liste non exhaustive)	
Novembre-décembre 2015 : COnférence mondiale des Parties (COP21) sur les changements climatiques, « Paris Climat 2015 », conférence visant à limiter le réchauffement climatique.	
2011 : le gouvernement américain doit approuver le projet de pipeline Keystone XL évalué entre 5 et 7 milliards de dollars, pipeline permettant de faire transiter 830 000 barils de pétrole brut de l'Alberta vers les raffineries du Texas.	
Janvier 2010 : un séisme de magnitude 7 a frappé Haïti plongeant la capitale dans le chaos, sous un épais nuage de poussières. Les dégâts sont considérables et le bilan des victimes dépasse 300 000 morts.	
20 avril 2010 : explosion de la plate-forme pétrolière Deepwater Horizon a provoqué la plus gigantesque marée noire mondiale dans le Golfe du Mexique.	
20 mars 2010 : le volcan islandais Eyjafjöll est entré en éruption le 20 mars 2010, provoquant un nuage de cendres rendant impossible tout trafic aérien au-dessus d'une grande partie de l'Europe.	
2010 : Exposition Universelle à Shanghai	
2009 : la conférence de Copenhague sur le climat a mis en lumière le rôle incontournable de la Chine et des Etats-Unis, les deux plus grands pollueurs de la planète. De vives tensions entre les 2 pays ont marqué la conférence.	
2008 : année des élections américaines. Un des thèmes majeurs de la campagne présidentielle porte sur l'environnement, notamment sur une limitation des émissions de gaz à effet de serre. Barack Obama défend la filière éthanol et les énergies renouvelables, tandis que McCain pousse à la construction de nouvelles centrales nucléaires.	
15 octobre 2007 : ouverture du 17ème congrès du parti communiste chinois à Pékin (un tous les cinq ans). Réélection du secrétaire général Hu Jintao qui affiche sa volonté de combattre les inégalités sociales et régionales.	
Octobre 2007 : le comité Nobel attribue le prix Nobel de la paix à Al Gore et au groupe d'experts intergouvernemental sur l'évolution du climat (GIEC). Ce prix récompense les efforts menés par les lauréats qui ont essayé de sensibiliser l'opinion mondiale à la gravité du réchauffement climatique.	
16 février 2005 : entrée en vigueur du protocole de Kyoto	
26 décembre 2004 : un séisme dans l'océan Indien au large de Sumatra d'une magnitude > 9. Le tremblement de terre a provoqué vingt minutes plus tard un tsunami avec des vagues allant jusqu'à plus de 30 mètres de hauteur, ce séisme a frappé un vaste territoire allant de l'Indonésie, les côtes du Sri Lanka et du sud de l'Inde, ainsi que l'ouest de la Thaïlande.	
1997 : signature du protocole de Kyoto (Dollfus, 1999 : 34), accord international visant à la réduction des émissions de gaz à effet de serre et venant s'ajouter à la Convention-cadre des Nations Unies sur les changements climatiques.	
Accidents et/ou incidents nucléaires (liste non exhaustive)[237]	
FRANCE	**gravité[238]**
28/04/2014 - Golfech : baisse de pression du circuit primaire	moyenne
27/04/2014 - Chinon : disfonctionnement de deux capteurs de mesure du niveau d'eau	moyenne
26/04/2014 - Cattenom : départ d'incendie	moyenne

[236] Source : les informations sont extraites de la revue Réseau « Sortir du nucléaire », www.sortirdunucleaire.org

[237] http://www.irsn.fr/FR/connaissances/Installations_nucleaires/La_surete_Nucleaire/echelle-ines/Pages/2-Incidents-accidents.aspx?dId=8a15297f-e5f9-42cd-9765-ed2049203773&dwId=a1de7c68-6d78-4537-9e6a-e2faebed3900 (consulté le 8/08/2015)

[238] Niveau de gravité classé selon INES - Institut Nuclear Event Scale - (échelle internationale de classement des événements nucléaires).

25/04/2014 - Romans-sur-Isère : découpe accidentelle d'un conteneur contenant de la poudre d'uranium enrichi	moyenne
23/04/2014 - Tricastin : modification temporaire des règles générales d'exploitation	moyenne
23/04/2014 - Civaux : dégagement de fumée sur un appareil de filtration de l'air de l'unité de production n° 1	moyenne
11/04/2014 - Fessenheim : erreur de réglage d'une vanne d'isolement de l'enceinte du réacteur n° 1	moyenne
9/04/2014 - Fessenheim : inondation interne dans la partie non nucléaire du réacteur n° 1	moyenne
8/04/2014 - Dampierre-en-Burly : défaut d'isolement extérieur de l'enceinte de confinement lors des opérations de redémarrage du réacteur.	moyenne
7/04/2014 - Blayais : défaut d'isolement sur un tableau d'alimentation électrique 48 volts	moyenne
3/04/2014 - Gravelines : fuite d'un robinet du système de production d'eau glacée	moyenne
2/04/2014 - Dampierre-en-Burly : défaut de la mesure du niveau d'eau dans la piscine du bâtiment réacteur pendant le rechargement en combustible	moyenne
2011 : Marcoule : explosion dans un four de retraitement des déchets nucléaires	1
2010 : incidents à la Hague à l'usine de retraitement des déchets radioactifs	
2002 : fuite radioactive d'un fût expédié de Suède et transitant par Roissy	3
2000 : incidents à répétition à Dampierre en Burly (Loiret)	
1980 : Saint-Laurent-des-Eaux : défaillance technique entraînant une inflammation locale du combustible	4
ROYAUME-UNI	
06/01/2014 - Heysham : EDF met le réacteur "Heysham 1" à l'arrêt suite à la perte d'une pompe du réacteur.	moyenne
2005 - Fuite nucléaire à Sellafield	moyenne
1957- Incendie à la centrale de Windscale renommée Sellafield	**5**
U.R.S.S	
2000 - Mer de Barents : poubelle nucléaire où vient de couler le sous-marin russe « Koursk », mais on retrouve beaucoup de déchets nucléaires radioactifs et irraidés.	**catastrophe**
1986- Catastrophe de Tchernobyl	**catastrophe (7)**
1957- Catastrophe de Kyshtym : Explosion d'un réservoir contenant des déchets radioactifs	**catastrophe (6)**
U.S.A	
08/04/2014 -Vogtle : arrêt automatique du réacteur n°2 suite à une baisse rapide du niveau de vapeur sur le circuit du générateur de vapeur	critique
03/04/2014 - Arkansas : arrêt automatique du réacteur	haute
02/04/2014 - Perry relâchement de gaz toxique intempestivement dans la zone protégée est du trichloréthylène	moyenne
02/04/2014 - Quad Cities : alerte déclarée suite à un incendie dans le bâtiment turbine du réacteur 2	haute
01/04/2014 - Seabrook : arrêt automatique du réacteur suite à un problème électrique	haute
31/03/2014 - Quand Cities : arrêt du réacteur n°2 suite à une fuite de pression	haute
29/03/2014 - Grand Gulf : arrêt d'urgence du réacteur suite à l'arrêt d'une turbine	haute
25/03/2014 - Clinton : arrêt manuel d'urgence du réacteur suite à la perte de vide au condenseur	haute
20/03/2014 - Fermi : alerte déclenchée suite à un incendie sur le calorifuge du groupe électrogène d'urgence	haute
19/03/2014 - Turkey Point : découverte d'une fuite potentielle autour de la réserve de chauffage du pressuriseur	moyenne
18/03/2014 - Browns Ferry : arrêt d'urgence du réacteur suite à l'impossibilité de contrôler le niveau du séparateur d'humidité de la turbine principale	haute
17/03/2014 - Fort Calhoun : arrêt automatique du réacteur suite à l'arrêt turbine initié par la perte de l'eau du refroidissement du stator.	haute
17/03/2014 - Grand Gulf : arrêt manuel d'urgence suite à une fuite de vapeur dans la ligne basse-pression de la turbine.	haute
10/03/2014 - Calvert Cliffs : Un incendie menace les équipements de mise à l'arrêt des réacteurs	moyenne
10/03/2014 - Nine Mile Point : Arrêt d'urgence du réacteur n° 2 suite à l'insertion de barres de contrôle	moyenne
04/03/2014 - Limerick : Arrêt d'urgence manuel suite à un arrêt rapide dû à un problème de turbine (défaillance sur le circuit électrique de commande hydraulique)	haute
26/02/2014 - Carlsbab : 13 employés exposés à des radiations suite à l'élévation du niveau de radioactivité artificielle contenu dans l'air du site de stockage de déchets radioactifs du Nouveau Mexique	haute
12/02/2014 - North Anna : les fûts de déchets radioactifs décalés par le séisme peuvent rester là où ils sont. Le tremblement de terre du 23 août 2011 (de magnitude 5.8) a décalé les fûts de déchets radioactifs pesant 115 tonnes	moyenne
04/02/2014 - Cooper : accident mortel d'un travailleur sous-traitant, la NRC se permet d'annoncer, avant même le lancement de l'enquête, que la cause du décès ne serait pas liée à l'activité professionnelle	haute
03/02/2014 - Brunswick : événement inhabituel lié à un relâchement de gaz toxique susceptible d'affecter les opérations normales du réacteur	moyenne
02/02/2014 - Diablo Canyon : arrêt automatique du réacteur n° 2. Durant un orage, la foudre s'est abattue sur une des phases d'alimentation électrique du réacteur	haute
31/01/2014 - Salem : arrêt manuel du réacteur n° 2 suite à l'atteinte de la limite basse de température	haute
31/01/2014 - Millstone : non-respect de prescriptions concernant la défaillance d'un groupe de chauffage du pressuriseur	haute
29/01/2014 - Palisades : problèmes identifiés sur les mécanismes de commande des barres de contrôle	moyenne
20/01/2014 - Perry : découverte de tritium dans les eaux souterraines	moyenne
18/01/2014 - South Texas : événement inhabituel engendré par un incendie en zone protégée	haute
1979- Catastrophe nucléaire de Three Mile Island. : dysfonctionnement du système de refroidissement	**catastrophe (5)**
CHINE	
03/02/2012 : selon un journal japonais, des problèmes seraient survenus sur un surgénérateur expérimental chinois, la Chine dément.	moyenne
JAPON	

2013 : fuite de 300 tonnes d'eau radioactive à la centrale de Fukushima	3
21/06/2012 - centrale nucléaire de Ooi : KEPCO a attendu 10 heures avant d'annoncer une alarme et une fuite, avec l'approbation de la NISA	moyenne
10/06/2011 - Fukushima : TEPCO a annoncé sa volonté de relâcher 3 000 tonnes d'eau « faiblement » radioactive dans la mer. La compagnie a ensuite fait état d'une fuite d'huile inexpliquée	haute
27/05/2011 - Fukushima : un incendie s'est déclaré dans le sous-sol d'une annexe à Fukushima Daini	haute
18/05/2011 - Fukushima : une fuite d'Arsenic 76 radioactif s'est produite sur le réacteur n° 5 de la centrale d'Hamaoka. 500 tonnes d'eau de mer avaient pénétré dans ce réacteur lors de sa mise à l'arrêt.	haute
02/05/2011 : une augmentation des niveaux de radioactivité a été constatée dans l'eau de refroidissement d'un réacteur à la centrale de Tsuruga, à quelques 350 kilomètres à l'ouest de Tokyo, indique Japan Atomic Power, l'exploitant de l'installation.	haute
09/04/2011 : la réplique sismique principale de jeudi 7 avril 2011 a neutralisé toutes les lignes à haute tension extérieures à la centrale nucléaire de Higashidori dans la préfecture d'Aomori.	haute
07/04/2011 : de l'eau s'échappe notamment des piscines de stockage de combustible usagé dans les réacteurs 1, 2 et 3 de la centrale. Trois autres fuites ont été signalées dans le réacteur 3	haute
11/03/2011- Accident nucléaire de Fukushima : le tremblement de terre du vendredi 11 mars 2011, au large des côtes japonaises, et le tsunami qui s'en est suivi n'ont pas fini d'avoir des répercussions sur l'île et dans le monde entier (désastre humanitaire et son terrible bilan, ainsi que la catastrophe nucléaire)	**catastrophe (7)**
16/07/2007 : nombreux dégâts et pollution environnementale suite au séisme qui a secoué les sept réacteurs de la centrale nucléaire de Kashiwazaki-Kariwa. C'est la première fois au monde qu'un tremblement de terre est identifié comme étant à l'origine d'un incendie.	haute
30/09/1999 - Tokaï-Mura : au cours d'une opération de chargement en oxyde d'uranium, une erreur d'accumulation de matière fissile a entrainé un accident dit de criticité : cet accident a entrainé une contamination à l'extérieur de l'usine.	haute
15/12/1995 - Monju : un incident grave est survenu dans le surgénérateur, situé à proximité de Tsuruga, sur la côte ouest du Japon, à quelque 300 km de Tokyo. Une fuite de sodium dans le circuit secondaire de refroidissement du surgénérateur, qui fonctionnait alors à 40% de sa capacité, a obligé les ingénieurs de garde à procéder à un arrêt manuel.	haute

C.2 L'échelle INES

Cette « norme » internationale, International Nuclear Event Scale, permet de classer les incidents et les accidents nucléaires par niveau de gravité. Sa description et sa modélisation sont disponibles dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-C-nucléaire-monde

C.3 La politique nucléaire mondiale après la catastrophe de Fukushima

Un bilan de la situation (2012-2013) de l'industrie nucléaire dans le monde est présenté dans l'article de Messieurs Schneider et Froggatt (Schneider et Froggatt, 2014), article disponible dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-C-nucléaire-monde. Les grandes décisions liées au nucléaire dans les pays concernés par cette énergie sont relatées dans le tableau C.2 disponible dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-C-nucléaire-monde.

Et aujourd'hui?

Le parc nucléaire mondial a fourni presque 11% de l'électricité produite dans le monde.

Nous avons synthétisé dans le tableau C.3, le nombre de réacteurs dans les 3 pays qui nous intéressent, la France, les Etats-Unis et la Chine. Ce tableau ainsi que la répartition de la part du nucléaire dans la production d'électricité sont disponibles dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-C-nucléaire-monde.

Lien du serveur dédié :
https://drive.google.com/folderview?id=0B8XHfHwNzWAAeDN6UEc1b2dGa1U&usp=sharing

Annexe D : le nucléaire et la politique française

Evénements marquants concernant la politique nucléaire civil

Un panorama de ces événements est disponible d'une part, dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-D-nucléaire-France, d'autre part sur le site vie-publique.fr[239].

(Lien du serveur dédié :
https://drive.google.com/folderview?id=0B8XHfHwNzWAAeDN6UEc1b2dGa1U&usp=sharing)

En résumé, les quelques lignes ci-dessous, exposent les grandes tendances de la politique nucléaire française. Dès les années 1960, les politiques se sont intéressés aux questions portant sur le nucléaire. D'abord, le général de Gaulle présente sa conception de l'indépendance énergétique qui passe, par la maîtrise de l'arme atomique et de l'énergie nucléaire : *« ...Nous avons décidé d'avoir ce qu'il nous faut pour nous défendre et d'autant plus que cette puissance nucléaire est liée directement à l'énergie atomique, elle-même qui est, comme vous le savez tous, le fond de l'activité de demain ...* »[240].

Ensuite, le contexte mondial joue en faveur du développement de l'énergie nucléaire. L'année 1973 marque le premier choc pétrolier et le commencement d'une hausse vertigineuse des carburants. Pierre Messmer dans son discours du 30 novembre 1973 parle *« d'accélérer le programme de réalisation de nos centrales d'électricité nucléaire …* » et annonce la construction d'une usine d'enrichissement d'uranium.

L'année 1986 est, quant à elle, marquée par une des premières grandes catastrophes nucléaires, Tchernobyl, catastrophe à l'échelle européenne, voire d'ampleur internationale. Cette catastrophe lance les débats sur les problèmes environnementaux, débats qui aboutiront à l'adoption d'un certain nombre de lois, notamment une loi relative aux recherches sur la gestion des déchets radioactifs en 1991, l'arrêt du surgénérateur Superphénix en 1997 et la création de l'Institut de radioprotection et de sûreté nucléaire en 2002. En 2005, une loi de programme fixant les orientations de la politique énergétique est adoptée.

Les présidents suivants poursuivent la politique du nucléaire, mais le débat fait toujours peur aux politiques français. Le président français Sarkozy «*… parcourt le monde à la manière du VRP d'une industrie nucléaire rutilante…* » (Schneider, 2008 : 4).

Durant la dernière campagne présidentielle française en 2012, le nucléaire est très présent dans les débats politiques. C'est en ces termes *« le progrès face au retour au Moyen-Age »* que Nicolas Sarkozy a posé le débat sur le nucléaire face à François Hollande. Pour Nicolas Sarkozy, *« notre parc nucléaire constitue une force économique et stratégique considérable pour la France. Le détruire aurait des conséquences dramatiques ».*[241] Quant à François Hollande, un de ses engagements concernant le nucléaire est le suivant : « *je préserverai l'indépendance de la France tout en diversifiant nos sources d'énergie. J'engagerai la réduction de la part du nucléaire dans la production d'électricité de 75 % à 50 % à l'horizon 2025, en garantissant la sûreté maximale des installations et en poursuivant la modernisation de notre industrie nucléaire. Je favoriserai la montée en puissance des énergies renouvelables en soutenant la création et le développement de filières industrielles dans ce secteur. La*

[239] http://www.vie-publique.fr/politiques-publiques/politique-nucleaire/chronologie/ (consulté le 08/08/2015)
[240] http://www.ina.fr, extrait d'un discours du Général de Gaulle le 25 septembre 1963 à Orange (Vaucluse).
[241] Extrait Le Monde.fr du 26 novembre 2011 par Vanessa Schneider – Pierrelatte, envoyée spéciale.

France respectera ses engagements internationaux pour la réduction des émissions de gaz à effet de serre. Dans ce contexte, je fermerai la centrale de Fessenheim en 2016 la plus vieille centrale nucléaire française encore en activité et je poursuivrai l'achèvement du chantier de Flamanville (EPR) qui devrait être opérationnelle en 2016 »[242].

Les conséquences de Fukushima sur la politique nucléaire française

Les pouvoirs publics ont demandé un audit général des installations nucléaires, mais ont réaffirmé leur confiance dans la sécurité du parc nucléaire français (Deveaux et Collignon, 2013). La filière nucléaire est toujours présentée comme un garant de notre indépendance en matière énergétique. Une interview d'Eric Besson, alors ministre de l'industrie et de l'énergie, est disponible dans notre serveur dédié, dossier Doc_Annexes : Doc-Annexe-D-nucléaire-France. Cependant, l'opposition émettait des doutes sur la fiabilité des installations face à des risques d'accidents (se reporter à l'annexe H, tableau H.1). « *Pour les uns, la sortie du nucléaire doit être programmée, pour les autres, une réorientation de la politique énergétique est nécessaire en arrêtant de prôner le tout nucléaire, en privilégiant les énergies renouvelables, (...), en favorisant l'émergence de nouvelles technologies plus propres et moins risquées (...) »* (Lacroix-Lanoë, 2013).

En 2015 qu'en est-il de la politique gouvernementale française sur le nucléaire ?

Dans un entretien à l'Usine Nouvelle[243], Ségolène Royal, ministre de l'Énergie, estime qu'*«il faut programmer la construction d'une nouvelle génération de réacteurs qui prendront la place des anciennes centrales lorsque celles-ci ne pourront plus être rénovées»*. A la question : « *Le nucléaire garde donc un avenir en France* », Ségolène Royal répond : *« L'énergie nucléaire est un atout, même si demeurent des questions sur la gestion des déchets et l'approvisionnement en uranium. Elle nous permet de réaliser la transition énergétique ... Depuis Fukushima, la demande mondiale de nucléaire a baissé, même si, dans la construction d'une économie décarbonée, le nucléaire est un atout évident… Il faut penser la demande nucléaire de manière intelligente dans un contexte de mix énergétique ...* ».

Qu'en est-il du lobby nucléaire en France ?

Le « (…) *lobby du nucléaire n'est pas transparent ni démocratique (...). Lorsque le lobby est constitué de hauts fonctionnaires, c'est l'Etat qui est son bailleur de fonds et son donneur d'ordre. Le lobby est alors plus dépendant de la puissance publique, qui décide de son budget, de ses lieux d'implantation, de son programme d'activité, et indirectement de ses effectifs, que la puissance publique n'est dépendante du lobby. Il n'y a donc pas de lobby nucléaire au sens où cette expression s'emploie habituellement (...)* » (Jancovici, 2003 : 13).

[242] http://www.parti-socialiste.fr/static/14423/les-60-engagements-pour-la-france-de-francois-hollande.pdf (consulté le 19/10/2014)
[243] L'Usine Nouvelle n°3406, Ségolène Royal : « Il faut bâtir de nouvelles centrales nucléaires », Propos recueillis par Olivier Cognasse, Ludovic Dupin et Pascal Gateaud, publié le 13 janvier 2015.

Annexe E : enquête Ifop réalisée du 7 au 10 mars 2011 et du 24 au 25 mars 2011

Une enquête Ifop a été réalisée du 7 au 10 mars 2011 auprès d'un échantillon de 1005 personnes représentatif de la population française âgée de plus de 18 ans et du 24 au 25 mars 2011 auprès d'un échantillon de 956 personnes, représentatif de la population française âgée de plus de 18 ans (méthode des quotas) (Bonneval, Lacroix-Lanoë, 2011 : 8).
La question posée est la suivante : *« Quels sont, parmi les suivants, les risques que vous jugez les plus préoccupants de nos jours ? ».*

Tableau E.1

Les risques environnementaux perçus comme les plus préoccupants (début mars 2001 vs fin mars 2011)

Risques	Avant Fukushima	Après Fukushima
Les risques liés aux changements climatiques	39% (1)	39% (2)
Les risques alimentaires	36% (2)	31% (4)
Les risques liés à la pollution des eaux	31% (3)	38 %(3)
Les risques liés à la pollution atmosphérique en ville	29% (4)	19% (6)
Les risques industriels	23% (5)	21% (5)
Les risques liés au nucléaire	18% (6)	40% (1)
Les risques liés à l'amiante	3% (7)	3% (7)

Avant le 11 mars 2011, ce sont les risques liés aux changements climatiques qui préoccupent les Français, suivis des risques alimentaires. Les risques liés au nucléaire arrivent en avant-dernière position sur les sept risques proposés par l'enquête.
Après le 11 mars, les données changent et les risques liés au nucléaire passent en première position avec 40% des réponses et deviennent une préoccupation majeure des Français. Les risques liés au changement climatique passent en deuxième position, mais demeurent toujours avec le même pourcentage de réponses.

Annexe F : la politique énergétique aux Etats-Unis

Une politique nucléaire difficile à mener face à l'hostilité de l'opinion publique américaine

L'opinion publique américaine est souvent hostile à l'utilisation de l'énergie nucléaire, une opinion qui pèse lourd lors des scrutins électoraux. Trois incidents majeurs touchant de près les Américains ont contribué à ce sentiment d'hostilité, sentiment entretenu par les médias. Le premier incident a eu lieu en 1979 à la centrale nucléaire de Three Mile Island, incident anodin puisque n'a été constaté ni décès, ni pollution radioactive, mais les Américains n'ont retenu que : « *Nuclear is no safe* » (Chavardès, 2009). Le deuxième, conséquence du premier, concerne le renforcement de la sûreté de toutes les centrales existantes, en construction, et en projet. « *Le coût du kWh nucléaire augmente et n'est plus compétitif » :* « *Nuclear is not competitive* » (Chavardès, 2009). La troisième raison concerne le problème de stockage et de traitement des combustibles irradiés. Le projet de stockage de Yucca Mountain est finalement abandonné, alors « *que l'étude de sûreté avait abouti laborieusement à son terme* » (Chavardès, 2009).

Mais l'opinion américaine commence à bouger, d'une part en prenant conscience des risques de changements climatiques comme l'augmentation en fréquence et en violence des tornades dans le sud-est des Etats-Unis et l'ouragan Katrina, d'autre part, les envolées des prix du pétrole et du gaz mettent en évidence l'importance du problème des approvisionnements énergétiques. Durant la campagne présidentielle, Barack Obama reconnaissait « *que le recours à l'électronucléaire devait faire partie d'une politique énergétique diversifiée permettant de réduire à terme les émissions de CO2 des Etats-Unis* ».

D'ailleurs, les Etats-Unis sont devenus un des deux pays avec la Chine, les plus pollueurs de la planète et sont conscients que quelque chose doit être fait à propos du réchauffement climatique. Ils n'avaient pas ratifié le protocole de Kyoto, mais ils adhèrent désormais à la COP et en deviennent un membre actif (COP21).

En matière d'énergie, ce sont plutôt les énergies renouvelables qui seront privilégiées. « *Les investisseurs, compagnies électriques et banques, se montrent prudents avant de se lancer dans des projets coûteux de construction de nouvelles centrales* » (Chavardès, 2009).

Quant au lobby nucléaire américain, « *il exerçait des pressions auprès des politiques, des médias mais aussi auprès de l'opinion publique très fortes dans les années 1970* ». Dans ces années glorieuses, quatre groupes industriels de constructeurs se partageaient les marchés, « *General Electric, Westinghouse, Combustion Engineering, Babcok&Wilcox, mais aujourd'hui, un seul constructeur américain General Electric a survécu, Westinghouse est la propriété de Toshiba, Combustion-Engineering a disparu après avoir transféré sa technologie au Coréen Doosan et la partie de Babckock&Wilcox/centrales nucléaires, a été rachetée par Framatome ; le lobby nucléaire purement américain n'a donc plus l'impact qu'il pouvait avoir autrefois sur l'Administration et le Congrès* » (Chavardès, 2009). « *L'opinion publique américaine semble résignée à accepter le retour de l'électronucléaire, mais ce sont les conditions pratiques et notamment financières qu'exige cette relance qui ne sont pas au rendez-vous* » (Chavardès, 2009).

Annexe F

Les Etats-Unis n'ont pas ratifié le protocole de Kyoto, mais désormais la question du réchauffement climatique préoccupe les politiques et les Américains. « *Depuis l'arrivée de Barack Obama au pouvoir, l'énergie et l'environnement sont devenus des secteurs prioritaires* » (Méritet, 2009 ; Le Leuch, 2010). L'organisation politique américaine est un peu complexe et les actions d'ensemble sur le réchauffement climatique restent difficiles à mener au niveau fédéral. Certaines mesures sont cependant prises au niveau des Etats (Massachusetts et Etat de Washington). Pour avoir plus de détails sur les principaux acteurs publics et privés dans le domaine des énergies ainsi qu'un bref rappel historique du nucléaire, se reporter dossier Doc_Annexes : Doc_Annexe-F-énergie-EtatsUnis disponible dans le serveur dédié.

« *Depuis 2009, de nombreux débats animent la vie politique autour du projet de loi Climate Change Bill . Ce projet oblige les Etats-Unis à réduire les émissions de gaz à effet de serre de 17% en 2020 par rapport à 2005* » (Méritet, 2009).

La transition énergétique

En ce qui concerne la transition énergétique aux Etats-Unis un bref exposé est disponible dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-F-énergie-EtatsUnis.

Lien du serveur dédié :
https://drive.google.com/folderview?id=0B8XHfHwNzWAAeDN6UEc1b2dGa1U&usp=sharing

Annexe G : la politique énergétique chinoise

La politique énergétique chinoise couvre de larges domaines, tant sur l'exploitation des sources d'énergie, de leur utilisation, de leur transport que sur le plan politique avec les choix et les orientations à donner, le tout s'insérant dans un contexte mondial.
Commençons par rappeler quelques chiffres (Boquet, 2009 ; Chavardès, 2004, 2010 ; OCDE, 2013 : 70) : une population de 1,351MM en 2012 dont un quart est citadine, une consommation énergétique qui ne cesse d'augmenter sachant que la croissance économique chinoise repose toujours sur l'utilisation massive des énergies fossiles, comme le charbon (Martin-Amouroux, 2004), première ressource énergétique du pays et première source d'émissions de CO2 dans l'atmosphère.

La Chine construit sa politique énergétique sur un dilemme, l'ouverture et l'autosuffisance énergétique, tout en devant s'adapter aux exigences internationales[244] comme la réduction des gaz à effet de serre (GES). Les investissements, notamment dans le nucléaire et l'énergie éolienne, pourraient prolonger encore quelques temps cette autonomie énergétique[245] (Domergue, 2005 ; Wang et Raes, 2012 ; Liebermann, 2014).
Une autre source de production de GES est l'augmentation permanente du nombre d'automobiles. Actuellement, l'estimation du parc automobile chinois avoisine les 100 millions d'unités, avec une prévision de quelques 200 millions de véhicules d'ici 2020 (Keppler et Méritet, 2004 ; Madslien, 2012). A Pékin, ce sont 2 000 nouveaux véhicules mis en circulation tous les jours qui provoquent une hausse de la pollution (Dong, 2007). La voiture représente le symbole de réussite sociale, c'est le « *cheng jiu gan* »[246]. Les Chinois, qui achètent une voiture, ne sont pas certains qu'elle sera immatriculée rapidement, il faut attendre de longs mois, voire des années avant que celle-ci puisse circuler.

Politique de coopération

La Chine poursuit une politique de coopération avec des pays émergents, comme par exemple le Bangladesh ou le Pakistan, mais également le Vietnam, le Zimbabwe, le Botswana, etc. D'autres partenariats sont signés avec l'Afrique du Sud à propos de la sécurité minière et de la production de carburants synthétiques. Elle participe également avec d'autres pays à la recherche de nouvelles *clean coal* technologies (Locatelli et Martin-Amouroux, 2005).

L'énergie nucléaire, un cas particulier

Dans les années 1955, la Chine s'est lancée dans le nucléaire militaire, mais aujourd'hui, le nucléaire civil est devenu la priorité. La Chine a choisi d'élaborer un programme nucléaire ambitieux (Wang et Raes, 2012), mais cette énergie n'a pas l'ambition de se substituer à d'autres formes d'énergie, en particulier le charbon (Tonnac de et Pervès, 2012). La politique chinoise de développement de l'énergie nucléaire s'accélère au début du 11ème Plan Quinquennal (2006-2011). En 2007, la production nucléaire est encore limitée à six implantations sur trois sites côtiers, Dayawan, Qinshan et

[244] La Chine adhère à la COP, un des objectifs de la COP21 (décembre 2015) est de limiter la hausse des températures à 1,5°C.
[245] Le Monde.fr, Florent D., « Politique énergétique chinoise : l'heure des choix », publié le 25/03/2015, http://www.lemonde.fr/idees/chronique/2010/03/26/politique-energetique-chinoise-l-heure-des-choix_1324520_3232.html (consulté le 10/08/2015)
[246] RFI, Stéphane Lagarde, correspondant à Pékin, l'eldorado chinois de l'automobile, 17/10/2010, http://www.rfi.fr/asie-pacifique/20101017-eldorado-chinois-automobile/ (consulté le 30/03/2015)

Tianwan (se reporter au chapitre 6 et à l'annexe J). Les objectifs de développement fixés pour la croissance de l'énergie nucléaire sont revus à la hausse en 2010, à l'occasion du 12ème Plan Quinquennal (Zhang, 2011).

Le président Xi Jinping déclare «*le pays adhère à l'usage pacifique et sûr de l'énergie nucléaire (...)* » et s'engage dans une lutte contre les émissions de CO2 et la pollution de l'air, quant au premier ministre Li Keqiang, *« Nous travaillerons à faire de notre pays un acteur de taille de l'industrie nucléaire »*.

Le gouvernement chinois avait interrompu la construction de nouveaux réacteurs après la catastrophe japonaise de Fukushima en 2011. Le renforcement de la sûreté nucléaire avec l'aide de la coopération internationale nécessite des investissements équivalant à 10 milliards euros d'ici 2020 (Laramée de Tannenberg, 2012). Mais une fois l'évaluation de la sûreté des réacteurs effectuée, la construction des réacteurs a repris[247] et les projets nucléaires aussi. Les débats publics consacrés à l'atome restent très limités (Lafargue, 2013 : 24) et la part de l'électricité nucléaire chinoise est en constante progression (se reporter au dossier, Doc_Annexes, Doc_Annexe-C-nucléaire-monde.docx, disponible sur notre serveur dédié[248]).

Un autre objectif, devenu prioritaire, est la modernisation des centrales de 2ème génération avec l'intégration des nouvelles normes de sécurité et de sûreté nucléaire suite à l'accident de Fukushima, 80 milliards de yuan (13 milliards de dollars) seront investis sur les trois prochaines années. Le développement nucléaire reste une priorité importante en Chine et absorbe environ 20% des investissements du secteur de l'énergie à fin 2012.

[247] Les ambitions de la filière nucléaire chinoise se précisent, publié le 19/01/2015, https://www.lenergieenquestions.fr/les-ambitions-de-la-filiere-nucleaire-chinoise-se-precisent/ (consulté le 12/04/2015)

[248] Lien du serveur dédié : https://drive.google.com/folderview?id=0B8XHfHwNzWAAeDN6UEc1b2dGa1U&usp=sharing

Annexe G

Quelques mots sur les réacteurs chinois

Brossons un rapide tableau de ces différents types de réacteurs (Jaouen et Béroux, 2012 : 113 ; Laponche, 2015).

Génération II : CPR-1000

Le CPR-1000 *(REP chinois amélioré) est un réacteur à eau pressurisée de génération II+, basé sur la conception du réacteur français REP de 900 MWe de Framatome (AREVA), amélioré pour atteindre une puissance électrique nette de 1 000 MWe et une durée de vie de 60 ans.*

Les quatre réacteurs à eau pressurisée en fonctionnement à la centrale nucléaire de Daya Bay, et à la centrale nucléaire de Ling Ao sont parfois dénommés CPR-1000, mais ils sont plus proches de la conception du REP de 900 MW français, avec une puissance nette inférieure à 1 000 MW et constitués essentiellement de composants importés. Ces quatre réacteurs ont été fournis par Framatome (AREVA).

Le CPR-1000 est construit et exploité par l'entreprise China General Nuclear Power Group (CGN), dénomination depuis 2013 de l'entreprise China Guangdong Nuclear Power Company (CGNPC), fondée en 1994. Pour la seconde tranche, 70 % des équipements sont fabriqués en Chine, avec un objectif de 90 % à terme. Le CPR-1000 est en développement rapide avec 15 tranches en construction en juin 2010. Le 15 juillet 2010, le premier CPR-1000 chinois, Ling Ao-3, est connecté au réseau. Avec CNNC et CGN, la troisième compagnie possédant la licence d'exploitation des centrales nucléaires en Chine est China Power Investment Corporation (CPIC).

Génération III importée

EPR d'AREVA et EDF *: en décembre 2009, AREVA et CGN (alors CGNPC) ont créé une coentreprise pour le développement de l'EPR en Chine, la Wecan JV, 55 % CGN et 45 % AREVA. Deux réacteurs EPR sont en construction à Taishan, d'une puissance électrique nette unitaire de 1 660 MW (début de construction respectivement en 2009 et 2010 ; des retards de l'ordre d'un à deux ans pour la date de démarrage sont actuellement signalés). En août 2008, EDF et CGN (alors CGNPC) ont créé une coentreprise Guangdong Taishan Nuclear Power Venture Company Limited (TNPC) pour une période de 50 ans (durée maximale pour une « joint-venture » en Chine) à 30 % EDF et 70 % CGN. TNPC est maître d'oeuvre pour la construction de la centrale (Taishan 1 et 2), en est propriétaire et l'exploitera.*

AP-1000 de Westinghouse : *Quatre exemplaires de l'AP1000, de puissance électrique nette de 1 000 MW, sont déjà en construction en Chine depuis 2009 (centrales de Haïyang et Sanmen). Des retards de l'ordre de deux ans de la date de démarrage sont actuellement signalés (problèmes de sûreté), ainsi qu'une augmentation des coûts. Un troisième acteur majeur dans le développement du nucléaire en Chine est State Nuclear Power Technology Corporation (SNPTC), créé en 2007. Westinghouse a consenti à transférer à SNPTC la technologie des quatre premiers AP-1000 construits en Chine, ce qui lui permettra de construire les suivants de façon indépendante.*

Génération III chinoise

ACPR-1000 de CGN : *ACPR-1000, (Advanced CPR-1000), est un modèle développé par CGN avec des partenaires chinois depuis 2009 ; c'est un REP avec double enceinte de confinement et récupérateur de corium. Ce modèle est dégagé de toute dépendance étrangère.*

ACP-1000 de CNNC : *la conception du type de réacteur avancé développé en Chine, l'ACP-1000 de CNNC, a passé avec succès l'examen générique de la sûreté des réacteurs de l'AIEA. L'Agence est ainsi arrivée à la conclusion que l'ACP-1000 était sûr et fiable. L'ACP-1000 est le premier type de réacteur développé en Chine ayant été soumis à un examen international.*

ACC-1000 ou Hualong-1 : *initialement, CNNC souhaitait utiliser son propre type avancé ACP-1000 pour les tranches en projet Fuqing 5 et 6. Mais en 2012, les autorités nationales ont demandé à CNNC et CGN, de rapprocher leurs programmes de développement nucléaire. L'ACP-1000 de CNNC et l'ACPR-1000 de CGN ont ainsi été regroupés pour devenir le type Hualong-1 (également appelé Hualong-1000 ou ACC-1000). Les réacteurs Hualong-1 seront utilisés pour la première fois dans le cadre des nouvelles constructions Fuqing 5et 6. Les réacteurs ACC-1000 possèdent des cycles de combustible compris entre 18 et 24 mois ainsi qu'une disponibilité élevée. Le réacteur contiendra probablement 177 assemblages combustibles et sera entouré de deux enceintes de confinement. Il sera en outre équipé de systèmes de sécurité actifs et passifs. Sa durée de vie escomptée est de 60 ans. En novembre 2014, un groupe d'experts est arrivé à la conclusion que la conception de l'ACC-1000 remplit toutes les exigences de sécurité de la troisième génération. La CNNC a informé que la conception pourra donc être exportée.*

CAP-1400 : le réacteur nucléaire chinois de 3ème génération avancée

En 2008 et 2009, Westinghouse a conclu des accords pour travailler avec la société d'État Nuclear Power Technology Corporation et d'autres instituts pour concevoir une version plus puissante que l'AP-1000, le réacteur CAP-1400, suivie éventuellement d'un modèle de puissance électrique 1 700 MW. La Chine possédera les droits de propriété intellectuelle pour ces designs plus grands, qui pourraient aussi être exportés ailleurs avec la coopération de Westinghouse. Ce réacteur fait partie des seize projets stratégiques inclus dans le plan de développement scientifique et technologique national. Les travaux préparatoires sur le site de construction de deux tranches de démonstration du CAP1400, dans la province chinoise de Shandong, sont déjà en cours. Cependant, la construction du premier réacteur CAP1400 est repoussée au plus tôt à 2015 ou 2016. Pour l'export, SNPTC serait associé à CNNC.

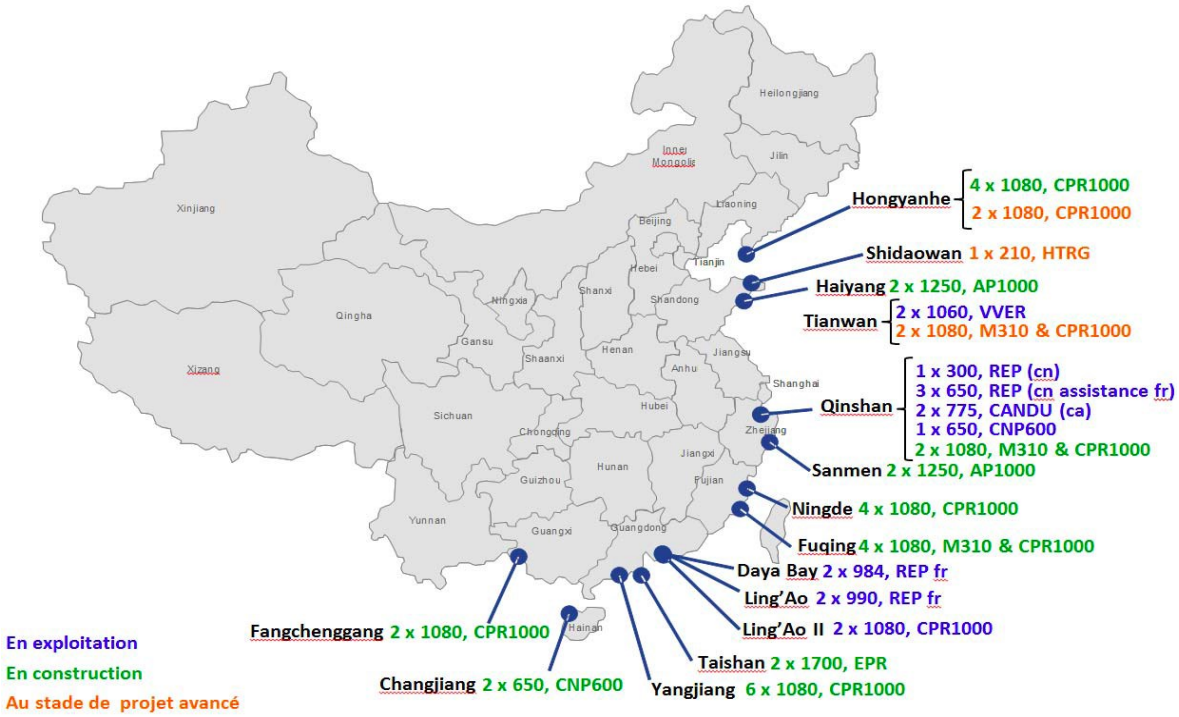


Figure G.2
Implantation des centrales nucléaires (de Tonnac et Pervès, 2012)

Une lente transition énergétique

Poursuivant les objectifs définis dans ses plans quinquennaux de *décarboner* sa production énergétique, l'Empire du Milieu est le premier marché en terme d'investissement dans les énergies vertes, il y a consacré 89,5 milliards de dollars en 2014 (Nodé-Langlois, 2015), soit une hausse spectaculaire de 32 % sur un an[249].

Les énergies renouvelables

Quelques éléments d'information concernant ces nouvelles énergies sont disponibles sur notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-G-énergie-Chine.

La pollution, les gaz à effet de serre, le réchauffement climatique

Les plans quinquennaux fixent des objectifs et des mesures concrètes commencent à être mises en place (Delbosc, 2011 ; Zhou et Delbosc, 2013).

Le 11ème plan quinquennal (2006-2010) comportait déjà plusieurs programmes ambitieux pour économiser l'énergie et diminuer les émissions de gaz à effet de serre. Une partie des actions était décidée au niveau local dans le cadre de la politique définie par le gouvernement central (Price, Levine et al, 2011). Le 31 janvier 2010, la Chine a fait part de ses propositions d'actions à l'horizon 2020 suite à l'Accord de Copenhague (Béduneau-Wang, Shan et al, 2010).

Le 12ème plan quinquennal laisse une place très importante aux enjeux environnementaux, d'une part réduire l'intensité en carbone de 17% par rapport au niveau de 2010 et d'autre part amener la part des énergies non fossiles à 11,4 %. La nouvelle équipe de Xi Jinping a déclaré la guerre à la pollution. Les dirigeants doivent faire évoluer les mentalités chinoises pour permettre à l'emploi à bon escient de

[249] http://www.lefigaro.fr/conjoncture/2015/01/10/20002-20150110ARTFIG00021-la-chine-devient-le-plus-gros-investisseur-dans-le-solaire-et-l-eolien.php (consulté le 15/07/2015).

l'énergie et le respect de l'environnement. *«L'ampleur de la pollution a réveillé les esprits, et il y a une forte prise de conscience de la société civile»,* souligne Chen Yan, directeur général du Forum China-Europa. Cette prise en compte du changement climatique a été confirmée le 5 mars 2012 lors des sessions de l'Assemblée nationale populaire (ANP) et de la Conférence consultative politique du peuple chinois (CCPPC).
Pour réduire les émissions de CO2, les deux options retenues sont le développement des énergies renouvelables, puis l'amélioration de l'efficacité énergétique[250].
Afin d'améliorer l'accès du pays aux technologies de pointe, Pékin a élaboré une stratégie de coopération internationale fondée sur deux concepts : « go global » [走出去] et « bringing in » [引进来] [251] .
Le problème de la pollution de l'air dans la capitale chinoise augmente sans cesse provoquant un brouillard fréquent, les autorités du pays ont fait fermer deux des principales centrales à charbon de Pékin. L'objectif est de réduire la pollution de l'air, mais celle-ci reste trop élevée avec une moyenne de pollution supérieure à dix fois le niveau recommandé par l'Organisation Mondiale de la Santé (Leplâtre, 2015).
Les centrales à charbon vont être remplacées progressivement par des centrales à gaz, moins nocives pour la population locale. La fermeture de la dernière centrale à houille de Pékin est d'ailleurs prévue pour 2016 (Leplâtre, 2015).
Le gouvernement de Xi Jinping s'est engagé avec les États-Unis à continuer la lutte contre l'effet de serre, avec un objectif de réduction de 26% entre 2005 et 2025. A l'heure actuelle, seules 8 des 74 villes chinoises surveillées par le Ministère pour la protection de l'environnement répondent aux normes requises pour la qualité de l'air en 2014 et 7 des 10 villes les plus polluées de Chine se trouvent dans la province de Hebei entourant Pékin.

La diplomatie et les partenariats
Aujourd'hui la Chine est une des plus grandes puissances au monde. Elle s'est ouverte vers l'extérieur et entretient désormais de nombreux partenariats notamment avec des pays fournisseurs d'énergie (Le Corre, 2006 ; Robert, 2010 ; Tan et al, 2013).
Des liens privilégiés sont entretenus avec les pays de l'Asie centrale dans les secteurs pétrolier et gazier (Chaudet, 2013 ; Fazilov et Chen, 2013). Mais des accords ont également abouti dans d'autres domaines, comme celui de la finance (Chambon et al, 2010 ; Lenglet, 2010).
Un résumé sur ce thème est disponible dans notre serveur dédié, dossier Doc_Annexes : Doc_Annexe-G-énergie-Chine.

Lien du serveur dédié :
https://drive.google.com/folderview?id=0B8XHfHwNzWAAeDN6UEc1b2dGa1U&usp=sharing

[250] Institut des Hautes Etudes pour la Science et la Technologie, « Les grandes tendances de la politique énergétique chinoise », date de mise en ligne le 21/11/2013, http://www.ihest.fr/IMG/article_PDF/article_a904.pdf (consulté le 01/04/2015)
[251] Il s'agit d'une stratégie industrielle coordonnée dotée de moyens financiers importants visant à accroître la capacité d'innovation de ses compagnies énergétiques et son niveau d'autonomie dans les équipements énergétiques de pointe. L'application de cette stratégie n'est pas toujours facile, face au refus des principaux détenteurs de la technologie occidentaux de partager leurs brevets ou de vendre leurs équipements de pointe par crainte de l'espionnage industriel (Alexeeva et Roche, 2014 : 2).

Annexe H: dictionnaire d'événements et restitutions par poly-cooccurrences des formes-pôles *nucléaire* et *énergie* pour le sous-corpus ENRG_FR pour la période du 24 septembre 1999 au 17 avril 2012

H.1 Forme-pôle *nucléaire* : constats pour ENRG_FR

L'illustration des poly-cooccurrences pour ENRG_FR se trouve dans la figure H.1 ci-après.

Afin de détecter un maximum d'informations, à l'aide de l'outil Trameur, les paramètres par défaut pour tous les calculs de poly-cooccurrences ont été maintenus : co-fréquence 2, seuil 10, contexte . !?

Le contexte . !? signifie approximativement dans le langage courant phrase.

Le retour au contexte s'avère indispensable pour attester toute affirmation présentée dans nos recherches ci-après.

Graphe N°1, poly-cooccurrences sur La forme « nucléaire »

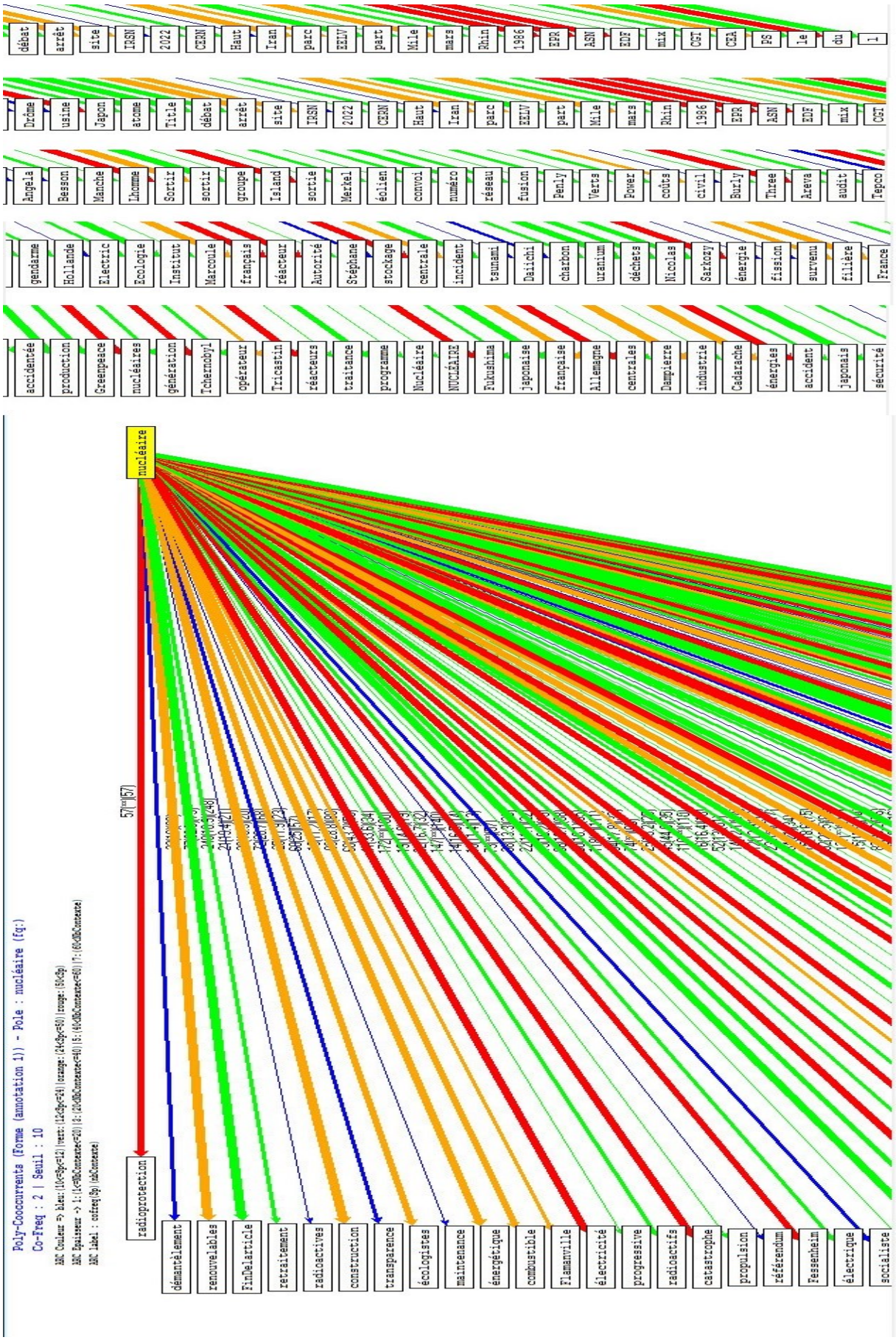


Figure H.1

ENRG_FR : poly-cooccurrences autour de la forme-pôle *nucléaire*

Après une première analyse de la figure H.1, ci-dessus, nous constatons notamment une forte présence de formes poly-cooccurrentes se rapportant à des toponymes, à la politique nucléaire, à des activités liées au domaine nucléaire, etc. Quelques explications sur le contenu de ces formes seront expliquées ci-après.

Les toponymes

Ces formes poly-cooccurrentes sont souvent associées à des sites français nucléaires, mais aussi à quelques sites étrangers. Ces lieux font souvent la Une des médias, soit de manière ponctuelle, soit de manière récurrente, pour des problèmes liés à des incidents de production voire à des accidents majeures ou à des catastrophes.

Flamanville *:* commune du département de la Manche, où sont installées deux centrales nucléaires, l'unité 1 a été construite sur une ancienne mine de fer sous-marine, mise en service en mars 1985 et l'unité 2 en novembre 1986. Chaque réacteur a une puissance de 1300 MW. Flamanville 3(EPR) une centrale de nouvelle génération en cours de construction, l'EPR (European Pressurized Reactor), construction qui a débuté en décembre 2007 et se poursuit aujourd'hui[252] (se reporter au tableau 5.14 du chapitre 5, les EPR dans le monde). Le coût de construction[253] de l'EPR dépasse largement les prévisions : « *Le groupe énergétique français EDF a annoncé, lundi 3 décembre 2012, avoir relevé de 2 milliards d'euros son estimation du coût de la construction du réacteur pressurisé européen (EPR) de Flamanville, portée à 8,5 milliards, inflation comprise ».*
Formes poly-cooccurrentes associées à *Flamanville* constatés dans les calculs de poly-cooccurrences : *centrale*, *nucléaire, incident*, *Manche, construction*, *EPR*, *Génération* (les réacteurs sont typés en fonction de leurs dates de mise en construction et de leurs systèmes de production génération 1 ou 2 ou 3 ou 4 (à l'horizon 2040 pour ce dernier), *coûts, Greenpeace* (des actions contre la construction de l'EPR sont menées périodiquement)[254], *site*, *numéro, réacteurs, réacteur, EDF, électrique.*

Fessenheim : centrale située dans le département du Haut-Rhin, « *a l'originalité d'avoir également des actionnaires allemands et suisses. Elle est constituée de 2 réacteurs à eau sous pression d'une puissance de 900 MW* »[255]. La plus ancienne des centrales françaises (mise en service en 1978) cumule les incidents ; en 2009, des manifestations se déroulaient pour fermer cette centrale.[256] Un des engagements de François Hollande durant la campagne présidentielle 2012 était sa fermeture en 2015, mais cette date sera semble-t-il décalée du fait du recul de la mise en service de l'EPR Flamanville.
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *centrale, site, nucléaire, fermeture, Hollande, démantèlement, incident, Rhin, arrêt.*

[252] http://energie.edf.com/nucleaire/carte-des-centrales-nucleaires-45738.html (consulté le 18/03/2015).
[253] Le Monde : « Le coût de l'EPR de Flamanville revu encore à la hausse », 03/12/2012, http://www.lemonde.fr/planete/article/2012/12/03/le-cout-de-l-epr-de-flamanville-encore-revu-a-la-hausse_1799417_3244.html (consulté le 18/10/2014)
[254] Source : Greenpeace, http://energie-climat.greenpeace.fr/epr-de-flamanville-greenpeace-bloque-le-chantier (consulté le 18/10/2014)
[255] Source : ASN, centrale nucléaire de Fessenheim, http://www.asn.fr/L-ASN/ASN-en-region/Division-de-Strasbourg/Centrales-nucleaires/Centrale-nucleaire-de-Fessenheim/(rub)/112582
[256] http://www.sortirdunucleaire.org/Dossiers?id_mot=72 (consulté le 15/10/2014).

Tricastin : centrale crée en 1980, connue pour ses incidents à répétition, notamment pendant l'été 2008, *« (...) un incident a eu lieu, lundi 8 septembre à 10 h 30, pendant un arrêt de tranche à la centrale nucléaire a indiqué un porte-parole de l'Autorité de sûreté nucléaire (...) ».*[257]
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *centrale, nucléaire, incident*, *Drôme*, *réacteurs,* réacteur, *site.*[258]

Dampierre-en-Burly (Loiret) et ***Penly*** (dans le pays de Caux) : mises en service en 1990 et 1992, ces deux sites abritent des centrales nucléaires exploitées par EDF, centrales qui connaissent des incidents à répétition.
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *centrale, nucléaire, incident, Burly*, *puissance, réacteurs, réacteurs, EPR.*

Cadarache : créé en octobre 1959, un des dix centres de recherche du Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), un centre important de recherche et développement technologique pour l'énergie en Europe[259]. Ce site est associé au programme de recherche ITER (projet de réacteur expérimental à fusion thermonucléaire - *International Thermonuclear Experimental Reactor*).[260]

Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, CEA, ITER*.

Marcoule : centre de recherches du CEA, a été associé à de nombreux programmes de recherche et de développement sur le cycle de combustible (par exemple le programme Phénix ou encore sur la fabrication de combustible MOX)[261].
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, CEA, uranium*.

Les catastrophes

Des sites étrangers apparaissent sur la figure H.1, ci-dessus, sites liés à des catastrophes de grande ampleur ayant une répercussion mondiale.

Three Mile Island : site aux Etats-Unis où s'est produite une des premières catastrophes le 28 mars 1979, catastrophe classée au niveau 5 sur l'échelle INES[262].
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, accident, réacteurs, catastrophe, mars, sites.*

Tchernobyl **:** deuxième accident, classée au niveau 7 sur l'échelle INES, survenu le 26 avril 1986 dans la centrale nucléaire de Tchernobyl, en Ukraine, a profondément marqué l'opinion publique européenne. « *Cet accident entraîne d'importants rejets radioactifs dans l'atmosphère, qui se dispersent au gré du trajet des masses d'air. Le panache radioactif ainsi formé finit par couvrir une bonne partie de l'Europe au cours des journées suivant l'accident. En France, les doses reçues par la*

[257] http://www.lemonde.fr/societe/article/2008/09/09/nouvel-incident-a-la-centrale-nucleaire-du-tricastin_1092996_3224.html (consulté le 18/10/2014)
[258] http://www.areva.com/FR/activites-852/le-site-nucleaire-du-tricastin-un-site-industriel-unique-en-europe.html (consulté le 18/10/2014)
[259] http://www.cea.fr/le-cea/les-centres-cea/cadarache (consulté le 18/10/2014)
[260] http://www.cadarache-communication.fr/communication/tag/fusion/ (consulté le 18/10/2014)
[261] http://www.cea.fr/le-cea/les-centres-cea/marcoule (consulté le 10/10/2014)
[262] INES : en anglais International *Nuclear Event Scale*, en français échelle internationale de classement des accidents et des incidents nucléaires

population en 1986 sont très faibles. C'est plus particulièrement la thyroïde qui est exposée, par ingestion d'iode 131 présent dans les aliments (...) Ce sont les enfants qui vivent dans l'Est de la France en 1986, territoire le plus touché par les retombées, qui reçoivent les doses les plus élevées ».[263]

Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, accident, réacteurs, catastrophe, 1986.*

Fukushima : le 11 mars 2011, un séisme de magnitude 9 se produit à 80 km au large de l'île d'Honshu au Japon. Ce séisme provoque un tsunami qui touche la côte nord-est du Japon, tsunami qui prive les centrales nucléaires de ses sources externes d'électricité et de ses moyens internes de refroidissement du cœur des réacteurs nucléaires de Fukushima-Daiichi et de Fukushima Daini, un niveau de gravité 6 sur l'échelle INES[264]. Naoto KAN, alors premier ministre japonais, déclare quelques mois plus tard en ces termes la catastrophe nucléaire de Fukushima résultat de deux causes majeures : *« (...) Inutile de le dire, la cause première a été la coupure totale de courant à Fukushima Daiichi, causée par l'énorme séisme et tsunami, les plus forts jamais survenus dans l'histoire du Japon. Toutefois, il y avait effectivement une autre cause majeure. Une telle coupure totale de courant et un tsunami aussi puissant n'ont jamais été anticipés. Aucun préparatif à une telle situation n'a jamais été fait en termes d'installations physiques ou de structure de la communication au sein du gouvernement. C'était, en d'autres termes, une cause d'origine humaine ... »* (Kan, 2013)
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, accident, réacteurs, catastrophe, Daiichi, tsunami, Japon, japonaise, réacteurs, réacteur.*

Daiichi : groupe pharmaceutique d'origine japonaise, crée en 1899, l'entreprise concentre ses activités sur la recherche et la commercialisation de médicaments innovants. Mais Daiichi possède des réacteurs et son complexe atomique a fait l'objet de diffusions d'images-chocs montrant les réacteurs vomissant des panaches radioactifs à la suite du tsunami et de Fukushima.
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : n*ucléaire, accident, tsunami, Fukushima, catastrophe, japon, japonaise, réacteurs, électricité.*

Les formes poly-cooccurrentes se rapportant aux activités du nucléaire

Sécurité : réactivé suite à l'accident du 11 mars 2011. Tchernobyl faisait figure d'accident exceptionnel, mais le renouvellement d'un nouvel accident d'une telle ampleur a relancé le débat sur la sécurité et la sûreté des centrales nucléaires, notamment, en France, où notre politique énergétique est tournée vers le nucléaire. Pour les Français c'est la sûreté des installations et des infrastructures qui est primordial. Toutes sortes d'actions sont menées après Fukushima, audits, débats,, sondages, comme par exemple dans le livre blanc réalisé sur la sûreté nucléaire en Cotentin[265]. Les autres pays européens ont réagi différemment en fonction de leurs préoccupations. Différentes enquêtes ont été menées auprès de cinq états européens, France, Allemagne, Royaume-Uni, Espagne et Italie

[263] http://www.irsn.fr/FR/connaissances/Installations_nucleaires/Les-accidents-nucleaires/accident-tchernobyl-1986/consequences-homme-environnement/Pages/1-Le_scenario_de_l_accident.aspx?dId=257bc933-f16a-4c99-819e-ca843037559c&dwId=45d76c24-3232-45e5-9089-89cf5d3e2b35#.VEI2-ktquw1
[264] Le Monde, « L'ASN classe l'accident nucléaire de Fukushima au niveau 6 », 15/03/2011, http://www.lemonde.fr/japon/article/2011/03/15/l-asn-classe-l-accident-nucleaire-de-fukushima-au-niveau-6_1493498_1492975.html (consulté le 18/10/2014).
[265] http://www.rmandie.com/news/n/_La_surete_nucleaire_au_crible_d_elus_locaux_et_d_ecologistes52051220131653.asp (consulté le 18/10/2014).

(Bonneval, Lacroix-Lanoë, 2011). La première enquête[266] menée, porte sur le critère de choix prioritaire dans le choix énergétique d'un pays. La question est la suivante : *« en matière de politique énergétique selon vous, parmi les suivants, quel est le critère de choix qui doit être prioritairement pris en compte ? ».*

Tableau H.1

Résultats sur les critères de choix en matière de politique énergétique

(France, UK, Allemagne, Italie, Espagne)

	France	UK	Allem.	Italie	Espagne
La sûreté des installations, des infrastructures	**25%(1)**	19%	15%	21%	16%
Le caractère de la source d'énergie	21%	**24%(1)**	**27%(1)**	**30%(1)**	**24% (1)**
L'indépendance énergétique qu'il procure au pays	19%	18%	16%	21%	20%
Le prix de revient	13%	10%	8%	9%	20%
L'impact sur le réchauffement climatique	11%	12%	13%	11%	11%
La sécurité d'approvisionnement : absence de coupures, de ruptures d'approvisionnement	11%	17%	21%	8%	9%

En ce qui concerne l'enjeu de la sécurité, l'opinion publique allemande s'avère une nouvelle fois plus critique qu'en France et qu'au Royaume-Uni. D'autres enquêtes ont été conduites en Europe, notamment sur le thème « les Européens et la sûreté nucléaire » (Lesourne, 2008), bien avant Fukushima.

Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : n*ucléaire, accident, stockage, retraitement, maintenance, audit.*

Coût : plusieurs interprétations possibles pour cette forme : soit le coût de l'électricité produite par le nucléaire, « *soit le coût démentiel de la construction de l'EPR et de l'ITER, soit le fiasco de Superphénix, soit le coût de démantèlement d'une centrale* » (Bourry, 2012) pouvant se découper de la façon suivante :

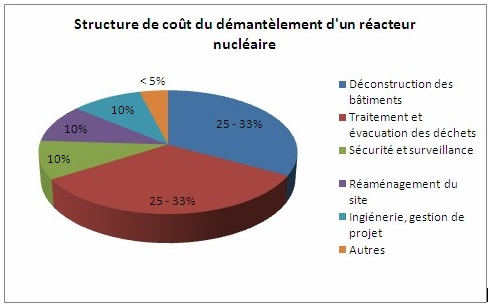


Figure H.2

Structure du coût de démantèlement d'un réacteur nucléaire[267]

[266] Enquête IFOP/le Monde réalisée du 21 au 27 juin 2011 auprès d'échantillons représentatifs des populations française (1006 personnes), allemande (603), espagnole (600), italienne (605), britannique (604) âgées de 18 ans et plus (méthode des quotas).

[267] Source : « Politiques, stratégies et coûts de démantèlement : un tour d'horizon international », AEN, décembre 2003.

Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, électricité, EPR, construction.*

Convoi : transport des déchets radioactifs et autres matières dangereuses. Les écologistes en font un peu leur cheval de bataille et c'est toujours sous « haute surveillance » que ces transports s'effectuent, transports souvent fortement médiatisés. Par exemple, l'automne 2010 a été marqué par des incidents entre manifestants antinucléaires et forces de l'ordre à l'occasion du convoi ferroviaire de 123 tonnes de déchets nucléaires (Bonnaure, 2011) entre la Hague dans le Cotentin et Dannenberg en Allemagne[268]. Un autre exemple : « *le 5 février 2013, un train en provenance des Pays-Bas transportant des déchets nucléaires destinés au centre de retraitement d'Areva de La Hague (Manche) a été immobilisé par des militants écologistes dans la banlieue de Lille. ... Le train a été d'abord arrêté à l'aide d'un fumigène lancé sur la voie, puis de nouveau un peu plus loin lorsque plusieurs militants se sont mis en travers de la voie. Les écologistes veulent montrer qu'on pouvait arrêter le train où on voulait ... »*[269]. En France, *« aux termes de la loi, les combustibles irradiés ne sont pas considérés comme des déchets car ils sont considérés comme potentiellement réutilisables. C'est l'exception française* »[270].
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, accident, Greenpeace, écologistes, EELV, Manche, déchets*.

Les formes poly-cooccurrentes associées aux mouvements écologistes

Les années 1970 voient l'apparition de l'écologie sur la scène politique avec la candidature de René Dumont à la présidentielle en 1974. Des candidats écologistes réalisent les premières percées, mais celles-ci restent localisées. Les luttes de terrain constituent l'essentiel de l'action : contre le nucléaire civil (Plogoff) et militaire (Larzac), contre les pollutions maritimes (Amoco Cadiz), contre les équipements inutiles et (projet de canal Rhin-Rhône)[271].

Verts et EELV : les « Verts » est un mouvement fondé en 1982, également l'ancien nom d'un parti politique et « *EELV*, Europe, Ecologie, Les Verts » désigne maintenant le nouveau nom du parti écologiste français.

Greenpeace[272] : ONG non violente, indépendante et internationale de protection de l'environnement, présente sur tous les continents et tous les océans grâce à ses 28 bureaux nationaux et régionaux et ses trois bateaux. La mission essentielle de Greenpeace est de travailler par rapport à des enjeux globaux (climat, énergie, biodiversité, etc.) qui peuvent avoir un impact direct ou indirect pour chaque habitant de notre planète.

Sortir, sortie, nucléaire : peut se traduire par l'expression « Sortir du nucléaire », une association indépendante[273] mais aussi par l'interprétation suivante *:* la « sortie du nucléaire de l'Allemagne »

[268] Le Monde : « le convoi nucléaire arrive à destination », 08/11/2010, http://www.lemonde.fr/planete/article/2010/11/08/face-a-face-tendu-entre-policiers-et-manifestants-devant-le-convoi-du-nucleaire_1436844_3244.html (consulté le 02/07/2014).
[269] Le Monde : « des militants écologistes immobilisent un convoi de déchets nucléaires », 06/03/2013, http://www.lemonde.fr/planete/article/2013/02/06/des-militants-ecologistes-immobilisent-un-convoi-de-dechets-nucleaires_1827636_3244.html (consulté le 02/07/2014)
[270] Le Figaro : « des déchets nucléaires arrivés dans la Hague », flash actu du 07/03 2012, http://www.lefigaro.fr/flash-actu/2012/03/07/97001-20120307FILWWW00457-des-dechets-nucleaires-arrives-a-la-hague.php (consulté le 02/07/2014)
[271] L'écologie, les Verts : « historique des Verts », http://www.lesverts.fr/spip.php?article232 consulté le 02/07/2014
[272] Greenpeace : http://www.greenpeace.org/france/fr/connaitre-greenpeace/ (consulté le 18/10/2014)
[273] Sortir du nucléaire : http://www.sortirdunucleaire.org/Nous-connaitre (consulté le 19/10/2014)

faisant suite à l'accident de Fukushima. Mycle Schneider analyse l'ensemble des « *scénarios de sortie du nucléaire en Allemagne et montre qu'un abandon rapide (1 à 2 ans) entraînerait une augmentation à court terme des émissions de gaz à effet de serre, mais que ces émissions repasseraient au bout de quatre ans en dessous du niveau de référence.*» (Schneider, 2000 : 15).
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *Allemagne, Angela Merkel, 2022, Nicolas Sarkozy, Hollande, Besson, atome, filière, industriels, électricité, nucléaire, accident, gendarme, convoi, stockage, réseau.*

Lhomme et ***Stéphane*** : Monsieur Stéphane Lhomme, porte-parole du réseau Sortir du nucléaire a été arrêté le 17 mai 2006. « *Les Verts déclarent être en possession du document classé secret défense ayant valu à Stéphane Lhomme d'être incarcéré. Ce document indique quels risques sont encourus en cas d'attaque aérienne contre le réacteur Nucléaire EPR* »[274].
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, accident, gendarme, sortir, EPR*.

Les autres formes poly-cooccurrentes

Hollande *:* S'agit-il de François Hollande ou des Pays-Bas ? L'un et l'autre peuvent être rapprochés de la forme-pôle nucléaire pour des raisons bien différentes.
Pendant la campagne présidentielle de 2012, *François Hollande* déclarait dans le 41ème engagement[275] : « *Je préserverai l'indépendance de la France tout en diversifiant nos sources d'énergie. J'engagerai la réduction de la part du nucléaire dans la production d'électricité de 75 % à 50 % à l'horizon 2025, en garantissant la sûreté maximale des installations et en poursuivant la modernisation de notre industrie nucléaire. Je favoriserai la montée en puissance des énergies renouvelables en soutenant la création et le développement de filières industrielles dans ce secteur. La France respectera ses engagements internationaux pour la réduction des émissions de gaz à effet de serre. Dans ce contexte, je fermerai la centrale de Fessenheim en 2016 la plus vieille centrale nucléaire française encore en activité et je poursuivrai l'achèvement du chantier de Flamanville (EPR) qui devrait être opérationnelle en 2016.*»
Autre homonyme ***Hollande*** : la *Hollande* a une centrale nucléaire dont les déchets sont retraités par la France, or ceux-ci sont acheminés par convoi ferroviaire, convois souvent « attaqués » par les mouvements écologistes (Bonnaure, 2011).
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, déchets, stockage, socialiste, programme, PS.*

Gendarme : peut désigner les interventions de la gendarmerie lors d'actions écologistes ou autres, mais aussi l'Agence internationale de l'énergie atomique (AIEA), parfois surnommée le « gendarme de l'atome », organisation créée en 1957 par l'Assemblée générale des Nations Unies pour encourager et faciliter, dans le monde entier, le développement et l'utilisation pratique de l'énergie atomique à des fins pacifiques, et la recherche dans ce domaine.[276]

[274] L'écologie, les Verts : « arrestation de Stéphane Lhomme porte-parole du réseau Sortir du nucléaire, http://www.lesverts.fr/spip.php?article2713 (consulté le 02/07/2014)
[275] Extrait du projet socialiste 2012 pour la campagne présidentielle 2012, http://www.parti-socialiste.fr/articles/les-60-engagements-pour-la-france-le-projet-de-francois-hollande (consulté le 02/07/2014)
[276] Source AIEA : http://www.un.org/fr/disarmament/instruments/iaea.shtml (consulté le 19/10/2014)

Annexe H

Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, incidents.*

Les formes poly-cooccurrentes se rapportant aux activités du domaine nucléaire

Retraitement, radioactif, stockage, déchets *:* liés au stockage des déchets radioactifs et à leur retraitement. En France, le site AREVA La Hague (situé près de Cherbourg) assure la première étape du recyclage des combustibles usés provenant des réacteurs nucléaires du monde entier, premier centre industriel de ce type dans le monde. *« AREVA La Hague accorde une importance majeure à la protection de l'environnement. L'établissement mène une politique continue de réduction de l'impact de son activité sur la santé et le milieu naturel. Il consacre un effort important à la surveillance de l'environnement : plus d'une centaine de prélèvements et d'analyses sont effectués chaque jour »*[277]. Cependant, des incidents sont signalés, comme par exemple le *5 novembre 2010, « où des effluents très faiblement radioactifs ont été transférés de l'atelier de réception et de déchargement de combustibles usés nucléaires vers la station de traitement des effluents sans que l'opération soit préalablement autorisée par l'équipe chargée de la conduite de cette station ».*[278]
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, convoi, gendarme, sécurité, Manche, combustibles, réacteurs, incident, site, usine, fusion, fission, uranium.*

Les opérateurs

EDF : « *EDF, premier électricien nucléaire mondial, gère en France un parc de production nucléaire composé de 58 unités de production réparties sur 19 sites. Le parc produit en moyenne 410 milliards de kWh par an et assure plus de 87% de la production d'électricité d'EDF qui propose ainsi à ses clients un kWh parmi les plus compétitifs d'Europe* »[279]. «*EDF apporte des solutions compétitives pour concilier durablement développement économique et protection du climat* »[280].

AREVA : multinationale française du secteur de l'énergie œuvrant principalement dans le domaine du nucléaire.

TEPCO (Tokyo Electric Power) : multinationale japonaise et avant sa nationalisation le plus grand producteur mondial privé d'électricité. « *Tepco, le gérant de la centrale de Fukushima, est nationalisée : Les actionnaires de Tokyo Electric Power (Tepco) ont approuvé lors d'une assemblée générale ordinaire, mercredi 27 juin, une augmentation de capital grâce à un apport de l'Etat, entérinant une nationalisation de facto de la compagnie gérante de la centrale nucléaire accidentée de Fukushima (…) . Tepco est actuellement dans une situation extrêmement délicate, avec des finances dévastées par les conséquences de la catastrophe nucléaire de Fukushima, provoquée par le séisme et le tsunami du 11 mars 2011 – la pire depuis celle de Tchernobyl, en Ukraine, en 1986* »[281]. *« Tepco a déjà dépensé 27 milliards de dollars depuis l'accident et doit faire face à de nouvelles obligations somptuaires pour déclasser la centrale et la mettre hors service, indemniser les 160.000 personnes évacuées et assumer les coûts de la décontamination de la région entourant le site*»[282].

[277] http://www.areva.com/FR/activites-1253/la-surveillance-de-l-environnement-de-l-usine-areva-la-hague.html (consulté le 18/10/2014)
[278] http://www.sortirdunucleaire.org/Incident-a-l-usine-Areva-de-La (consulté le 18/10/2014)
[279] Site EDF : http://energie.edf.com/nucleaire/accueil-45699.html (consulté le 18/10/2014)
[280] Site EDF : http://presentation.edf.com/profil/histoire/1990-a-nos-jours-40182.html (consulté le 18/10/2014)
[281] Le Monde, « Tepco, le gérant de la centrale de Fukushima, est nationalisée », 27/06/2012, http://www.lemonde.fr/japon/article/2012/06/27/tepco-qui-gere-la-centrale-de-fukushima-nationalisee-par-le-japon_1725037_1492975.html (consulté le 18/10/2014)
[282] http://fr.reuters.com/article/frEuroRpt/idFRL5N0IK02820131030 (consulté le 18/10/2014)

Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *nucléaire, accident, électricité, production, réacteurs, catastrophe, convoi, Fukushima, Daiichi, Japon, japonaise, arrêt*.

Les organismes de surveillance

ASN : Autorité de Sûreté Nucléaire **«** *assure, au nom de l'État, le contrôle de la sûreté nucléaire et de la radioprotection en France pour protéger les travailleurs, les patients, le public et l'environnement des risques liés à l'utilisation du nucléaire ... Elle contribue à l'information des citoyens ».*[283]

IRSN : Institut de radioprotection et de sûreté nucléaire est l'expert public en matière de recherche et d'expertise sur les risques nucléaires et radiologiques et a été créé en mai 2001. Le champ de compétences de l'IRSN couvre l'ensemble des risques liés aux rayonnements ionisants, utilisés dans l'industrie ou la médecine, ou encore les rayonnements naturels[284].

CERN : (Conseil européen pour la Recherche nucléaire), une organisation européenne pour la Recherche nucléaire. Cet organisme à l'origine provisoire institué en 1952 avait pour mandat de créer en Europe une organisation de rang mondial pour la recherche en physique fondamentale et se consacre à la recherche scientifique fondamentale est souvent appelé Laboratoire européen pour la physique des particules.[285]
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *autorité, nucléaire, accident, incident, sécurité.*

Les formes en rapport avec les types d'énergies
Renouvelable : leur part en Europe progresse lentement. *« Selon les estimations publiées par Eurostat, l'office statistique de l'Union Européenne, en 2011, l'énergie provenant de sources renouvelables a représenté seulement 13 % de la consommation finale brute d'énergie dans l'UE, contre 7,9% en 2004 et 12,1% en 2010 (...) Au niveau européen, en 2011 (...) 13% était issu des énergies renouvelables avec un peu plus de 8% pour la biomasse et 2,5% pour l'hydroélectricité. C'est pourquoi, l'UE s'est donnée comme objectifs d'atteindre 20 % d'énergies renouvelables dans la consommation totale d'énergie en 2020, mais aussi de diminuer de 20 % la consommation d'énergie et de réduire 20% ses émissions de gaz à effet de serre (...) ».* Comme Pierre Bauquis l'explique : *« ... parmi les questions que pose l'avenir des énergies renouvelables, une des plus importantes est celle des types d'aides qu'il convient de mettre en œuvre pour accélérer leur développement. Les énergies renouvelables aujourd'hui requièrent des efforts de recherches, mais ces efforts pour être efficaces doivent relever d'une logique très décentralisée et irriguer une multitude de petites équipes (...)»* (Bauquis, 2001, 2006).
Mais que pensent les Français de ces nouvelles énergies ? Dans le cadre d'un sondage réalisé par l'IFOP/Le Monde entre le 21 et le 27 juin 2011, la question suivante a été posée : *« Pensez-vous qu'il serait possible, en France/Royaume-Uni/Allemagne/Italie/Espagne, de produire quasiment toute l'électricité nécessaire au pays à l'aide d'énergies renouvelables comme l'énergie solaire, éolienne, hydraulique ?».*

[283] Site ASN : http://www.asn.fr/Presse/L-ASN-en-bref (consulté le 18/10/2014)
[284] Site IRSN : http://www.irsn.fr/FR/IRSN/presentation/Pages/Presentation.aspx#.VEO3O0tquw0 (consulté le 18/10/2014)
[285] http://home.web.cern.ch/fr/about (consulté le 19/10/2014)

Annexe H

Tableau H.2
Avis de l'opinion publique face à l'utilisation d'énergies renouvelables[286]
(Italie, Allemagne, Espagne, Royaume-Uni, France)

Pays	Oui certainement	Oui probablement	Non probablement pas	Non certainement pas
Italie	**35 (1)**	**37 (4)**	19	9
Allemagne	29	**41 (1)**	23	7
Espagne	29	**41 (1)**	23	7
Royaume-Uni	18	39	**33 (2)**	10
France	**15 (5)**	**37 (4)**	**37 (1)**	11

Les Français et les Britanniques semblent les plus sceptiques quant à l'utilisation d'énergies renouvelables.

Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *éolien, parc d'éoliennes, énergies, charbon.*

Nous constatons l'absence de certaines formes poly-cooccurrentes susceptibles d'être associées à la forme-pôle *nucléaire* comme le mot santé ou tout autre mot à consonance médicale (bactériologie, leucémie, cancer, épistémologie, etc.). Pourquoi cette absence : un mot tabou ? Les politiques n'en parlent pas ou très peu. Comme le mentionne Guillaume Roquette cité ci-dessus dans son édito du figaro magazine du 14 mars 2014, « *combien de personnes sont mortes dans la catastrophe de Fukushima ? Trois à la suite d'une des explosions (ils n'ont donc pas été irradiés). Bien sûr ce sont trois morts de trop mais disons-le c'est un chiffre dérisoire par rapport aux 17 000 victimes du tremblement de terre et du tsunami qui l'a suivi (...) ».* Et Tchernobyl*, « 28 ans après le drame l'organisation Mondiale de la santé dénombre une cinquantaine de morts par contamination directe et indique que le bilan définitif sera d'environ 4 000 décès (...) ».* Pourtant le nucléaire semble aujourd'hui une énergie moins dangereuse qu'une énergie traditionnelle comme le charbon « *dont l'extraction tue plusieurs milliers de mineurs chaque année et dont les micro-particules empoisonnent littéralement l'atmosphère* ».[287] Les conséquences de telles catastrophes sur l'environnement et sur les populations tout comme l'impact de la présence de centrales nucléaires en fonctionnement et du stockage des déchets sur l'environnement humain, animal, végétal, etc. sont mal maîtrisées. Diverses études bactériologiques ont été réalisées, comme celle parue en 2010 sur les taux de leucémie chez l'enfant[288]. Une augmentation du nombre de leucémies chez les enfants et les jeunes adultes habitant à proximité d'installations nucléaires (par exemple Sellafield en Angleterre ou Dounreay en Ecosse, Krümmel en Allemagne ou à proximité de l'usine de retraitement de La Hague en Cotentin) est constatée. Cette étude ne permet pas d'établir une relation de cause à effet entre les doses attribuées aux rejets radioactifs et les cas de leucémie. Une autre étude (Laurier, Rommens et al, 2000) avait été lancée en 1997 *« afin d'évaluer de façon réaliste l'exposition aux rayonnements ionisants de la population des 0-24 ans résidant à proximité de l'usine de retraitement de La Hague, et d'en déduire le risque de leucémie associé entre 1978 et 1996 ».* Les résultats portent sur *« 6 656 individus nés entre 1954 et 1996 et ayant résidé pendant au moins un an dans le canton entre 1978 et 1996 avant l'âge de 25 ans, soit un total de 69 308 personnes-années. Sur la base de l'estimation des doses reçues, le nombre de cas de leucémie radio-induite au sein de cette population, attribuable à l'exposition due aux rejets des installations nucléaires locales, est inférieur à 0,002 sur la période 1978-1996 ».* En conclusion*, « ... Le nombre de leucémies radio-induites estimé est faible en regard des 4 cas de*

[286] Source : enquête IFOP/ le Monde réalisée du 21 au 27 juin 2011
[287] Guillaume Roquette édito du Figaro Magazine du 14 mars 2014
[288] Article : « Info + les leucémies chez l'enfant : mécanismes et causes », lettre d'information de l'unité Prositon, CEA n°6, septembre 2010, p.5.

leucémie observés sur la même période. Il est donc très peu probable que l'exposition due aux installations nucléaires locales puisse être impliquée de façon notable dans l'incidence élevée de leucémie observée chez les jeunes dans le canton de Beaumont-Hague ». Diverses études bactériologiques sont réalisées périodiquement (Laurier, 2007). Autres mesures plus récentes effectuées par l'Association pour le contrôle de la radioactivité dans l'Ouest[289] (ACRO) affirment qu'il y a une concentration en tritium de 110 becquerels par litre d'eau dans la baie d'Ecalgrain située à quelques encablures de l'usine Areva de la Hague dans la Manche. Autre cas, les essais nucléaires militaires dans les années 1960, comme par exemple à Mururoa, n'ont pas été sans conséquence sur l'état de santé de la population locale. Mais les rapports établis par l'Armée française n'ont jamais été divulgués.

H.2 Forme-pôle *énergie* : constats pour ENRG_FR

Les figures H.1 et H.3 se rapportant aux deux formes-pôles *nucléaire* et *énergie* sont strictement identiques tant par leur nombre de formes poly-cooccurrentes que par leur contenu.

[289] GOUIN, Simon, article paru le 29/03/2013, « Agir contre la contamination radioactive en France dans la Manche, plus de radioactivité qu'à Fukushima ? », http://www.bastamag.net/Dans-la-Manche-plus-de (consulté le 26/03/2015).

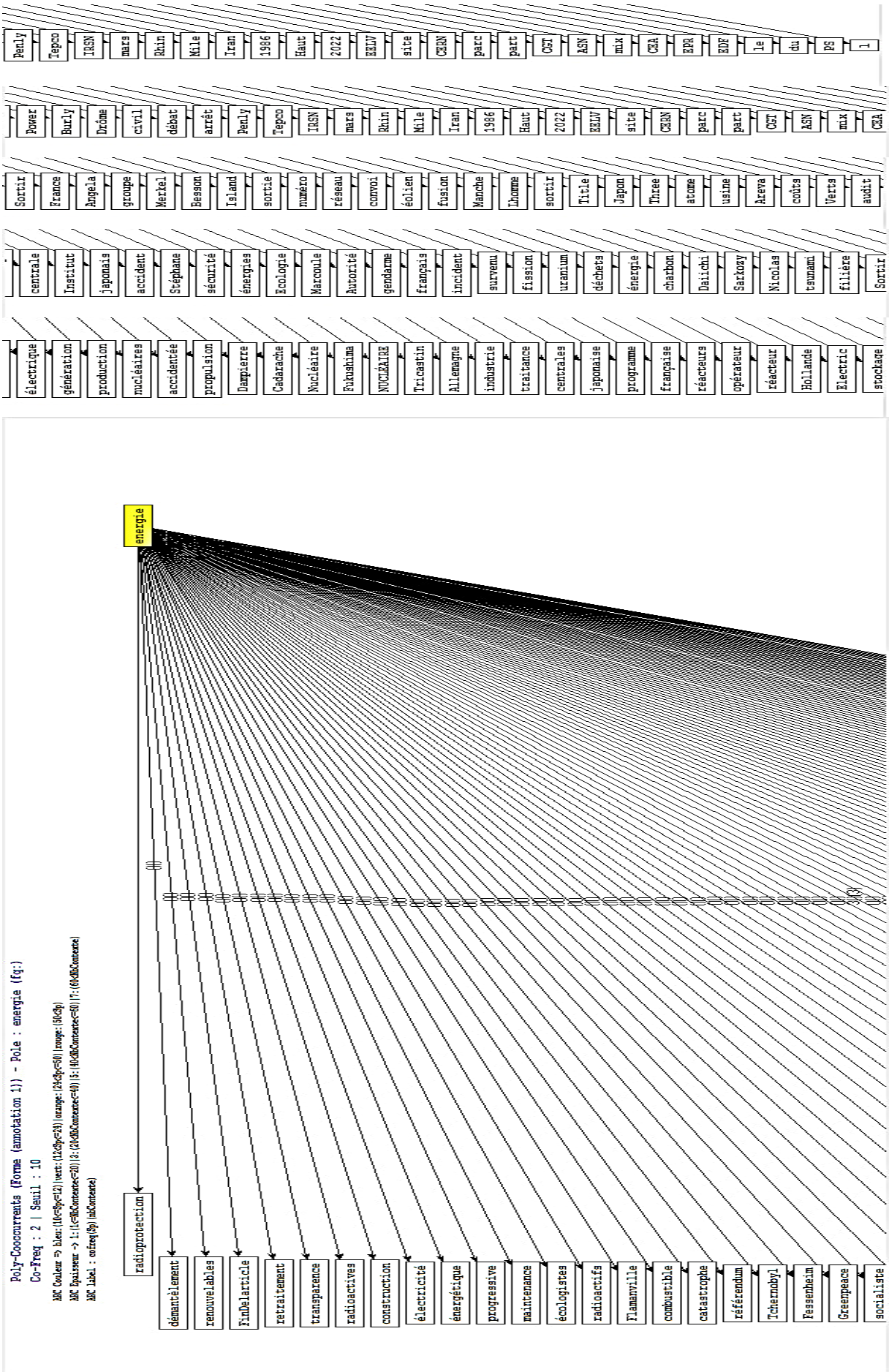


Figure H.3

ENRG_FR : poly-cooccurrences autour de la forme-pôle *énergie*

H.3 Etude de la ventilation de la spécificité des formes-pôles *énergie, énergies, nucléaire* par année pour ENRG_FR

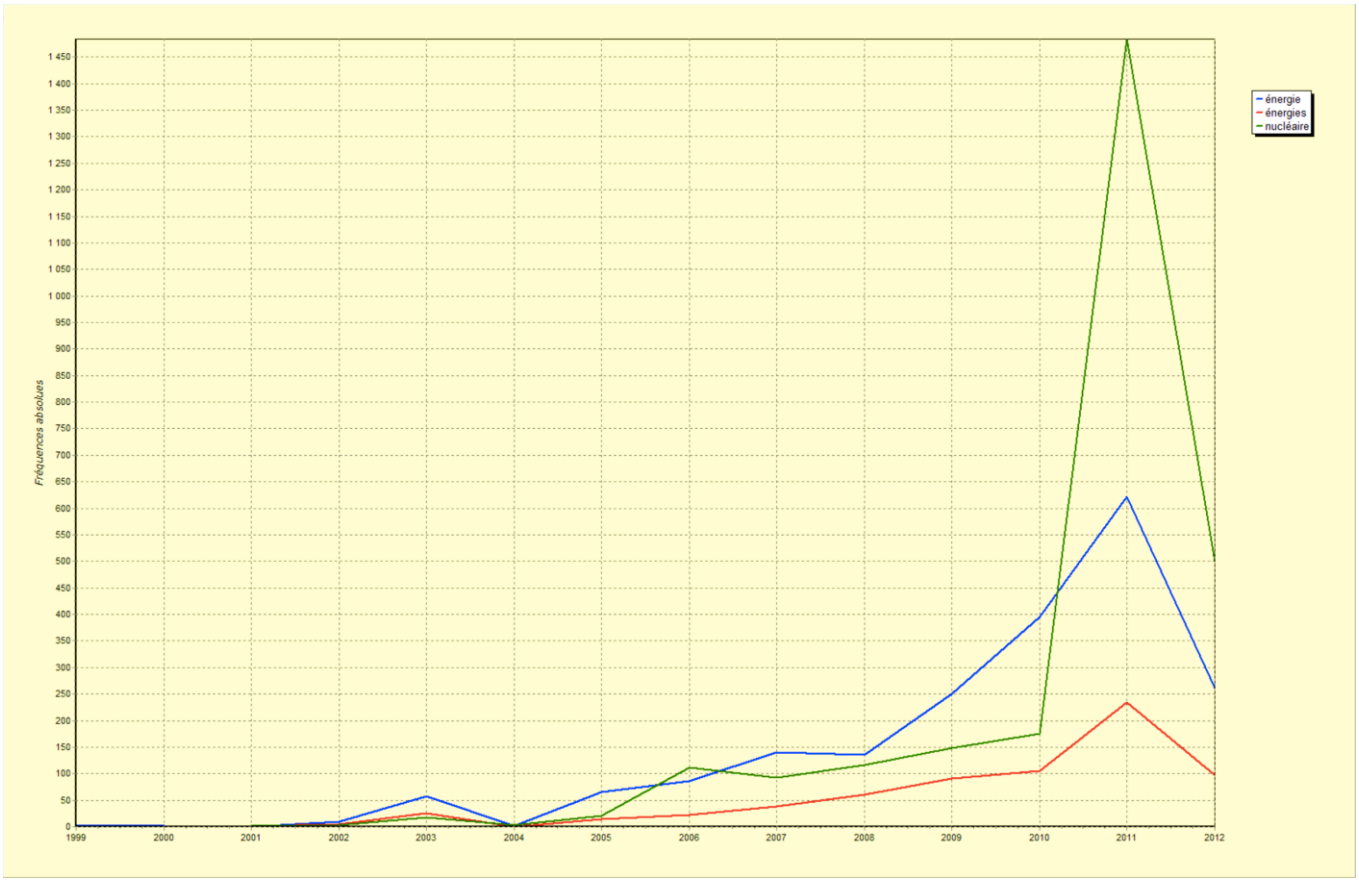


Figure H.4
Ventilation des fréquences absolues par année des formes-pôles *énergie, énergies, nucléaire* du 24/09/1999 au 17/04/2012

L'analyse de la figure H.4 nous montre trois courbes présentant les mêmes tendances, mais leurs fréquences absolues ne sont pas identiques. Les trois courbes démarrent à la mi-2002, avec un premier « pic » en 2003, puis une chute en 2004 et ensuite une tendance à la hausse progressive pour atteindre 3 « pics » en 2011. Ensuite, les trois courbes s'infléchissent. Les catastrophes naturelles à l'origine de ces accidents nucléaires sont inimaginables. La plus terrifiante de ces dernières années a été celle de Fukushima (11 mars 2011), le « pic 2011 » correspondant à un accident de niveau gravité 7 classé selon l'INES. Tous ces événements sont bien sûr fortement médiatisés. Quelques mois plus tard, la courbe s'infléchit, mais l'événement reste présent en mémoire.

Revenons au début des années 2000, rien ou peu de choses sur le nucléaire, ce n'est pas la préoccupation actuelle. L'année 2002 voit le passage à l'euro en Europe préoccupe davantage les esprits que le nucléaire, mais un soubresaut en 2003, peut-être lié au problème de la centrale de Paks en Hongrie (fuite radioactive classé niveau 3 par l'INES a vraisemblablement mis en danger la population environnante).

L'année 2005 voit un autre accident nucléaire à la centrale de Sellafield en Grande-Bretagne, accident de gravité 3, mais cette centrale a déjà connu un très grave incident nucléaire qui a duré plusieurs jours pendant lesquels des produits de fission sont rejetés à l'extérieur. Le nuage radioactif a parcouru l'Angleterre porté par les vents puis touche le continent sans que la population ne soit avertie. Après cet accident, l'usine qui s'appelait Windscale est débaptisée pour devenir Sellafield.

La période 2004-2007 correspond au débat sur l'EPR et à l'implantation de ce premier réacteur à

Flamanville dans la Manche. Cette commune du Cotentin à quelques encablures de l'usine de retraitement des déchets radioactifs de la Hague, figure parmi les lieux d'implantation du programme de développement nucléaire des années 1970. En 1975, une consultation publique est organisée à Flamanville ; 63,7% des Flamanvillais se prononcent en faveur de l'implantation de la centrale nucléaire[290].

En conclusion, les analyses des figures H.1 et H.3, nous apprennent que les formes-pôles *nucléaire et énergie,* sont rattachées à de nombreuses autres formes poly-cooccurrentes liées aux domaines du nucléaire et de l'énergie en général quel que soit sa source et son type (charbon, gaz, nucléaire, hydraulique, éolienne, etc.), mais aussi à tous les organismes étatiques et privés, à la politique nucléaire et énergétique menée par les différents gouvernements français et à tous les événements se rapportant à ces deux formes-pôles *nucléaire* et *énergie*.
Quant à la figure H.4, elle permet un suivi de la chronologie des événements liés aux formes et à la période retenue.
Ces conclusions relèvent d'une étude empirique et peuvent être vérifiées par un retour au contexte.

[290] https://www.edf.fr/sites/default/files/contrib/groupe-edf/producteur-industriel/carte-des-implantations/centrale-flamanville/presentation/dossier_de_presse_2015_centrale_de_flamanville.pdf (consulté le 20/08/2015).

Annexe I : dictionnaire d'événements et restitutions par poly-cooccurrences des formes-pôles *nuclear* et *energy* pour le sous-corpus ENRG_US pour la période du 26 janvier 2005 au 18 avril 2012

I.1 Forme-pôle *nuclear* : constats pour ENRG_US

L'illustration des poly-cooccurrences pour ENRG_US se trouve dans la figure I.1 ci-après.

Afin d'attraper un maximum d'informations, à l'aide de l'outil Trameur, les paramètres par défaut pour tous les calculs de poly-cooccurrences ont été maintenus : co-fréquence 2, seuil 10, contexte . !?

Le contexte . !? signifie approximativement dans le langage courant phrase.

Le retour au contexte s'avère indispensable pour attester toute affirmation effectuée à partir d'une veille aveugle et présentée dans nos recherches ci-après.

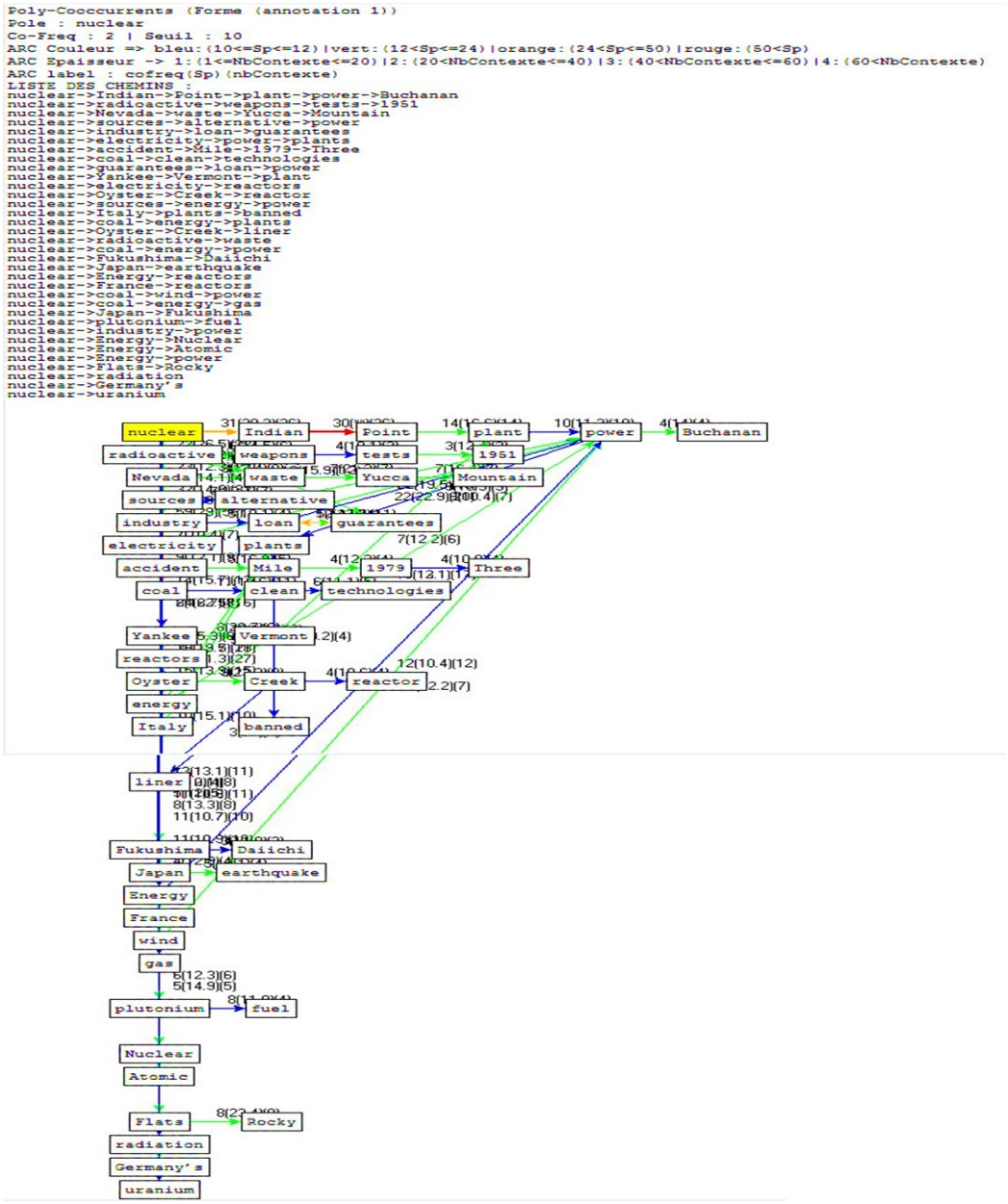


Figure I.1
ENRG_US : poly-cooccurrences autour de la forme-pôle *nuclear*

Suite à notre veille aveugle et au vu des figures relevant les formes-pôles des sous-corpus ENRG_FR et ENRG_US pour les formes *nucléaire / nuclear* et *énergie / energy*, l'approche anglophone aux Etats-Unis ne semble pas la même que l'approche française. Un bref rappel historique du nucléaire accompagné de quelques dates clés, d'une présentation des acteurs et organismes étatiques ou privés est disponible dans l'annexe F, la politique énergétique aux Etats-Unis et dans le dossier Doc_Annexes, Doc_Annexe-F-énergie-Etatsunis.docx disponible sur le serveur dédié[291]et permet de cerner le contexte énergétique américain.
Nous allons maintenant examiner les formes-pôles *nuclear* et *energy* à l'aide des figures I.1 et I.2 représentant les poly-cooccurrences.

[291] Lien du serveur dédié : https://drive.google.com/folderview?id=0B8XHfHwNzWAAeDN6UEc1b2dGa1U&usp=sharing

Une analyse de la figure I.1, ci-dessus, nous montre une présence de formes poly-cooccurrentes se rapportant à quelques toponymes (mais en nombre moins important que pour ENRG_FR) et de formes évoquant des sources d'énergies fossiles, renouvelables et vertes. Par contre, contrairement à ENRG_FR, nous constatons une quasi-absence de formes associées à la vie politique, que ce soit des noms d'hommes ou de partis politiques, d'associations, mais aussi de formes traitant des activités du nucléaire et de tout sujet autour de ce thème.
Deux autres constats, le premier est le nombre de formes poly-cooccurrentes dans ENRG_US beaucoup moins élevé que dans ENRG_FR, dans chacune des figures relatives aux poly-cooccurrences par pôle et le deuxième porte sur le contenu des mots obtenus entre les figures I.1 et I.2 des formes-pôles *nuclear* et *energy* d'ENRG_US.

Les toponymes

Les toponymes reflètent principalement quelques sites de centrales ou de stockage atypiques dans le panorama du nucléaire.
Nuclear >Nevada > waste > Yucca > Mountain : centrale Yucca Mountain, située à 130 kilomètres au Nord-Ouest de Las Vegas dans le Nevada. Cette montagne, d'origine volcanique, avait été choisie par le gouvernement fédéral comme site d'enfouissement : les déchets nucléaires produits par les réacteurs américains devaient y être enterrés 300 mètres sous terre. Les opposants à ce projet multiplièrent les procédures judiciaires. L'administration Obama semble confirmer le futur abandon du projet de stockage des déchets radioactifs de Yucca Mountain[292]. En mai 2007, une nouvelle carte géologique montre le site plus vulnérable.
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *radiation, reactors, energy*.

Nuclear > accident > Mile > 1979 > Three : quatre mots accident, Mile, 1979, Three situent l'événement concernant l'accident nucléaire survenu le 28 mars 1979 à la centrale de Three Mile en Pennsylvanie. *« (...) Une série de défaillances matérielles et humaines provoquaient la fusion partielle du cœur du réacteur nucléaire. Le choc provoqué par cet accident a été considérable et les enseignements tirés ont été nombreux, notamment en France (...) »*[293].
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *radiation, reactors, energy*.

Nuclear > Indian > Point > plant > power > Buchanan : centrale nucléaire d'Indian Point, centrale équipée de trois réacteurs nucléaires (forme *power*) et implantée (forme *plant*) à Buchanan (Etat de New-York), une centrale également à risques : *« (...) Jusqu'au 11 septembre 2001, ce n'était qu'une centrale nucléaire un peu vétuste, tout près de New York. Aujourd'hui, elle cristallise la peur et les hantises des Américains. Elle cumule le plus grand nombre d'incidents de toute l'industrie nucléaire du pays. Elle se trouve à peine à 40 kilomètres au nord de la ville de New York, dont elle fournit un tiers de l'électricité. Plus de 20 millions d'habitants vivent dans un rayon de 80 kilomètres (...) »* (Leser et Thompson, 2003).
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *radiation, reactors, energy*.

[292] Source : Ambassade de France aux Etats-Unis, BE Etats-Unis 159 (27/03/2009), http://www.bulletins-electroniques.com/actualites/58419.htm (consulté le 20/10/2014)
[293] Site IRSN, http://www.irsn.fr/FR/connaissances/Installations_nucleaires/Les-accidents-nucleaires/three-mile-island-1979/Pages/sommaire.aspx#.VETnGktquw1 (consulté le 20/10/2014)

Nuclear > Oyster > Creek > reactor : centrale américaine la plus ancienne située dans l'état du New Jersey. Plusieurs incidents ont été signalés sur ces deux dernières années[294].
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *liner, radiation, reactors, energy.*

Nuclear > Yankee > Vermont : centrale américaine implantée dans le Vermont, qui a fait l'objet d'une mise en vente en novembre 2010, un événement atypique, *« Vermont nuclear plant up for sale*[295]» (Wald, 2010).
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *reactors.*

Nuclear > Flats > Rocky : laboratoire installé près de Denver dans le Colorado et dont la principale activité était la production d'armes nucléaires et la fabrication des cœurs d'ogives nucléaires au plutonium. Il a été opérationnel du début des années 60 à 1989.
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *plutonium, radiation, uranium.*

Radioactive > weapons > tests > 1951: quatre mots constituant une chaîne dans la filière du nucléaire américaine. Cette date de 1951 marque le début d'une série de cinq essais atomiques atmosphériques sur le site d'essais du Nevada[296].
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *radiation, plutonium, uranium, energy.*

***N**uclear > Fukushima - >Daiichi* : deux noms liés puisque c'est le nom du lieu où une centrale a été touchée par la catastrophe auquel est associé le nom de l'exploitant.
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *Japon.*

Nuclear > Japon > earthquake : l'activité sismique est particulièrement importante au Japon et il est nécessaire de surveiller cette activité en permanence afin de prévenir d'éventuels accidents[297].
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *Fukushima, Daiichi, radiation, reactors.*

Les types et sources d'énergies

Nuclear > coal > wind > power : une énergie en pleine expansion aux Etats-Unis[298].
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *clean, technologies, sources, energy, power.*

Nuclear > coal > energy > gas : la veille aveugle ne permet pas de savoir de quel type de gaz, dont il est question. Gaz de schiste ? Les Etats-Unis ont connu une période d'euphorie avec le gaz de schiste[299], puis un déclin semble s'amorcer, « (…) *Les champs de Barnett et de Haynesville, dans le*

[294] Source : « Sortir du nucléaire », http://www.sortirdunucleaire.org/spip.php?page=recherche-risques&recherche=oyster+creek (consulté le 20/10/2014)
[295] « la centrale nucléaire de Vermont à vendre »
[296] Source : http://nuclearweaponarchive.org/Usa/Tests/ (consulté le 18/10/2014)
[297] Source : http://www.jma.go.jp/en/quake/quake_singendo_index.html (consulté le 20/10/2014)
[298] Extrait article : « L'énergie éolienne en pleine expansion aux Etats-Unis », 09/08/2013, http://iipdigital.usembassy.gov/st/french/article/2013/08/20130809280337.html#axzz3Gluwfl4j (consulté le 20/10/2014)
[299] gaz naturel souvent enfoui à très grande profondeur (1 500 à 3 000 mètres) dans des roches compactes et imperméables

Sud des Etats-Unis, ont franchi leur pic de production respectivement en novembre et décembre 2011. Ces puits ont fourni jusqu'ici près de la moitié de la production américaine de gaz de schiste, etc. »[300].
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *source, alternative, energy, power, technologies, clean.*
Avec la forme *nuclear,* nous trouvons l'association de la forme *coal* sachant que les Etats-Unis sont le deuxième pays producteur et consommateur de charbon.

Nuclear > coal > clean > technologies,

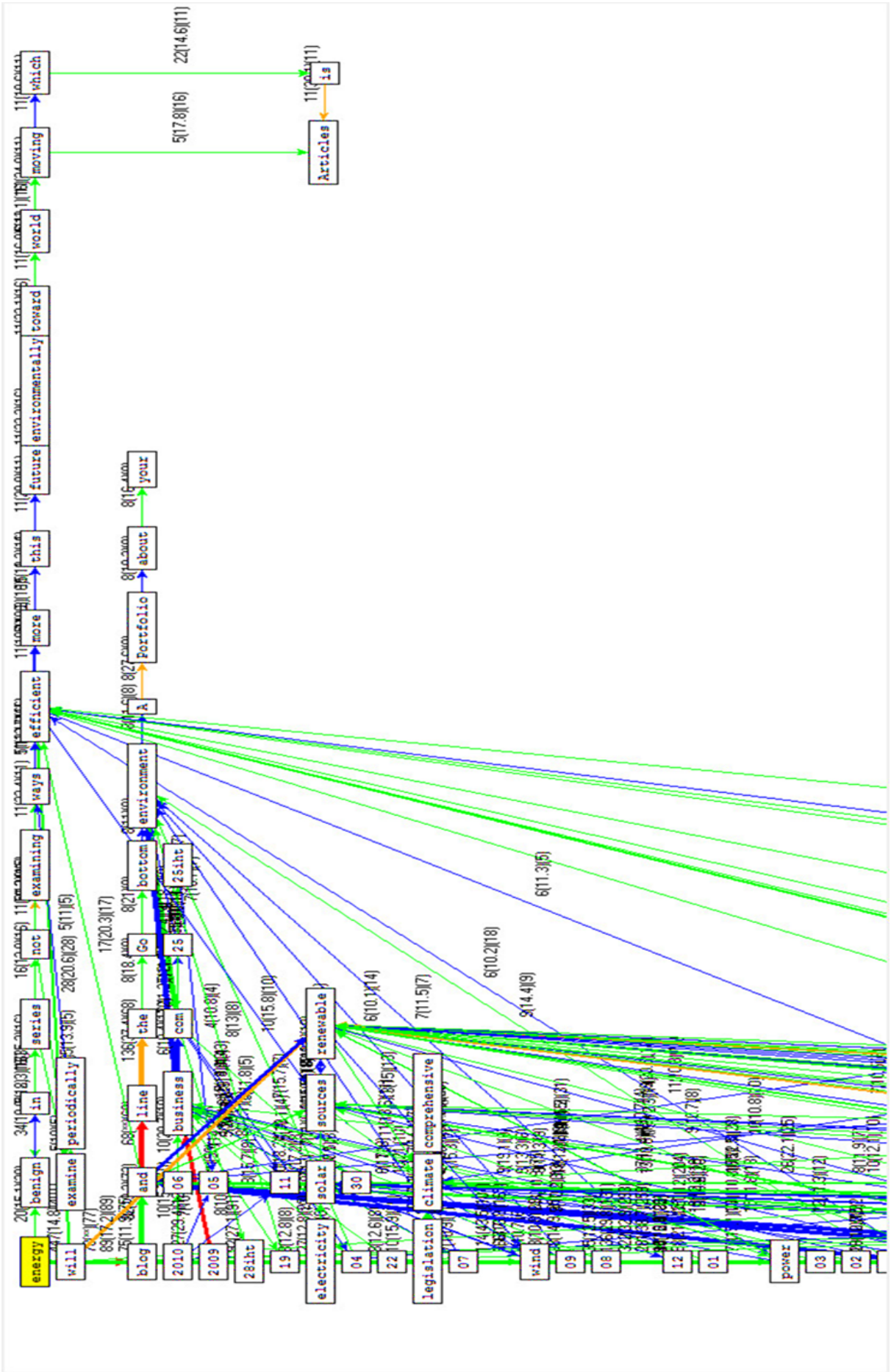
Des énergies émergeantes[301]
Nuclear >sources >alternative >power

Des associations de mots pour un pays qui dispose du plus grand parc nucléaire au monde
Nuclear > Energy > Nuclear
Nuclear > Energy > Atomic
Nuclear > Energy > Power
« (…) *Avec 104 réacteurs produisant environ 20 % de l'électricité du pays, le Département de l'Énergie Américain (DOE) prévoit que les besoins en électricité des États-Unis vont augmenter de 25 % d'ici 2030. Autrement dit, 35 nouvelles centrales nucléaires seraient nécessaires, afin de maintenir les 20 % de production d'électricité assurés par l'énergie nucléaire. Le nucléaire est la seule source d'énergie à faible émission de gaz à effet de serre, pouvant être exploitée pour subvenir aux besoins en énergie grandissants des Etats-Unis.* Les *États-Unis ont d'importants besoins énergétiques au vu de leurs 304 millions d'habitants (…) »*[302].

Nuclear > industry > loan > guarantees : l'année 2010 voit la relance du programme nucléaire américain avec l'annonce par le président Obama de construction de nouvelles centrales : **«** *(…) Nous annonçons environ 8 milliards de dollars en garanties de prêts pour entamer la construction de la première centrale nucléaire dans notre pays depuis près de 30 ans* ». Monsieur Obama déclare que *« ce soit dans l'énergie nucléaire, solaire ou éolienne, si nous n'investissons pas dans ces technologies aujourd'hui, nous les importerons demain (…) ».*[303]

I.2 Forme-pôle *energy* : constats pour ENRG_US
Voir les images ci-dessous :

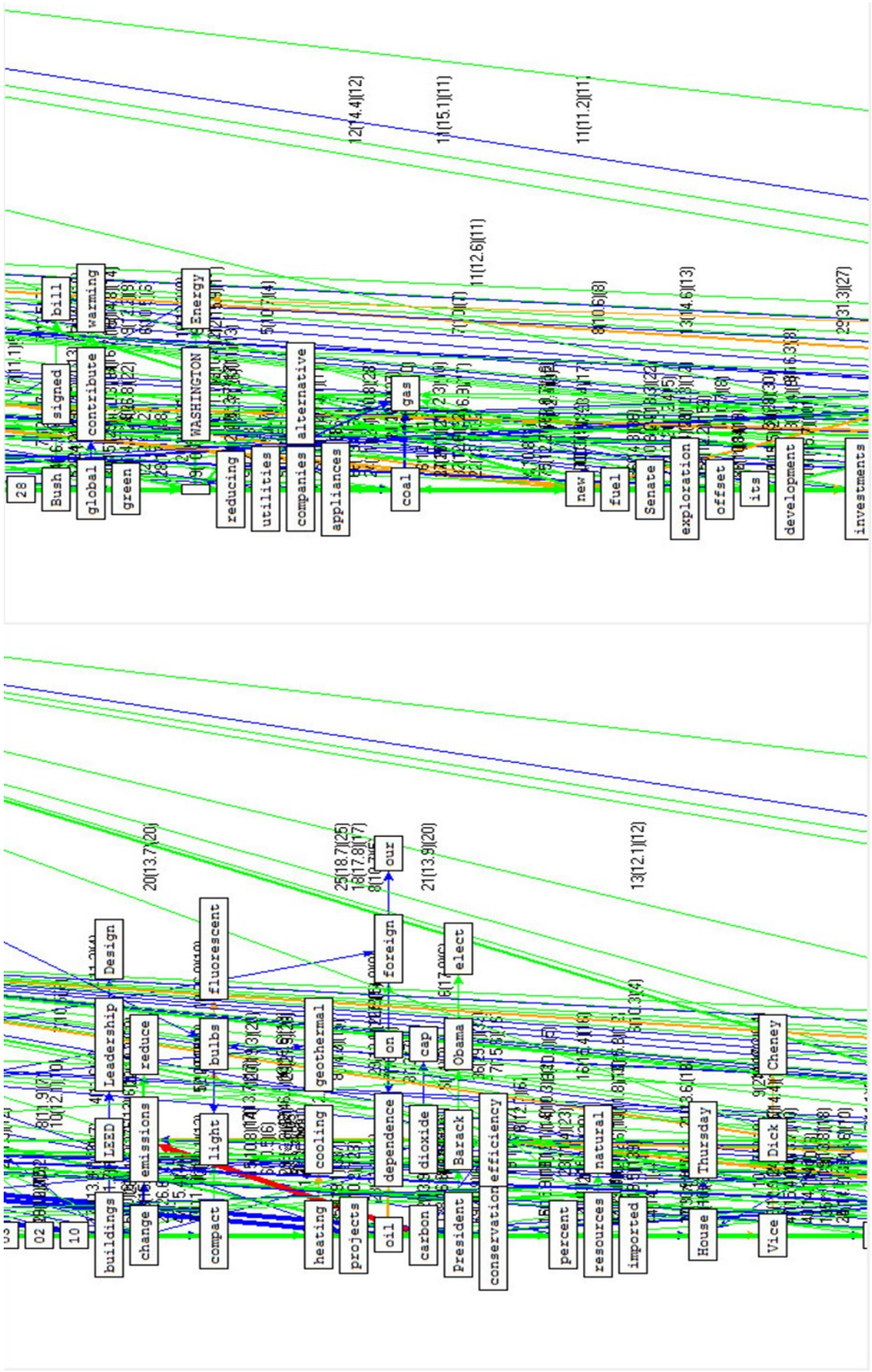
[300] Extrait blog Le Monde Matthieu Auzanneau : « gaz schiste : premiers déclins aux Etats-Unis », 01/10/2013, http://petrole.blog.lemonde.fr/2013/10/01/gaz-de-schiste-premiers-declins-aux-etats-unis/ (consulté le 19/10/2014)
[301] Extrait article : «le secteur de l'énergie renouvelable et la satisfaction de la demande d'énergie des Etats-Unis », 09/05/2007, http://iipdigital.usembassy.gov/st/french/article/2007/05/20070509163532saikceinawz0.5519373.html#axzz3Gluwfl4j (consulté le 19/10/2014)
[302] Source : site Areva, http://www.areva.com/FR/groupe-1505/les-etatsunis-une-forte-puissance-nucleaire-104-reacteurs-en-service.html (consulté le 18/10/2014)
[303] L'Express /L'Expansion, publié le 17/10/2010, http://lexpansion.lexpress.fr/actualite-economique/obama-relance-la-construction-des-centrales-nucleaires_1425482.html (consulté le 17/10/2010)



Partie 1

La lecture de l'image ci-dessous se fait de gauche à droite

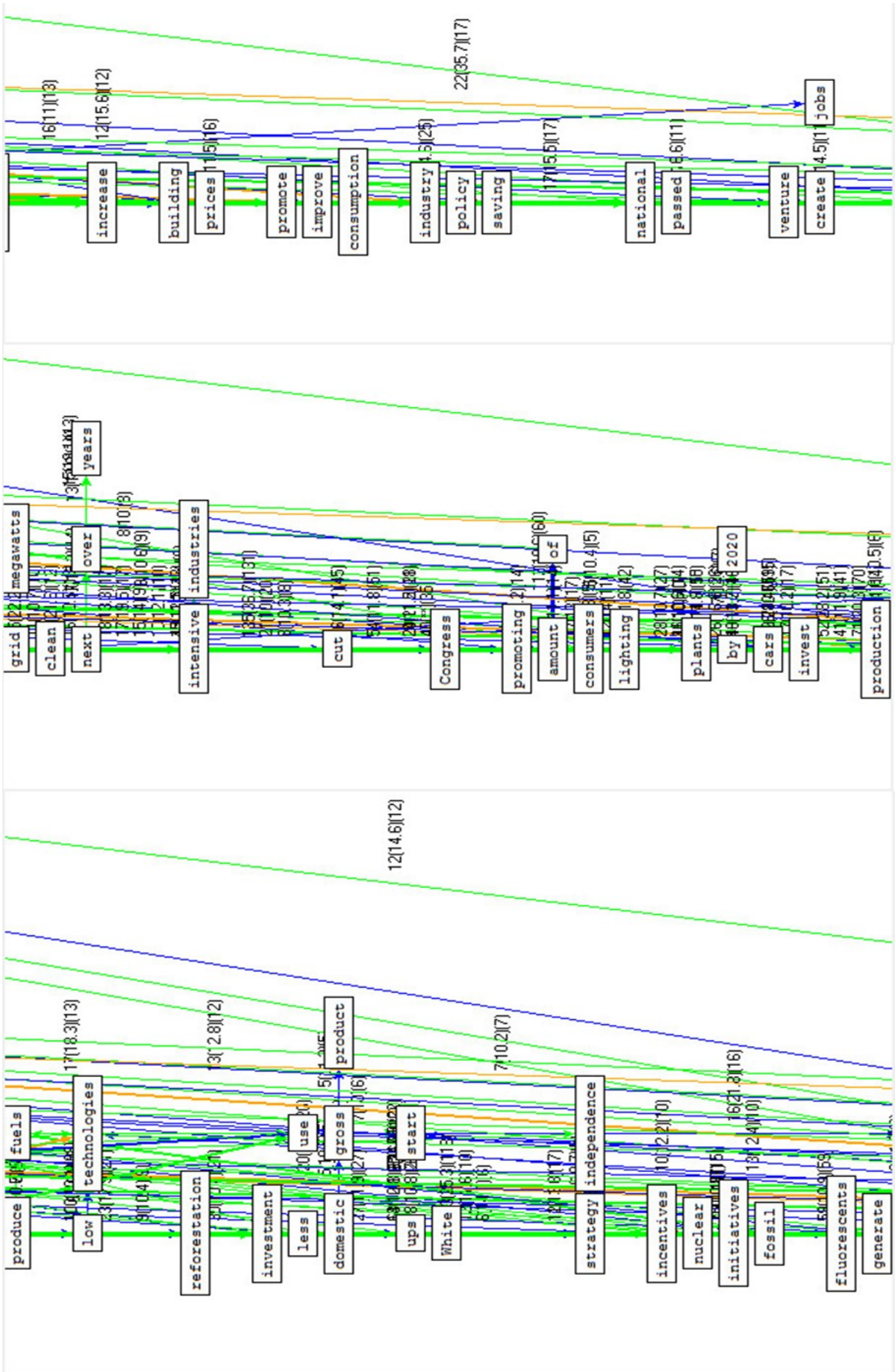
Annexe I



Partie 2

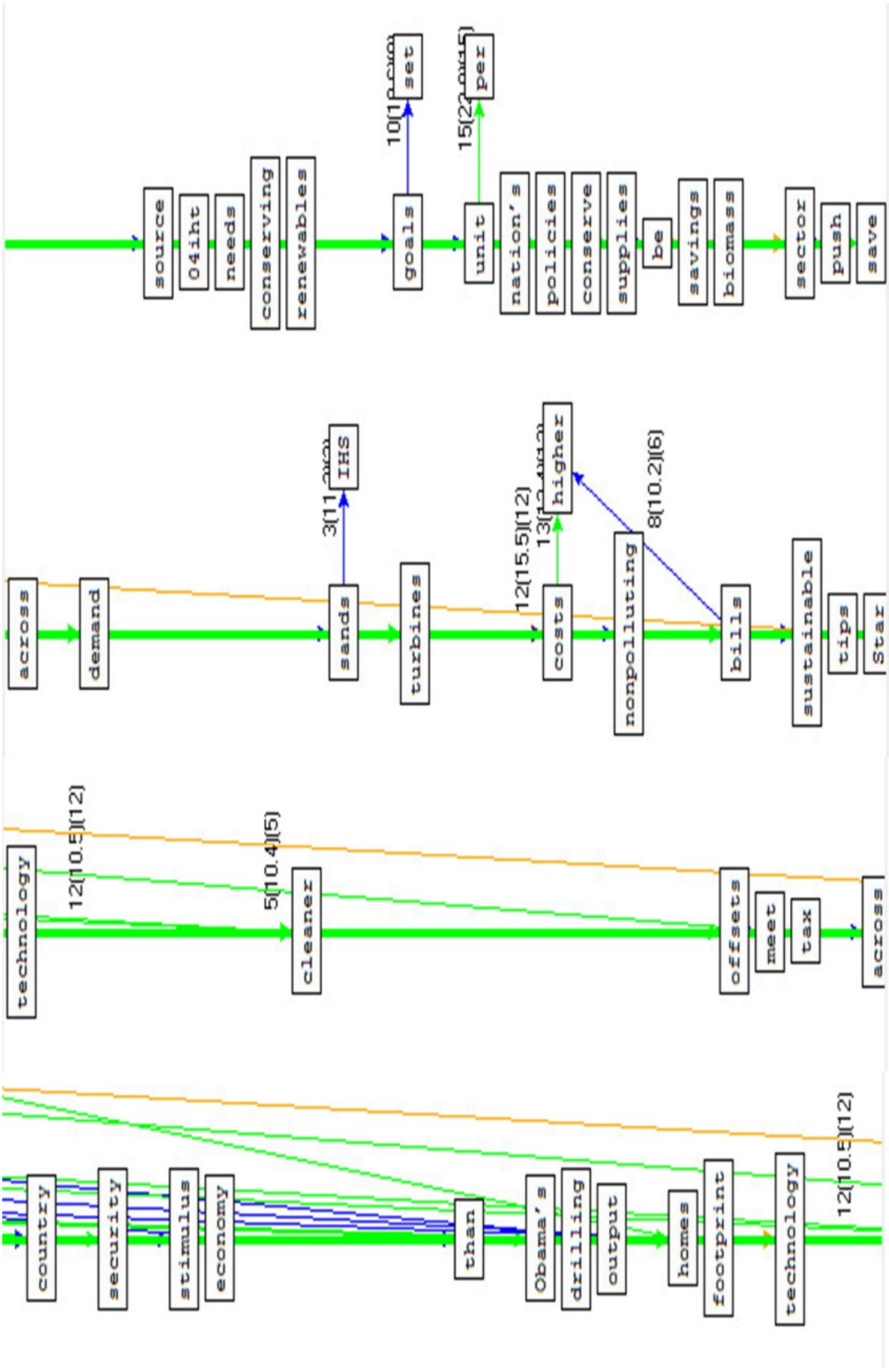
La lecture de l'image ci-dessous se fait de gauche à droite

Annexe I



Partie 3

La lecture de l'image ci-dessous se fait de gauche à droite



Partie 4 de la Figure I.2
ENRG_US : poly-cooccurrences autour de la forme-pôle *energy*

Annexe I

Une présentation de la situation énergétique américaine est consultable dans le chapitre 2, Energies et environnement dans le monde. L'analyse des mots associés à cette forme-pôle laisse la place à une présence politique bien marquée, tant sur le plan des hommes politiques que sur la politique énergétique menée aux Etats-Unis, alors que celle-ci était totalement absente de la figure I.1, à la différence de la France où les mots rattachés aux hommes politiques se trouvent à la fois dans la forme-pôle *nucléaire* et la forme-pôle *énergie*.
Energy > Président > Barack > Obama > elect ou Energy > Obama's : en 2010, Obama relance la construction des centrales nucléaires, une politique déjà énoncée pendant la campagne présidentielle, alors que les Etats-Unis boudaient l'énergie nucléaire depuis 30 ans. Le président mise sur cette technologie pour réduire les émissions de carbone ainsi que la dépendance énergétique de son pays. *« Il nous faut construire une nouvelle génération de centrales nucléaires sûres et propres aux Etats-Unis ».*[304]

Energy > Vice > Dick > Cheney : Dick Cheney alors vice-président des Etats-Unis[305] est chargé par le Président Bush de définir la politique américaine de l'énergie et évoque à nouveau la possibilité de construire des centrales nucléaires. Le vice-président a ensuite abordé le cas des énergies fossiles (gaz, pétrole, charbon), énergies produisant du dioxyde de carbone qui a pour effet d'augmenter l'effet de serre, avant de s'arrêter sur le nucléaire. « *Nous avons, après tout, mis au point une technologie qui ne cause aucune émission polluante et ne produit aucun gaz à effet de serre, c'est l'énergie nucléaire* (...). *Si nous voulons être sérieux en parlant de protection de l'environnement, nous devons nous interroger sur les raisons de nous détourner d'une source d'énergie qui a montré qu'elle est sûre, propre et abondante* », a déclaré Dick Cheney. Mais les Etats-Unis doivent résoudre le problème du stockage des déchets.

Energy > Bush > signed > bill : le protocole de Kyoto fixe des objectifs de réduction de gaz à effet de serre à une quarantaine de pays industrialisés. Mais les Etats-Unis n'ont pas ratifié ce protocole, protocole pourtant signé sous la présidence Clinton en 1997, mais ce protocole n'a pas été ratifié par le Sénat américain. Le 13 mars 2001, le Président George W. Bush a dénoncé le protocole, annonçant qu'il privilégiait de nouvelles approches pour combattre les gaz à effet de serre. Les Etats-Unis sont avec la Chine les principaux pays émetteurs de gaz à effet de serre dans le monde.
Formes poly-cooccurrentes associées dans les calculs de poly-cooccurrences : *relance, investments, invest, strategy, independance, Congress, Washington, Senate, White, House.*

Energy > change > emission > reduce : les Etats-Unis commencent à mettre en place quelques programmes de réduction des gaz à effet de serre (par exemple le Regional Greenhouse Gas Initiative, le Midwestern Greenhouse Gas Reduction Accord ou le Western Climate Initiative).

Energy > Green
Energy > Clean
Energy > Electricity > solar > sources > renewable

[304] L'Express/L'Expansion, publié le 17/02/2010, http://lexpansion.lexpress.fr/actualite-economique/obama-relance-la-construction-des-centrales-nucleaires_1425482.html (consulté le 20/10/2014)
[305] Les Echos, 02/05/2001, http://www.lesechos.fr/02/05/2001/LesEchos/18395-035-ECH_dick-cheney-envisage-de-relancer-le-nucleaire-aux-etats-unis.htm (consulté le 20/10/2014)

Ces formes poly-cooccurrentes concernent les énergies renouvelables[306] et les énergies alternatives[307]. Les Etats-Unis doivent absolument diminuer leurs importations d'énergie et limiter les émissions de gaz carbonique. Pour ce faire, ils développent les *green* (vertes) énergies et les *clean* (propres) énergies.

Les résultats de ces figures sur les formes-pôles nucléaire/nuclear, énergie/energy et de la veille aveugle démontrent un contexte énergétique français bien différent de celui des Américains. En France, la politique énergétique est depuis fort longtemps orientée vers le « tout nucléaire », d'une part nos ressources fossiles étaient limitées, d'autre part des choix d'indépendance énergétique ont être retenus par nos politiques (se reporter à l'annexe D, le nucléaire et la politique française).

I.3 Etude de la ventilation de la spécificité des formes-pôles *energy, energies, nuclear* par année pour ENRG_US

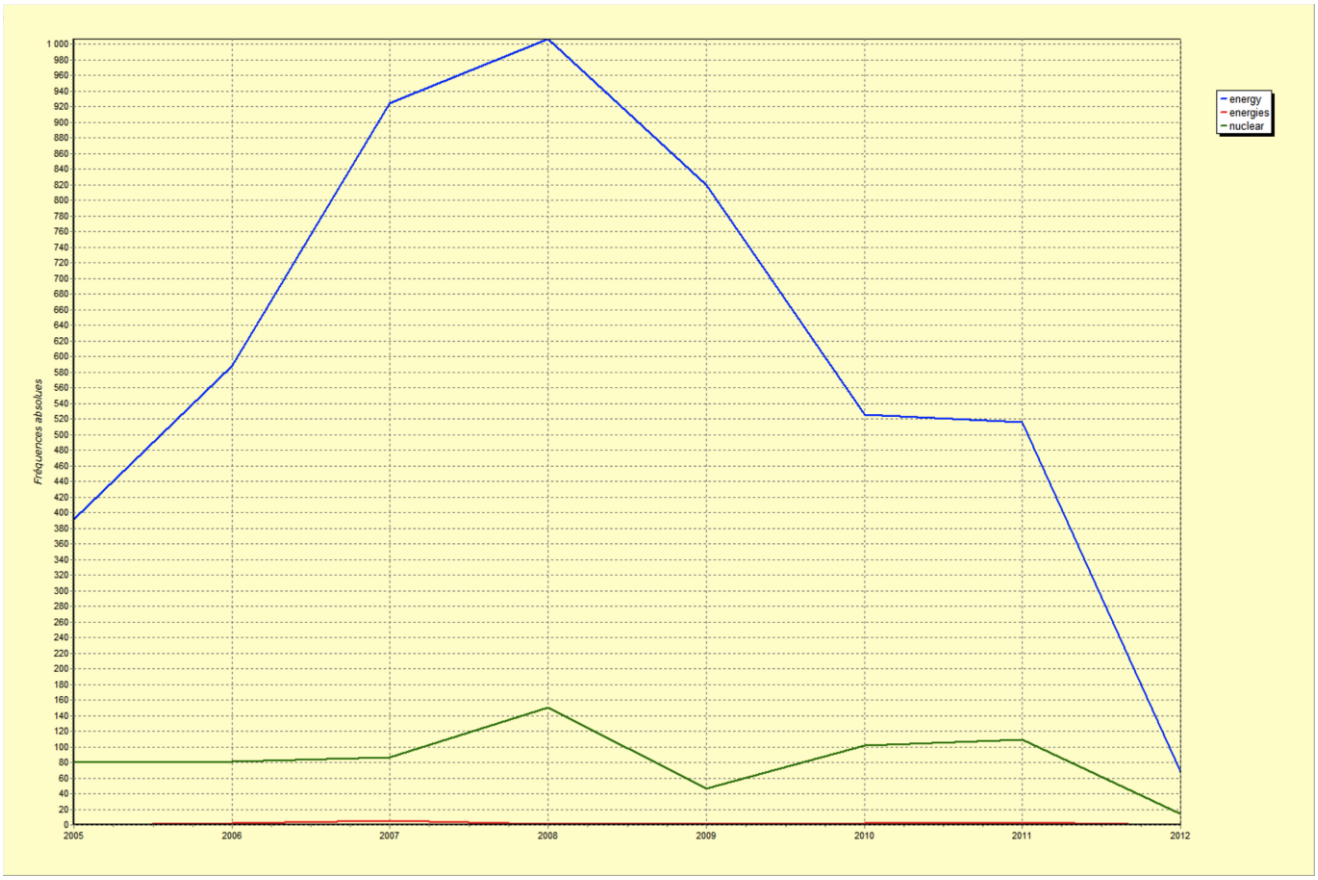


Figure I.3
ENRG_US : ventilation des fréquences absolues par année
des formes-pôles *energy, energies, nuclear*
du 26/01/2005 au 26/01/2012

La figure I.3 montre une fréquence absolue très importante de la forme-pôle *energy* par rapport à celle de la forme-pôle *nuclear*. Quant à la forme-pôle *energies, la courbe est quasiment plate,* ce constat peut s'expliquer par la non-utilisation du mot energy au pluriel dans la langue anglaise.

Par contre, pour le sous-corpus ENRG_FR, le constat est différent : c'est la forme *nucléaire* qui

[306] Energies renouvelables : énergies inépuisables qui se constituent ou se reconstituent plus rapidement qu'elles ne sont utilisées (énergie éolienne, énergie hydraulique, la biomasse produite par photosynthèse et une partie des énergies marines, énergie marémotrice), énergie géothermique, http://www.energies-renouvelables.org (consulté le 20/08/2015).

[307] Energies alternatives : énergies qui se substituent aux énergies fossiles et qui produisent une quantité faible de polluants, http://www.cea.fr/energie/energies-climat-les-defis-de-la-recherche/les-energies-alternatives-au-cea (consulté le 20/08/2015).

l'emporte sur la forme *énergie*. Quant à la forme *énergies*, la courbe existe, avec un « pic » en 2011 expliqué en partie par Fukushima.

Concernant la spécificité des formes-pôles *energy* et *nuclear* cette figure I.3 est marquée par deux « pics » : un en 2008 et un en 2011, ceux-ci peuvent s'expliquer par les raisons suivantes :

- l'année 2008 est marquée par deux accords de coopération nucléaire : d'abord en juin la signature (des États-Unis) avec la Turquie de transferts de technologie, de matériel, de réacteurs et de l'aide en particulier dans le développement d'applications nucléaires dans les domaines de la médecine et de l'agriculture ; puis un deuxième accord en septembre avec l'Inde pour une durée de 40 ans pour des transferts de technologies et fournitures de matériel nucléaire et non-nucléaire,
- l'année 2011 : Fukushima et ses conséquences sur le nucléaire à l'échelle mondiale,
- l'année 2010[308] marque un nouveau démarrage du nucléaire : le président Obama effectue un voyage à New Delhi pour relancer la coopération nucléaire civile.

Pour la forme-pôle *nuclear*, l'augmentation du nombre d'occurrences repart à la hausse dès l'année 2009 contrairement à la forme-pôle *energy* qui continue de décroître jusqu'en 2010. Cette tendance peut s'expliquer de la manière suivante : en janvier 2009, c'est la prise de fonctions du président Obama, qui annonce des changements importants dans la politique énergétique des Etats-Unis. Un projet de loi sur l'énergie prévoyant des mesures pour accroître l'efficacité énergétique et limiter le réchauffement climatique est voté par la Chambre des Représentants en juin 2009, mais rejeté par le Sénat. Autre événement en 2009, les Etats-Unis et l'Union Européenne souhaitent coopérer dans le secteur de l'énergie depuis de nombreuses années. Juin 2009 voit l'aboutissement de cette coopération par la « *proposition des Etats-Unis de créer un conseil de l'énergie UE-Etats-Unis* ».[309]

[308] Centre d'études de sécurité internationale et de maîtrise des armements (Cesim) ONP n°77 Nucléaire, http://www.cesim.fr/observatoire/fr/77/article/120 (consulté le 20/10/2014)
[309] Commission européenne, Energie, http://ec.europa.eu/energy/international/bilateral_cooperation/usa_fr.htm (consulté le 19/10/2014)

Annexe J : dictionnaire d'événements et restitutions par cooccurrences des formes-pôles *nucléaire* et *énergie nucléaire* pour le sous-corpus ENRG_CN pour la période 2010 - 2012

La politique chinoise face à la situation énergétique, à l'augmentation de la consommation d'énergie et aux problèmes environnementaux est présentée succinctement dans l'annexe G, intitulée la politique énergétique chinoise ainsi que dans le chapitre 2, Energies et environnement.

Pour le sous-corpus ENRG_CN, il n'a pas été possible d'obtenir les poly-cooccurrences des formes-pôles *nucléaire* et *énergie nucléaire* principalement pour les raisons suivantes, un volume de données trop important et un grand nombre d'erreurs détectées dû à des signes de ponctuation chinois non reconnues par l'outil Trameur. Pour tous les calculs de cooccurrences, nous avons maintenu les paramètres par défaut : co-freq 2, seuil 10, contexte . !?

Dans les analyses suivantes, nous appliquons notre technique de veille *aveugle*, c'est-à-dire, à partir des formes-pôles *nucléaire* et *énergie nucléaire*, nous rechercherons la signification contextuelle des mots associés, mais le retour au contexte s'avère indispensable pour vérifier toute affirmation présentée ci-dessous.

J.1 Veille et restitutions par cooccurrences de la forme *nucléaire*

Nous allons commencer par l'examen des cooccurrents de la forme-pôle *nucléaire* en un seul caractère chinois. Les résultats sont disponibles dans le chapitre 6, tableau 6.15. Ces cooccurrents peuvent être associés à quatre grandes familles :

- les toponymes associés à des sites nucléaires,
- les grandes catastrophes nucléaires,
- l'activité nucléaire, sa technologie, son fonctionnement,
- les réacteurs.

A la différence des analyses concernant ENRG_FR et ENRG_US, ENRG_CN présente de nombreux cooccurrents se rapportant à l'ensemble des filières nucléaires.

1. Les toponymes

Sites de centrales nucléaires chinoises :

Taishan, Dayawan, Yangjiang, Ling-ao et ***Mayak*** complexe nucléaire russe.

Des noms de pays touchant de près le nucléaire pour diverses raisons :

Iran : les relations diplomatiques entre l'Iran et les autres pays ont connu de graves tensions au sujet de la politique nucléaire menée par l'Iran après le 11 septembre 2001. La communauté internationale soupçonne l'Iran de vouloir se doter de l'arme nucléaire et de mettre au point des armes de destruction massive, et s'inquiète, craintes confirmées par l'AIEA. Les négociations ont été longues pour aboutir à un compromis à l'été 2015.

Chine : elle développe un ambitieux programme nucléaire (voir annexe G)

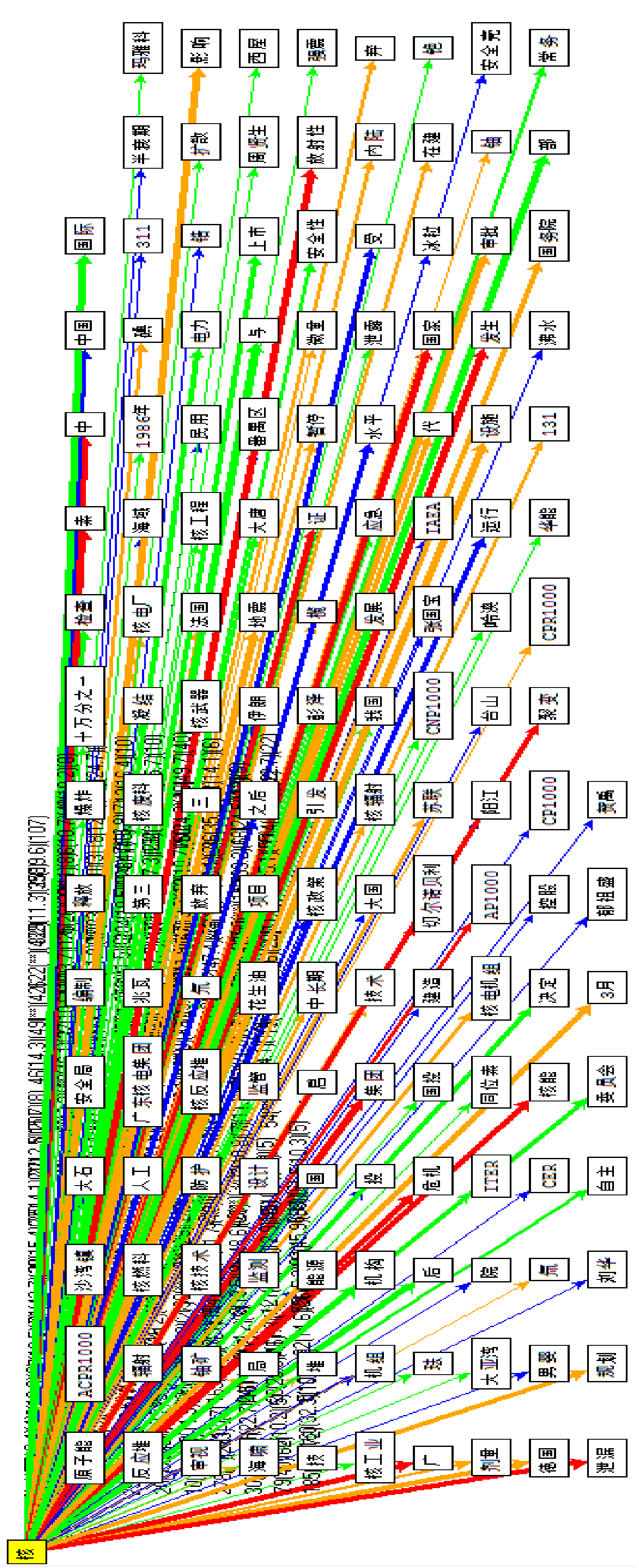


Figure J.1

ENRG_CN : réseau cooccurrentiel du nucléaire 核/hé/nucléaire pour la période de 2010 à 2012

Tableau J.1

ENRG_CN 2010, 2011 et 2012 : équivalent en français des cooccurrents de la forme 核/hé/nucléaire en un seul caractère

原子能 énergie atomique	ACPR1000	沙湾镇 Shawan comté	大石 (大石街) Dashi (comté du district Pānyú)	安全局 Agence nationale de la sécurité (NSA)	编制 système	释放 libérer	爆炸 explosion	十万分之一 un sur dix mille	检查 inspection	素 élément	中 Chine ou milieu ou pendant	中国 Chine	国际 international		
反应堆 réacteur	辐射 radiation ou radioactif	核燃料 combustible nucléaire	人工 artificiel	广东核电集团 CGNP	兆瓦 mégawatt(s) MW	第三 troisième	核废料 déchet nucléaire	凝结 coagulation	核电厂 centrale nucléaire	海峡 détroit	Année 1986	碘 iode	311 (11 mars, date de Fukushima)	半衰期 La demi-vie (d'une substance)	玛雅科 Complexe nucléaire Maïak (Mayak) en Russie
审视 examiner	铀矿 mine d'uranium	technologie nucléaire	防护 protection	核反应堆 réacteur nucléaire	氘 deutérium	放弃 abandonner	三 trois	核武器 armement nucléaire	法国 France	核工程 projet nucléaire	民用 civile	电力 électricité	锆 (gào) zirconium	prolifération ou propagation	影响 affecter ou Influence
海啸 tsunami	局 bureau, office	监测 surveillance	设计 conception	监管 contrôle	花生油 huile d'arachide	项目 Projet	之后 après	伊朗 Iran	地震 séisme	大唐 Datang (producteur d'électricité)	番禺区 Pānyú district	与 avec	上市 coté à la bourse	周贤生 Zhou Xiansheng (homme)	西屋 Westinghouse (US - AP1000)
技 (术) technologie	堆 réacteur	能源 énergie	国 pays	启(动) démarrer	中长期 à long terme	核政策 politique nucléaire	引发 suicider	彭泽 Pengze (comté, site candidat pour centrale nucléaire)	榄 (核) Lǎn Hé (comté)	证 certificat	暂停 suspendre	微量 Oligo- (peu nombreux)	安全性 sécurité	放射性 radioactivité	强震 choc violent
核工业 Industrie nucléaire	机组 unité de	机构 établissement	投 investissement	集团 groupe	技术 technique	大国 grand(s) pays	核辐射 radioactivité nucléaire	我国 la Chine	发展 développement	应急 urgence	水平 niveau	泄漏 fuite	受 à cause de	内陆 intérieur du continent	弃 abandonner
长 centrale ou usine	(阿海) 珐 Areva	后 après	危机 crise	国投 investissement national	建造 construire	切尔诺贝利 Tchernobyl	苏联 URSS	CNP1000	张国宝 Zhang Guobao (ministre énergies)	IAEA	代 génération	国家 pays ou national	冰粒 grésil	在建 en cours de construction	铯 césium
计量 dose	大亚湾 Dayawan Bay (ville)	院 institut	ITER	同位素 isotope	核电机组 système de production de l'électricité nucléaire	AP1000	阳江 Yangjiang (ville)	台山 Taishan (ville)	岭澳 ling-ao (ville)	运行 opération	设施 équipement	发生 avoir lieu	审批 approbation	铀 uranium	安全壳 couvercle de confinement
德国 Allemagne	男婴 bébé de sexe masculin	氚 tritium	CER (Certified Emission Reduction)	核能 énergie nucléaire	决定 décider	控股 holding	CP1000	聚变 fusion	CPR1000	华能 Huaneng groupe (producteur d'électricité)	131	废水 eaux usées	国务院 conseil d'état	部 département	常务 comité permanent
泄漏 fuite	规划 planification	刘华 Liuhua (ingénieur en Chef)	自主 autonome	委员会 comité	三月 mars	郁祖盛 Yu Zusheng (ministre env)	贺禹 Hè Yù (PDG de CGN)								

Allemagne : après Fukushima, le pays décide de sortir du nucléaire
France : le pays du nucléaire par excellence
Cooccurrents associés : *pays, national, etc.*

2. Les catastrophes

Fukushima **:** mentionné sous un **numéro 311** qui représente la date de la catastrophe de Fukushima le 11 mars.
Cooccurrents associés : *tsunami, sécurité, surveillance, radiation, fuite, explosion, mars.*
Tchernobyl : autre catastrophe nucléaire mentionné.
Cooccurrents associés : *URSS, année 1986*, *radiation, radioactivité.*

3. L'activité nucléaire

Cooccurrents associés à cette activité :
Industrie nucléaire, énergie atomique, centrale, technologie nucléaire
Conséquences des catastrophes :
Radiation, surveillance, protection, sécurité, radioactivité, IAEA, couverture de confinement

Sécurité : l'accident de Fukushima a marqué les esprits et le gouvernement chinois a réagi face à cette catastrophe en prenant certaines mesures au niveau de la sûreté ou de la sécurité. Des accords ont été signés sur la base de partenariat pour des partages d'expériences et de savoir-faire comme par exemple en janvier 2015 entre EDF et CGN ou encore entre EDF et Huadian, un des premiers électriciens chinois, « *(…) afin de partager leur retour d'expérience sur l'exploitation et l'ingénierie des parcs nucléaires existants et pour maintenir les plus hauts niveaux de sûreté (…) accords signés avec nos partenaires chinois historiques viennent approfondir des coopérations existantes et poser les bases de nouveaux projets communs (…) »* a déclaré le Président-Directeur Général d'EDF Jean-Bernard Levy[310].

Centrale *:* destinée à produire de *« l'électricité à partir d'un combustible nucléaire (…), (…) il existe plusieurs familles de réacteurs, que l'on appelle filières. Quatre constituants principaux sont nécessaires pour concevoir un cœur de réacteur : un combustible dans lequel se produit la fission ; un fluide caloporteur qui transporte la chaleur hors du réacteur ; un modérateur (sauf pour les réacteurs à neutrons rapides) qui permet de ralentir les neutrons ; des barres de commande qui contrôlent la réaction en chaîne (…) »*[311].
Cooccurrents associés *: uranium, mine d'uranium, combustible nucléaire, tritium*[312]*, isotope*[313]*, deutérium, césium*[314]

4. La technologie des réacteurs et leurs constructeurs (se reporter à l'annexe G)

[310] https://www.lenergieenquestions.fr/tag/chine/ (consulté le 1/03/2015).
[311] http://www.cea.fr/jeunes/themes/l-energie-nucleaire/le-fonctionnement-d-un-reacteur-nucleaire/les-differents-types-de-reacteurs (consulté le 27/08/2015).
[312] deutérium et tritium : atomes très légers, tous deux isotopes de l'hydrogène
[313] isotopes : atomes qui possèdent le même nombre d'électrons, mais un nombre différent de neutrons
[314] césium : un élément radioactif

Annexe J

Réacteurs[315] ***nucléaires*** (Jaouen et Béroux, 2012 : 113) : classés par type, par *puissance* exprimée en *MWe* (mégawatts électrique) et par génération : 1ère génération, 2ème génération, *3ème génération* et 4ème génération.

- Réacteur à eau pressurisée (ou REP) : eau sous pression et le combustible utilisé est de l'uranium enrichi, réacteur qui met l'accent sur la *sûreté* et à la *sécurité* (résistance renforcée aux agressions externes, type chute d'avion, *choc violent*) :
  - REP réacteur à eau pressurisée : type *CPR1000, AP1000, CAP1400,*
  - PWR pressurized water reactor en anglais
  - EPR réacteur pressurisé européen
- Réacteur à eau bouillante (ou REB) : eau mais non pressurisée et combustible utilisé est de l'uranium enrichi.
- Réacteur à eau lourde : combustible utilisé est de l'uranium naturel.
- Réacteur à neutrons rapides (ou RNR) : ont été conçus pour utiliser la matière fissile (l'uranium et le plutonium) comme combustible nucléaire,
- Réacteurs à caloporteur gaz (ou RCG) : susceptibles de permettre la réalisation d'unités de petite taille (de 100 à 300 MWe), économiques et sûres.

AP1000 (Harari et Chauvin, 2012 :142) : un nouveau type de réacteur *de* 3ème génération + développé par la compagnie américaine Westinghouse Electric Corporation, le premier de cette génération, un réacteur à eau pressurisée qui fonctionne suivant les mêmes principes que ceux du parc nucléaire français. Toutes les unités AP1000, soit quatre en Chine devraient être opérationnelles d'ici 2016. La construction des unités, deux à Sanmen et deux à Haiyang dans la province du Shandong, ont été autorisées par Westinghouse et son partenaire (le groupe Shaw), en septembre 2007. Ce réacteur est plus compact que les autres, ce qui permettrait d'utiliser moins de béton et de ferraillages pour sa construction. Quatre exemplaires de l'AP100 sont déjà en construction en Chine depuis 2009 (centrales de Haïyang et Sanmen).
Cooccurrents associés à *AP1000 : Westinghouse, réacteurs*

CAP1400 : autre type de réacteur[316], développé par la Chine, sa conception est basée sur le réacteur AP1000 de Westinghouse Electric Co. La Chine possède les droits de propriété intellectuelle sur les CAP1400, ce qui permet d'exporter le réacteur. *«La technologie est en cours d'évaluation (...) pourrait être construit d'ici fin 2013 au plus tôt*», a déclaré Gu Jun, directeur général de State Nuclear Power.
Autres types de réacteurs, autres *technologies* : *ACPR1000, CP1000, CNP1000* (se reporter à l'annexe G).
Nous remarquons que la forme EPR est absente de la figure J.1, alors que deux EPR sont en cours de construction en Chine (en construction, couverture de confinement). Par contre, la forme AREVA est présente.

Westinghouse : cette technologie AP1000 est liée à Westinghouse Electric Company[317], premier fournisseur au monde de la technologie nucléaire sûre et innovante, entreprise américaine créée en 1886.
AREVA *:* constructeur français développant entre autres la technologie EPR, « (…) *le groupe propose aux électriciens une offre qui couvre toutes les étapes du cycle du combustible, la conception et la*

[315] http://jeunes.edf.com/article/les-differents-types-de-reacteurs-nucleaires,64 (consulté le 27/08/2015).
[316] Source : le Quotidien du peuple en ligne, 04/02/2013, http://french.peopledaily.com.cn/Economie/8120966.html (consulté le 15/02/2015).
[317] http://www.westinghousenuclear.com (consulté le 19/02/2015).

construction de réacteurs nucléaires ainsi que les services pour leur exploitation. AREVA investit également dans les énergies renouvelables afin de développer en partenariat des solutions à fort contenu technologique (…) »[318].

CGN : construit deux EPR à Taïshan, mais un doute subsiste sur la résistance de l'acier de ses cuves où se produira la fission atomique puisqu'elles ont été forgées en France, comme celle de Flamanville[319].
Cooccurrents associés : *réacteurs, réacteurs nucléaires, trois, génération, troisième, technologie*, etc.

Producteur d'électricité

Datang : un des cinq grands producteurs d'électricité à partir du charbon, et un peu hydraulique, situé dans le nord da la Chine, société qui alimente en électricité toute la région de Pékin, Tianjin et côtée en bourse.
D'autres cooccurrents sont présents dans le tableau 6.15 du chapitre 6, mais ceux-ci sont difficilement associables à une famille sans un retour au contexte.

J.2 Veille et restitutions par cooccurrences de la forme *énergie nucléaire*
Nous allons continuer notre étude par l'examen des cooccurrents de la forme *énergie nucléaire* en un seul caractère chinois. Les résultats sont disponibles dans le chapitre 6, tableau 6.13. Ces cooccurrents peuvent être aussi associés à quatre grandes familles :

- les types et sources d'énergie,
- les grandes catastrophes nucléaires,
- l'activité nucléaire, sa technologie, son fonctionnement,
- l'électricité.

Les types et sources d'énergie représentent une proportion importante de cooccurrents.

5. Les types et sources d'énergie[320]

Fossile : ce sont les sources d'énergie principales en Chine
Cooccurrents associés : *charbon, houille, fuel, gaz naturel, gaz naturel, fuel (mazout), CCS.*
Pétrole et gaz naturel sont importés de l'étranger pour 30% de la consommation (Guermond, 2007 ; Guermond et Ma, 2011). Les pays, partenaires privilégiés de la Chine, sont majoritairement situés en Asie centrale (Lafargue, 2013 : 24) comme le Turkménistan.
La mise en valeur des gaz non conventionnels comme le gaz de schiste est assez compliquée du fait de l'absence d'eau dans les régions d'extraction. Des aides gouvernementales seront apportées à ce secteur (Grésillon, 2012).
Renouvelable :
Le pays a lancé ses politiques d'énergies renouvelables dès 2000 et s'est fixé comme objectif une couverture de la demande à concurrence de 10% en 2010, et de 15% d'ici 2020 (Chang, Zhao et al, 2012). L'Empire du Milieu serait ainsi en mesure d'augmenter la part du renouvelable dans son mix énergétique de 13 à 26% à l'horizon 2030 (Hill, 2014). L'opinion publique a évolué devant l'ampleur

[318] http://www.areva.com/FR/groupe-57/leader-mondial-des-metiers-de-l-energie-nucleaire-et-energies-renouvelables.html (consulté le 19/02/2015).
[319] Le Monde Economie, la Chine lance un concurrent de l'EPR français, publié le 6/05/2015, http://www.lemonde.fr/economie/article/2015/05/06/nucleaire-la-chine-lance-un-concurrent-de-l-epr-francais_4628880_3234.html (consulté le 27/08/2015).
[320] Pour plus de renseignements sur la question des types et sources d'énergie, se reporter à l'annexe G

de la pollution et les conséquences (Chen, Ebenstein et al, 2013) en sont parfois dramatiques (nappes phréatiques souillées, rejets toxiques dans les fleuves, etc.), (se reporter au chapitre 6)
Cooccurrents associés : *nouvelle, nouveaux, biologie*, *énergie solaire*, é*nergie éolienne*, é*nergie de l'eau*, é*nergie atomique, industrie nucléaire, technologie nucléaire, uranium, énergie de l'eau, fossile, biologie, renouvelable, nouvelle, propre, substitution*
Energie de l'eau :
En dépit de sa part de marché relativement modeste, le secteur de l'énergie éolienne offshore a continué de croître, en s'appuyant sur l'utilisation de turbines plus volumineuses (Sawin, Bhattacharya, Martinot et al, 2012).

6. Les catastrophes nucléaires

Cooccurrents associés : *accidents, Tchernobyl, URSS, Fukushima, Naoto Kan, 11 mars, Japon, crise, fuite*

7. L'activité nucléaire, sa technologie, son fonctionnement

Cooccurrents associés : *centrale nucléaire, technologie nucléaire, sécurité, futur, génération, réacteur, sûreté, industrie nucléaire, surgénération, nucléaire, Allemagne, France*

8. L'électricité

Production électricité, Energie électrique : la Chine possédait, à la fin 2011, une capacité électrique issue des énergies renouvelables supérieure à toutes les autres nations, selon les estimations, 282 GW (Sawin, Bhattacharya, Martinot et al, 2012).

Cooccurrents associés : *électricité, électricité nucléaire, produire de l'électricité, centrale d'électricité, alimentation, développement,*

Annexe K : enquête de terrain sur le thème nucléaire dans le Cotentin

J'ai eu l'occasion lors d'un séjour dans le Cotentin de rencontrer des Cherbourgeois qui m'ont fait part de leurs témoignages concernant le nucléaire dans leur région.

« (…) Au milieu des années 60, le nucléaire arrive dans le nord Cotentin, plus exactement dans La Hague, une terre « perdue » au bout du monde, le Far-West comme disaient certains! Utilisons plutôt le terme géographique de finisterre pour désigner cette extrémité de terre plongeant dans la mer. Ce site est aussi choisi pour son socle géologique ancien et stable à l'abri des tremblements de terre, et aussi pour ses courants marins violents avec le raz Blanchard, pour ses marées à fort marnage et pour ses vents forts propices à la dispersion et à l'évacuation des effluents radioactifs (…).

Les landes sont achetées à « bon prix » à leurs propriétaires trop « heureux » de pouvoir se débarrasser de ces terres incultes (source familiale).

Mais au fait vendre ces terres pour quoi faire ? Les Cotentinois disaient « c'est pour construire une usine atomique ! ». A part, Hiroshima, Nagasaki, la bombe atomique, tout cela est bien loin des préoccupations normandes. On parlait alors de l'atome, du CEA, de déchets, mais le mot nucléaire était peu utilisé. En bref, quelque chose d'inconnu, de mystérieux, mais qui rapportait !!! Ainsi l'usine de retraitement des déchets radioactifs est sortie peu à peu de « terre » et devient opérationnelle en 1966. Elle retraite notamment les déchets français mais aussi les allemands, italiens, hollandais, japonais entre autres.

Les « gens du cru » étaient très contents de cette usine, car elle apportait du travail dans un endroit « déshérité ». De nombreux locaux y trouvent un emploi, une véritable aubaine ! De plus, des personnels extérieurs arrivent également. C'est une petite ville qui se créée à Beaumont.

Peu de contestations au départ, puis début des années 70 les militants « antinucléaires » s'organisent et créent le comité contre la pollution atomique dans la Hague rejoint par Didier Anger, membre fondateur du parti « les Verts ». Les premiers rassemblements, surtout constitués par des étudiants et des intellectuels, ont un impact très faible, la population locale ne suit pas (Zonabend, 2014) *; cette usine est « une manne d'or » pour l'économie locale qui souffre de désertification mais aussi pour toutes les communes environnantes qui bénéficient des taxes versées par l'usine. Celle-ci emploiera jusqu'à 6000 employés et c'est toute l'économie locale qui se développe avec la mise en place d'infrastructures modernes (constructions de routes, création de logements, d'écoles, de centres sportifs, agrandissement des services publics, etc.). Le « bout du monde » sort de l'ombre !*

Les mouvements antinucléaires s'intensifient dans les années 75-80 avec l'arrivée des déchets japonais et du mouvement Greenpeace. A chaque fois, des opérations « coup de poing » sont menées. Par exemple, celle dans les années 90, lors d'un débarquement de containers japonais en gare maritime de Cherbourg, les anti-nucléaires avaient pris d'assaut une grue servant à leur déchargement et s'en était suivie une bataille rangée avec les forces de l'ordre. Autres questions, dont on reparle périodiquement, les rejets d'eaux polluées dans le courant du raz Blanchard, mais aussi des fuites sur le site de stockage des fûts radioactifs à Digulleville et de taux de radioactivité parfois élevé autour du site. Mais à quoi bon !

Et puis ce fut la construction à Flamanville de l'usine d'électricité nucléaire EDF sur le site d'une ancienne mine de fer sous-marine qui fonctionna jusqu'au début des années 1960, puis c'est la décision de la construction d'un EPR.

Actuellement, les régionaux s'opposent surtout à l'implantation des lignes à haute tension qui provoquent des perturbations dans le monde agricole.

Annexe K

Mais en cas d'incident ou d'accident grave, que doit faire la population ? Cherbourg et son agglomération ne sont situées qu'à une vingtaine de kilomètres de la Hague.
Les habitants ont-ils reçus des consignes particulières en « cas de pépins » ? La réponse est NON, en tout cas pour l'agglomération de Cherbourg, aucune consigne particulière n'a été donnée.
Et pourtant, on n'a pas peur d'y vivre, ni même de se baigner dans la baie d'Ecalgrain au cœur du nucléaire bas-normand. C'est un « petit coin », où il fait bon vivre.
Voilà en quelques mots le ressenti sur le nucléaire! ».

Annexe L : tableau récapitulatif des formes communes des trois sous-corpus ENRG

Tableau L.1

ENRG 2010, 2011 et 2012 : groupes de formes communes (liste non exhaustive)

	ENRG_FR forme *nucléaire* = forme *énergie* (poly-cooccurrences identiques)	**ENRG_US forme *nucléaire* (peu de poly-cooccurrences**	**ENRG_US forme *énergie* (beaucoup de poly-cooccurrences)**	**ENRG_CN forme *énergie nucléaire* par cooccurrences**	**ENRG_CN forme *nucléaire* par cooccurrences**
Toponymes	en grand nombre, nom sites de centrales et de centres de recherche	en petit nombre, seulement noms de sites de centrales	Aucun	Noms de pays	Taishan Yangjiang Dayawan LingAo Complexe de Mayak (Russie) Autres villes
Noms des catastrophes	Three Mile Island Tchernobyl Fukushima tsunami	Three Mile Island Tchernobyl Fukushima	Aucun	accidents, Tchernobyl, URSS, Fukushima, Naoto Kan, 11 mars, Japon, crise, fuite	urgence, fuite, Tchernobyl Année 1986 311 = 11 mars Fukushima, séisme, tsunami
Hommes politiques Institutions étatiques	principalement français : François Hollande Eric Besson Nicolas Sarkozy Angela Merkel	Aucun	Barak Obama Dick Cheney Bush Congress White house Senate Washington		3 noms de politiques chinois, de dirigeants de société Conseil d'état
Noms de pays	Allemagne, France, Iran, Hollande, Japon	Germany, Italy, France, Japon	Aucun	Pays (nom générique), France, URSS, Chine	
Politique	Référendum Socialiste CGT	Aucun	législation climat, Bill signed tax carbon dioxyde Investissements Policy Bills	Naoto Kan	Investissements Investissements/national Politique nucléaire International
Puissance	Electrique	Electricité	Electricité		Production électricité
Opérateurs Sociétés	Areva, EDF, TEPCO	Aucun	Aucun		CGNP, Huaneng, Datang, Guotou
Mouvements écologistes	EELV, Greenpeace, Ecologie	Aucun	Aucun	Aucun	
Organismes dont surveillance	ASN IRSN CERN	Aucun	Aucun		IAEA,
Termes en rapport avec l'actualité nucléaire	convoi coût sécurité, audit démantèlement incident accident sortir du nucléaire	Aucun	Aucun		Sécurité, Sûreté, industrie nucléaire, centrale, énergie atomique, mine d'uranium, combustible nucléaire, technologie nucléaire, surveillance, tritium, protection, césium, uranium, deutérium, isotope, fusion, bébé de sexe masculin, système de production d'électricité nucléaire
Termes nucléaires	Fusion	Plutonium,		Fission	

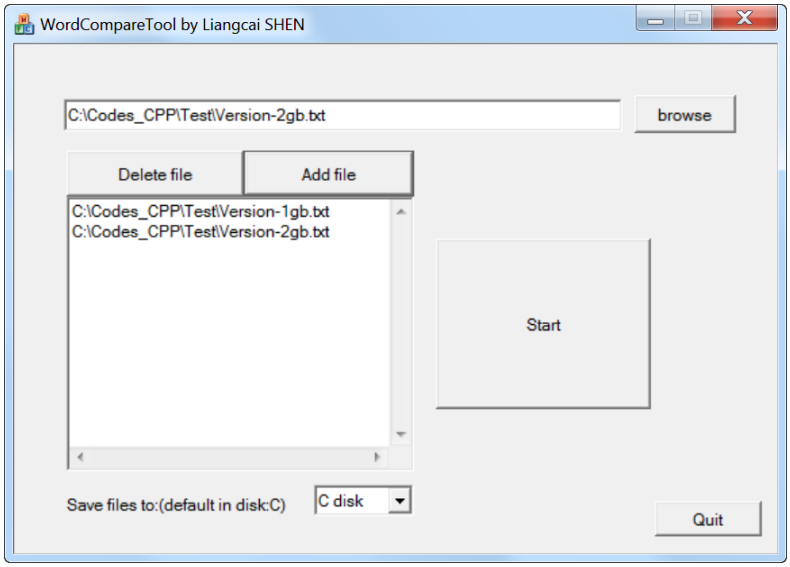
	Fission Uranium Radioactivité Nucléaire Propulsion Combustible Atome	Uranium, radiation, nuclear Atomic		Atomique Radioactif Uranium Tritium Radioactivité Armement nucléaire Césium Nucléaire	
Termes liés aux activités	retraitement Déchets Stockage Centrales	Aucun	Aucun	Production d'électricité, électricité, électricité nucléaire, centrale d'électricité, alimentation, développement, centrale nucléaire, technologie nucléaire, sécurité, futur, génération, réacteur, sûreté, industrie nucléaire, surgénération, nucléaire, Allemagne, France	Usine Industrie nucléaire Technologies/nucléaire Protection
Réacteurs	Réacteurs Nucléaire Power Génération Turbine EPR	réacteurs power aucun type	Aucun		Réacteurs Trois Troisième Génération ACP1000 (CGN) AP1000 (Toshsiba) CAP1400 CPR1000 CP1000 CNP1000 MWe (mégawatt) Unité de Areva (constructeur) Westinghouse (constructeur) CGN (China General Nuclear constructeur) Conception ITER En cours de construction
Sources et types d'énergie	Energies charbon Renouvelables éolien parc d'éoliennes	Energy coal gas sources alternatives clean technologies wind	green, clean solar, sources renewable, geothermal (heating), resources, natural, fossils, oil, coal, Gas, Exploration, Reforestation, biomass, Fuel	énergie atomique, uranium, énergie de l'eau, fossile, biologie, renouvelable, nouvelle, propre, substitution, gaz naturel, fuel (mazout), charbon (houille), énergie éolienne, énergie solaire, CCS	

Annexe M : programmes informatiques

Programmes d'extraction et de construction du corpus parallèle chinois-anglais

Comparateur de segmentation du chinois

WordCompareTool est un outil spécifique codé en C++ et conçu dans l'objectif de comparer et afficher les différences des deux résultats de textes chinois segmentés en mots par deux segmenteurs différents. Notre outil peut également comparer simultanément, deux à deux, plusieurs fichiers de résultats.



Fenêtre de saisie

Constitution des deux corpus ENRG et CLRG

De nombreux programmes informatiques ont été nécessaires pour la constitution et les traitements de nos corpus. L'analyse des procédures d'aspiration automatiques étant semblable dans les outils conçus, seule l'explication détaillée relative au corpus parallèle sera décrite.

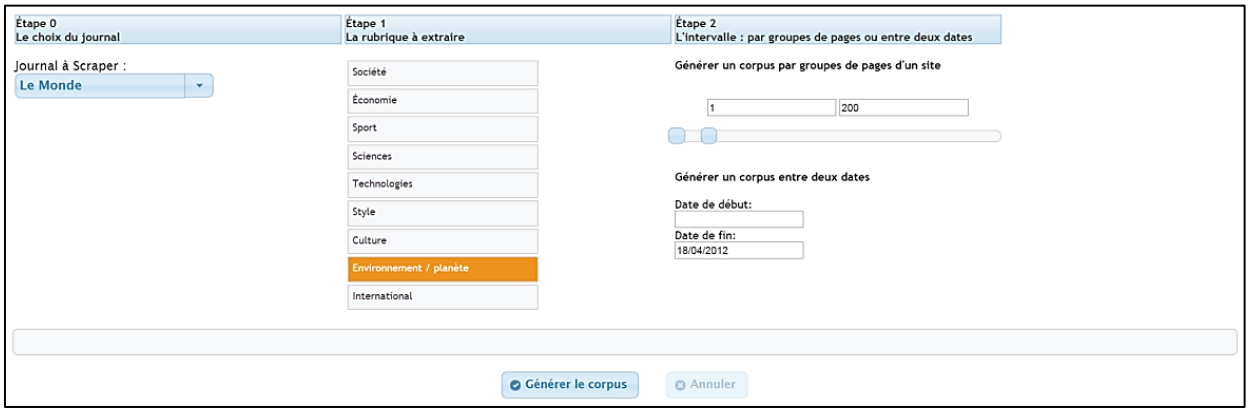
Choix du langage de programmation

Dans les traitements de données textuelles, nos expériences révèlent que par rapport aux langages de programmation C et C++, malgré leur robustesse, le langage Python présente un bon nombre d'avantages en particulier une écriture simple et efficace des codes informatiques. De plus, un grand nombre de bibliothèques et modules informatiques en *Open Sources* sont disponibles sur la toile permettant ainsi d'accélérer l'avancement de nos développements informatiques. Les projets dédiés aux *Web Scraping* et *Comparateur de segmentation du chinois* en C++ furent laborieux à développer, notamment à cause de la complexité du langage, c'est pourquoi, par la suite tous les autres programmes de traitements et post-traitements (nettoyage, conversion de signes, etc.) sont codés en Python avec l'interface graphique Qt et dans l'environnement Eric (semblable à l'environnement Eclipse ou NetBeans).

Corpus comparable

Le corpus comparable constitué de trois sous-corpus, à partir des journaux Le Monde en français et le *New York Times* en anglais, quant au sous-corpus chinois, celui-ci a dû être réalisé en deux temps : d'abord, à partir du site Sina.com.cn, puis le site QQ.com.

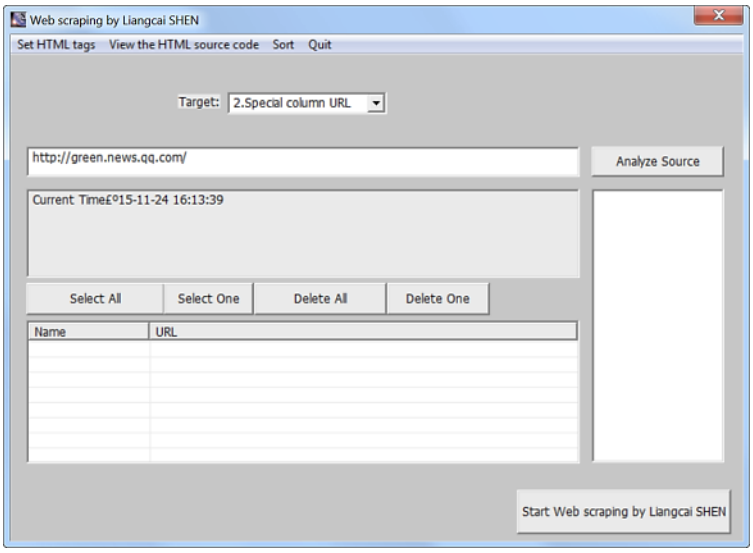
Pour ce faire, nous avons créé un ensemble de programmes nommé *Serveur Scraping*, opérationnels sous serveur local de type Apache, codés en Perl, PHP et Ajax avec une interface graphique écrite en HTML avec l'encodage UTF-8, jQuery AJAX.



Fenêtre de l'interface graphique du *Serveur Scraping*

A partir de l'interface ci-dessus, trois programmes spécifiques en Perl, PHP et Ajax correspondant aux trois sources ont été exécutés successivement afin de constituer la totalité des sous-corpus français, anglais (ENRG_FR et ENRG_US) et une partie du sous-corpus chinois.

La partie restante du sous-corpus chinois provient du site QQ.com. En raison de ces origines éditoriales diverses et variées, un logiciel spécifique en C++, langage reconnu pour sa robustesse, a été conçu, appelé *Web Scraping*. Les deux parties en chinois forment le sous-corpus ENRG_CN.



Fenêtre de saisie des paramètres du logiciel Web Scraping

Ainsi, le corpus comparable ENRG est composé des différents résultats obtenus par ces programmes.

Cependant, au fur et à mesure de l'avancement du projet, nous nous sommes heurtés à des difficultés de structuration des données textuelles obtenues. En effet, les sous-corpus générés par les programmes sont stockés sous format .TXT, format exigé par Lexico 3 et Trameur. Afin de faciliter les traitements de tri, les données ont été basculées dans une base de données de type MySQL-Serveur. Néanmoins, ce format manifeste des contraintes techniques et de manipulation, en particulier dans le cas des doublons et de la recherche d'informations. Pour ces raisons, nous avons élaboré d'autres programmes en Python permettant de structurer de manière normalisée et de trier chronologiquement ces données au format XML.

Avantages du format XML

Le balisage libre du langage XML permet de stocker et transmettre tous types de données entre applications, de faciliter les recherches et d'enrichir le Web sémantique. Son système d'annotation peut être proche de celui du *Dublin Core* et ainsi nous avons créé une norme pour nos corpus.

Tableau des balises pour nos corpus

Meta-information des articles des journaux	Balises en XML dans nos corpus	Types de données stockées entre les balises
Titre	<title>	Type texte
Auteur	<author>	Type texte
Corps de l'article	<body>	<p> pour paragraphe, contenu en type texte
Date de publication	<publish_date>	Temps au format ISO 8601
Nom du média	<media>	Type texte
URL	<url>	Uniform Resource Locator
Editeur	<editor>	Type texte
Photographe	<photographer>	Type texte
Langue de rédaction	<language>	Balise aux normes ISO 639-1

Exemple d'un corpus au format XML

```
<?xml version="1.0" encoding="UTF-8"?>
<article>
  <header>
    <title>Titre text</title>
    <publish_date>2008-08-01</publish_date>
    <author>Auteur Nom</author>
    <editor>Editeur Nom</editor>
    <photographer>Photographe Name</photographer>
    <media>Media Name</media>
    <language>Français</language>
    <url>http://www.sample.org</url>
  </header>
  <body>
    <p>Texte line1</p>
    <p>Texte line2</p>
    <p>Texte line3</p>
  </body>
</article>
```

Le balisage peut se préciser pour également des hyperliens tels que les médias, une annotation @url entre les balises <media> indiquera son adresse.

Structure normalisée de nos corpus

Les deux corpus sont constitués majoritairement par les articles des journaux, il est donc nécessaire de créer une structure normalisée et générique afin de faciliter les post-traitements, manipulations, conversions, et diffusions qui les concernent. Cette structure, ainsi retransformée et uniformisée au format XML en UTF-8, illustrée dans la figure[321] ci-dessous, permet également de corroborer les développements informatiques.

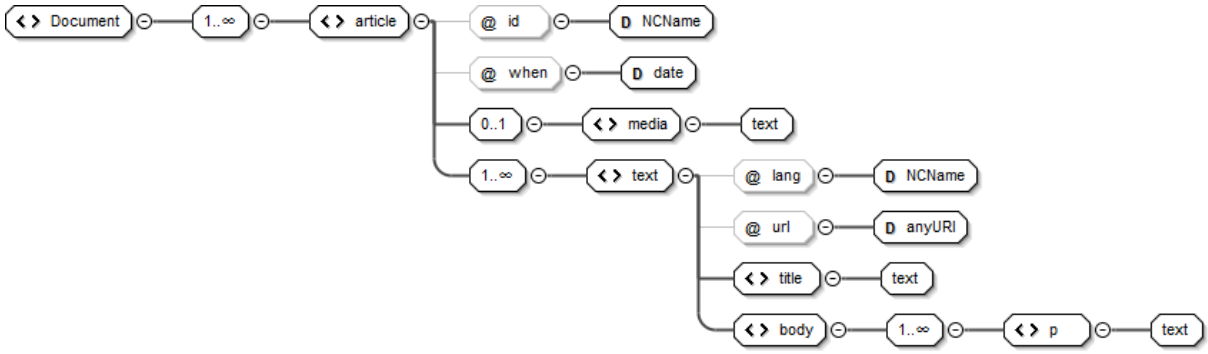


Figure de la structure générique des corpus

- <Document> est la racine du fichier.
- <article> indique le nœud de la racine et stocke le texte de chaque article.
- @id est une clé à valeur alphanumérique unique attribuée à chaque article des corpus, permettant ainsi de les repérer ultérieurement.
- <publish_date>, représenté par l'attribut @when dans la figure, est la balise portant la propriété temporelle de cette <article>, dénote ainsi la date à laquelle l'article est publié.
- La balise <article> est transformée en <text> qui marque le bloc où sont stockés le titre et le corps de chaque article.
- Les balises <language> et <url> sont devenues attributs de la balise <text>, et représentées par @lang et @url. Cette pratique nous permet d'enregistrer un corpus en plusieurs langues.
- La balise <title> devient un nœud de la balise <text> et stocke le titre de chaque article <text>.
- <body> est réservé au corps de chaque article.
- <p> désigne un paragraphe d'un article.

Un exemple concret d'un corpus au format XML est illustré dans la figure ci-dessous :

[321] Figure générée par *Oxygen XML Editor*, http://www.oxygenxml.com

```
<Document>
  <article id="lemonde_1" when="2012-04-17">
    <media>CNN</media>
    <text lang="fr"
      url="http://www.example.com/1.html">
      <title>Dernier vol...</title>
      <body>
        <p>Attachée sur le dos d'un Boeing 747 modifié..."</p>
        <p>Peu après 15 heures (GMT), la plus ancienne...</p>
        <p>Discovery avait été lancée pour la première...</p>
      </body>
    </text>
  </article>
  <article id="lemonde_2" when="2012-04-17">
    <media>BBC</media>
    <text lang="fr"
      url="http://www.example.com/2.html">
      <title>Carburants : les prix...</title>
      <body>
        <p>Une essence à deux euros le litre, c'est ...</p>
        <p>Pourquoi les prix des carburants atteignent-ils...</p>
        <p>Lire : "Printemps arabe et embargo iranien...</p>
      </body>
    </text>
  </article>
</Document>
```

Un exemple concret d'un corpus au format XML

Corpus parallèle

Grâce aux expériences lors de la construction du corpus comparable ENRG, nous avons appliqué les mêmes concepts, normes et format pour la constitution du corpus parallèle CLRG.

Analyse des pages de navigation pour l'aspiration automatique des articles bilingues

Chaque journal et/ou média présente une politique spécifique relative à la construction de leurs pages de navigation et/ou de rubriques. Une analyse approfondie de ces pages est la clé de chaque conception informatique. La figure ci-dessous montre quelle est la logique de leur construction et comment les pages de navigation se définissent dans le site bilingue *Chinadialogue.*

https://www.chinadialogue.net/article?page=1

https://www.chinadialogue.net/article?page=2

https://www.chinadialogue.net/article?page=3

...

Figure illustrant la logique de construction du site *Chinadialogue*

En effet, les articles bilingues sont rangés et classés sur ces pages de navigation, et c'est en analysant leurs codes sources en HTML que nous pouvons obtenir les informations concernant les URL, langues, titre, résumé, corps, auteur, version, etc., comme l'illustrent les figures ci-dessous, permettant par la suite de les aspirer automatiquement.

Extrait d'un article bilingue du site *Chinadialogue*

URL d'un article du site *Chinadialogue*

Grâce à l'outil de développement de Chrome, nous pouvons repérer assez rapidement les différentes balises où sont stockées les différentes parties des articles qui nous intéressent mais éparpillés dans l'arborescence des codes HTML.

La bibliothèque *Beautiful Soup* de Python, et en particulier la méthode *get_urls()*, permet de parcourir, analyser et récupérer efficacement les différentes parties textuelles (titres, paragraphes, etc.) et informations (date, URL, etc.) de nos articles. Les codes sont disponibles dans le fichier Spider.py.

Analyse des pages des articles

Pour ce faire, il faudrait commencer par comprendre où sont stockés les différents textes de chaque article en deux langues dans une page HTML, puis les récupérer et les sauvegarder proprement dans nos fichiers XML.

Titre d'un article du site *Chinadialogue*

Une fois que les balises spécifiques aux contenus textuels des articles sont repérées, des analyses des codes sources par l'outil *Beautiful Soup* seront effectuées. La méthode *make_xml()* va permettre de nous aider à générer le corpus au format XML selon les normes que nous avons définies.

L'extrait ci-dessous illustre un article bilingue sauvegardé dans un fichier XML en UTF-8 selon nos méthodes informatiques. En effet, le repérage de la langue s'effectue au début de l'analyse des URL des pages ; par la suite, le stockage des textes se fait en deux langues parallèlement. Ainsi, la balise <text> est annotée par @lang qui permet de distinguer les deux langues de rédaction.

```
<?xml version="1.0" encoding="UTF-8"?>
<Document>
 <article id="chinadialogue-00001" when="2006-06-23">
  <text lang="zh"
   url="https://www.chinadialogue.net/article/show/single/ch/132-Climate-Change-the-
real-threat-to-security">
   <title>气候变化：真正的安全威胁</title>
   <body>
    <p>- 气候变化的影响，可能会导致居民不...</p>
    <p>- 气候变化给所有国家的安全带来的长...</p>
    <p>- 但是，我们不应该将增加核能的利用...</p>
   </body>
  </text>
  <text lang="en"
   url="https://www.chinadialogue.net/article/show/single/en/132-Climate-Change-the-
real-threat-to-security">
   <title>Climate Change: the real threat to security</title>
   <body>
    <p>- The effects of climate change are likely ....</p>
    <p>- This has long-term security...</p>
    <p>- However, the response to climate …</p>
   </body>
  </text>
 </article>
</Document>
```

Extrait d'un article du corpus bilingue CLRG

Développement des outils de conversion au format Lexico 3 et de tri chronologique

Les logiciels Lexico 3 et Trameur exigent un format spécifique avant tout traitement textométrique.

Afin d'uniformiser le format de nos corpus générés par différents outils et de faciliter les post-traitements de tri, extractions et regroupements, un premier outil spécifique a été conçu en Python, nommé *LexicoConvertor*.

```
<year=2006>
<article=chinadialogue-00004>
title:Global warming: a clear and present danger
url:https://www.chinadialogue.net/article/show/single/en/76-G
# The science of climate change is not a new subject. Indeed,
# An Irish-British scientist, John Tyndall, discovered in 186
# The first global warming calculations were offered in 1896 
# He wasn't far out. The most recent calculation, based on en
# At the higher end, the impact of such a temperature rise wo
# I set out this history to make the point that our current u
# Scientific understanding has been enhanced greatly by the w
# Nonetheless, it is often reported that scientists themselve
# Beyond any reasonable doubt, climate change is happening. M
# Unmitigated climate change will both magnify humanity's exi
# So if scientific opinion is so united on these points why d
# Part of the answer is in the nature of the media itself, wh
# There is also an issue that some, including some politician
# Sceptics and evidence
# A few words are appropriate on the theme of the "climate ch
# First, there is a very small group of serious scientists wh
# Second, there is another small group of scientists who appe
# Third, there is a very vocal group of professional lobbyist
# In summary, it is quite clear that the balance of internati
# The Author: Professor Sir David King is the chief scientifi
# _end_
```

Figure d'un article anglais extrait de *Chinadialogue.net* avant conversion

Comme nous pouvons le constater, dans la figure ci-dessus, chaque article des corpus généré par nos outils au format Lexico3 est marqué par une balise de début <article=…>, puis les paragraphes débutent par le signe #, et l'article se termine par une chaîne spécifique de caractères _end_.

L'outil *LexicoConvertor* permet d'extraire les textes de nos corpus par année, par mois ou par jour puis de les trier chronologiquement.

Fenêtre de saisie de l'outil *LexicoConvertor*

Prérequis pour le développement

Annexe M

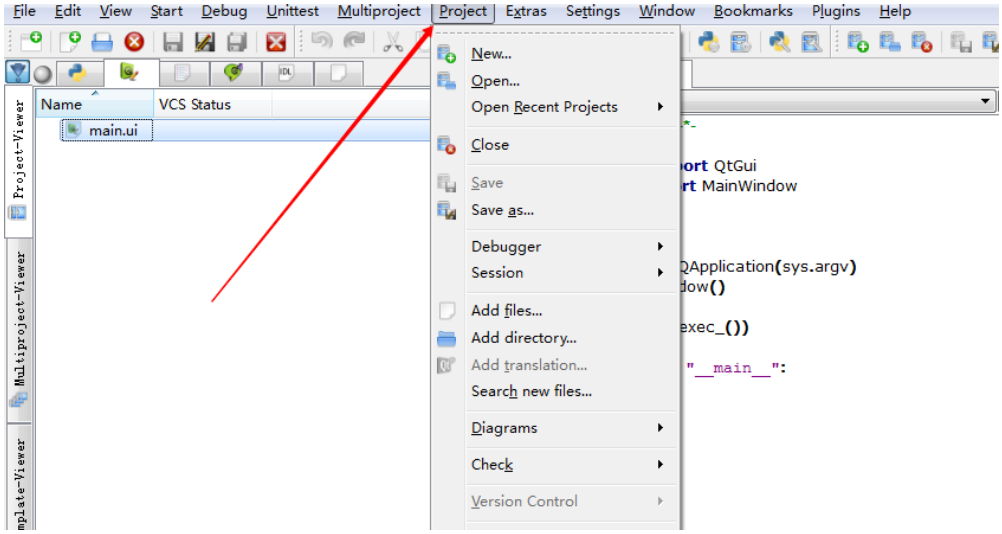
Un environnement de développement a été défini :

PyQt4: cette version de Qt Python permet de concevoir l'interface graphique (GUI). https://riverbankcomputing.com/

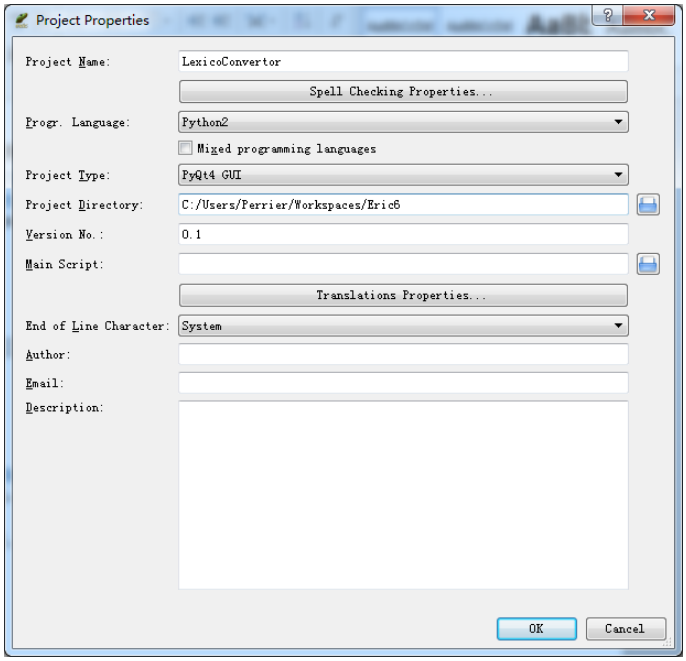
Eric6: Cette version d'Eric Python IDE fonctionne souvent avec PyQt4 et permet de compiler les codes en Python. http://eric-ide.python-projects.org/

PyInstaller : ce programme permet de générer les programmes exécutables de nos codes écrits en Python. https://github.com/pyinstaller/pyinstaller

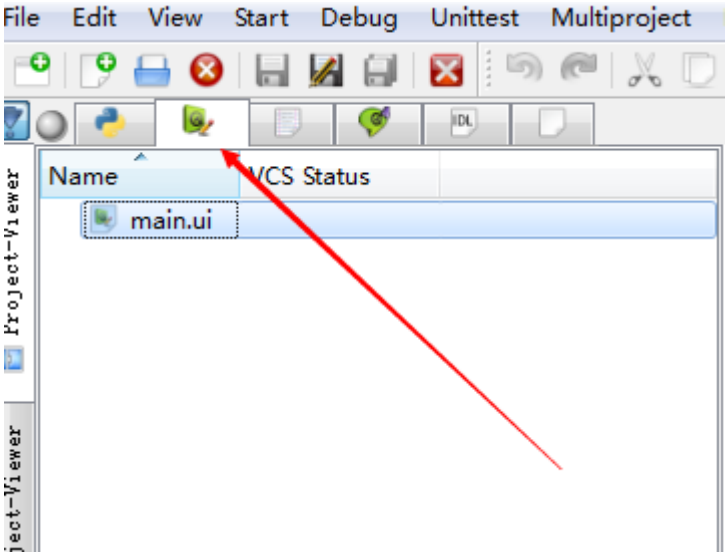
Comme l'illustre la figure ci-dessous, une fois l'environnement prêt, nous créons un projet *LexicoConvertor*.



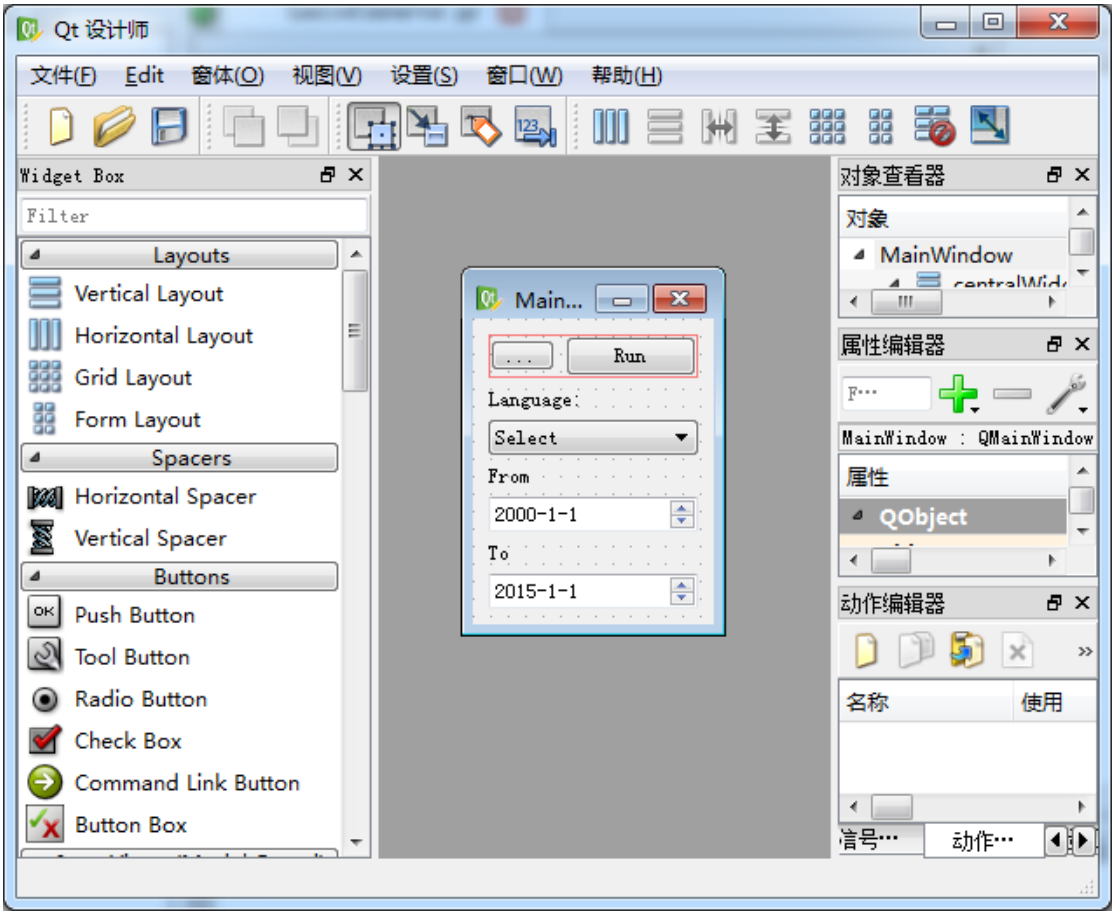
Nous paramétrons le projet en saisissant les différents champs concernés :



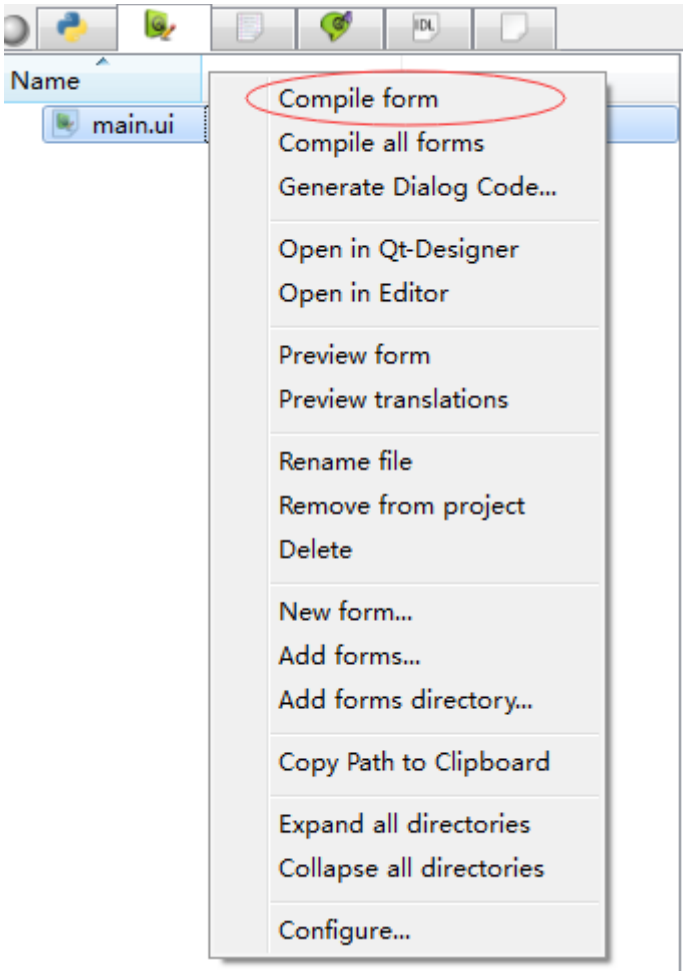
Nous créons une interface graphique sous le nom *main.ui* :



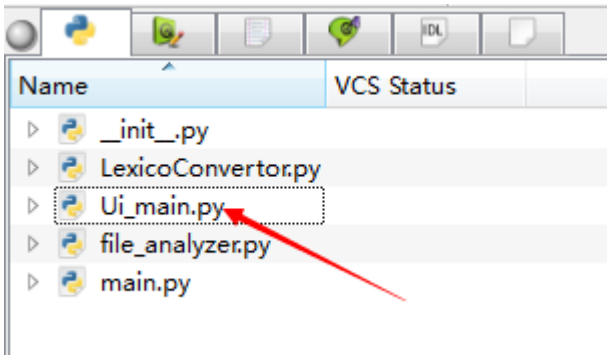
Ensuite, double clic sur *main.ui* pour ouvrir *Qt Designer*, nous commençons alors le paramétrage de cette interface en créant les champs suivants : un bouton fichier-entrée, un bouton *Run* pour l'exécution du programme, une liste déroulante pour la langue et deux champs de saisie pour les dates (début et fin) :



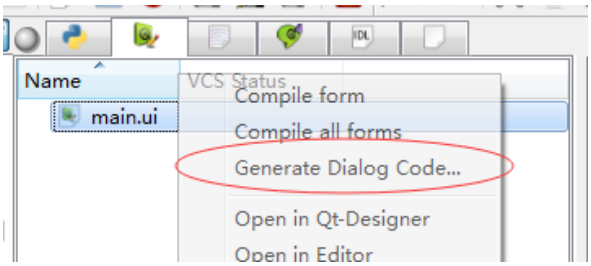
Une fois l'interface créée, nous revenons à l'environnement *Eric*, dans la barre *UI*, clic droit sur *main.ui*, sélectionner *Compile form* afin de générer l'interface graphique pour *Python UI_main.py*, comme illustrée ci-dessous :



En effet, comme illustrée ci-dessous, l'interface graphique est représentée dans le projet comme un fichier de programme sous le nom *Ui_main.py*.



Comme l'illustre la figure ci-dessous, pour configurer les différentes fonctions de cette interface, il faut procéder à un clic-droit sur *main.ui*, puis sélectionner *Generate Dialog Code*, par la suite, le fichier de code *main.py* sera généré.



Le contenu partiel du fichier *main.py* est présenté ci-dessous :

```
class MainWindow(QMainWindow, Ui_MainWindow):
    """
    Class documentation goes here.
    """
    def __init__(self, parent = None):
        """
        Constructor
        """
        QMainWindow.__init__(self, parent)
        self.setupUi(self)
        self._start = self.dateEdit_start.dateTime().toMSecsSinceEpoch()
        self._end = self.dateEdit_end.dateTime().toMSecsSinceEpoch()
        self._lang = 0
        self._sorted_type = 0

    @pyqtSignature("QDateTime")
    def on_dateEdit_end_dateTimeChanged(self, date):
        """
        update ended time
        """
        self._end = date.toMSecsSinceEpoch()

    @pyqtSignature("QDateTime")
    def on_dateEdit_start_dateTimeChanged(self, date):
        """
```

Les deux paramètres *@pyqtSignature("QDateTime")* et *on_dateEdit_end_dateTimeChanged(self, date)* contiennent les valeurs dates correspondant à la dernière modification du projet. Leurs procédures d'exécution se trouvent dans le fichier *file_analyzer.py*, nous y avons également mis les programmes de conversion de format XML au format Lexico3.

Après avoir réalisé ces précédents programmes, l'interface graphique sera opérationnelle et contrôlée par le fichier LexicoConvertor.py.

Enfin, le fichier exécutable sera généré par l'outil *Pyinstaller*, par la ligne de commande *pyinstaller –w LexicoConvertor.py*.

Annexe M

Liste des programmes de post-traitements :

align_checker.py	Vérifier la conformité des fichiers XML avant l'exécution des autres programmes
seg_xml.py	Segmenter les textes chinois dans les fichiers XML
remove_double.py	Supprimer les articles en double du corpus bilingue issus du site *chinadialogue*
conversion_signeschiAndinEN.py	Convertir tous les signes de ponctuation du chinois au format occidental (programme s'exécute avec le fichier signesChi.py).
align-raw.py	Uniformiser la longueur des paragraphes entre les deux langues au format Lexico3 (txt).
segmentor.py	Segmenter paragraphe par paragraphe les textes chinois des fichiers au format Lexico3.
recovery.py	Reconvertir un fichier du format Lexico3 en XML
stat.py	Recenser les nombres d'articles par jour, par mois et par année.
readline.exe	Vérifier dans le corpus parallèle en deux langues si les deux volets sont alignés (programme s'exécute avec le fichier align.py).
LexicoConvertor.exe	Convertir les fichiers XML au format Lexico3 (ce programme réalise la fonctionnalité inverse du programme recovery.py).

Annexe N : graphies et tableaux des résultats

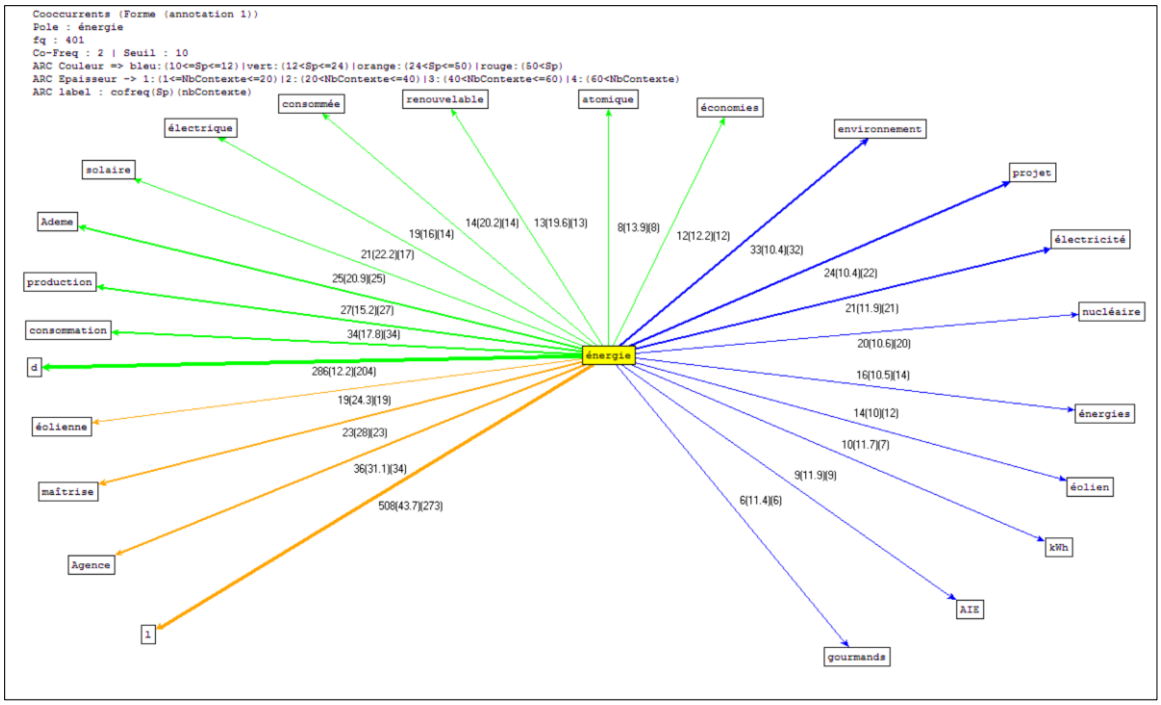


Figure N.1

ENRG_FR en 2010 : réseau cooccurrentiel autour de la forme-pôle *énergie*

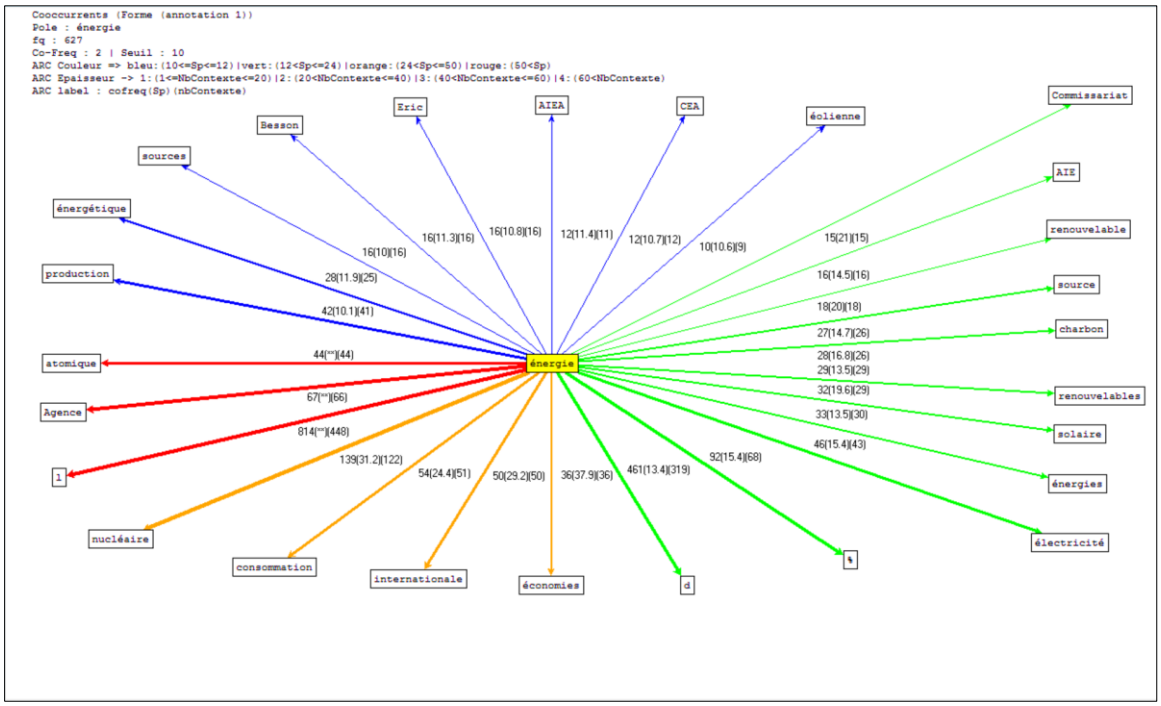


Figure N.2

ENRG_FR en 2011 : réseau cooccurrentiel autour de la forme-pôle *énergie*

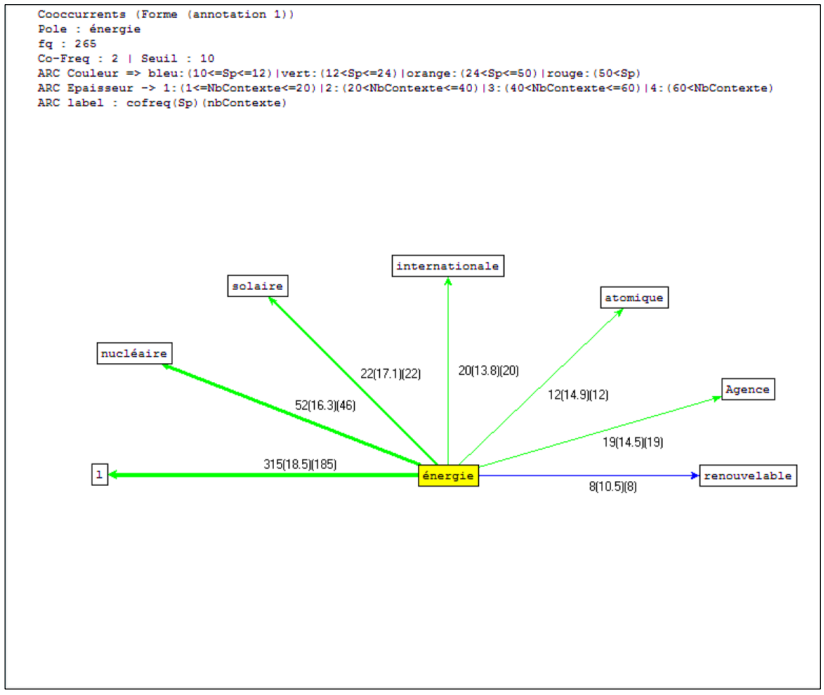


Figure N.3

ENRG_FR en 2012 : réseau cooccurrentiel autour de la forme-pôle *énergie*

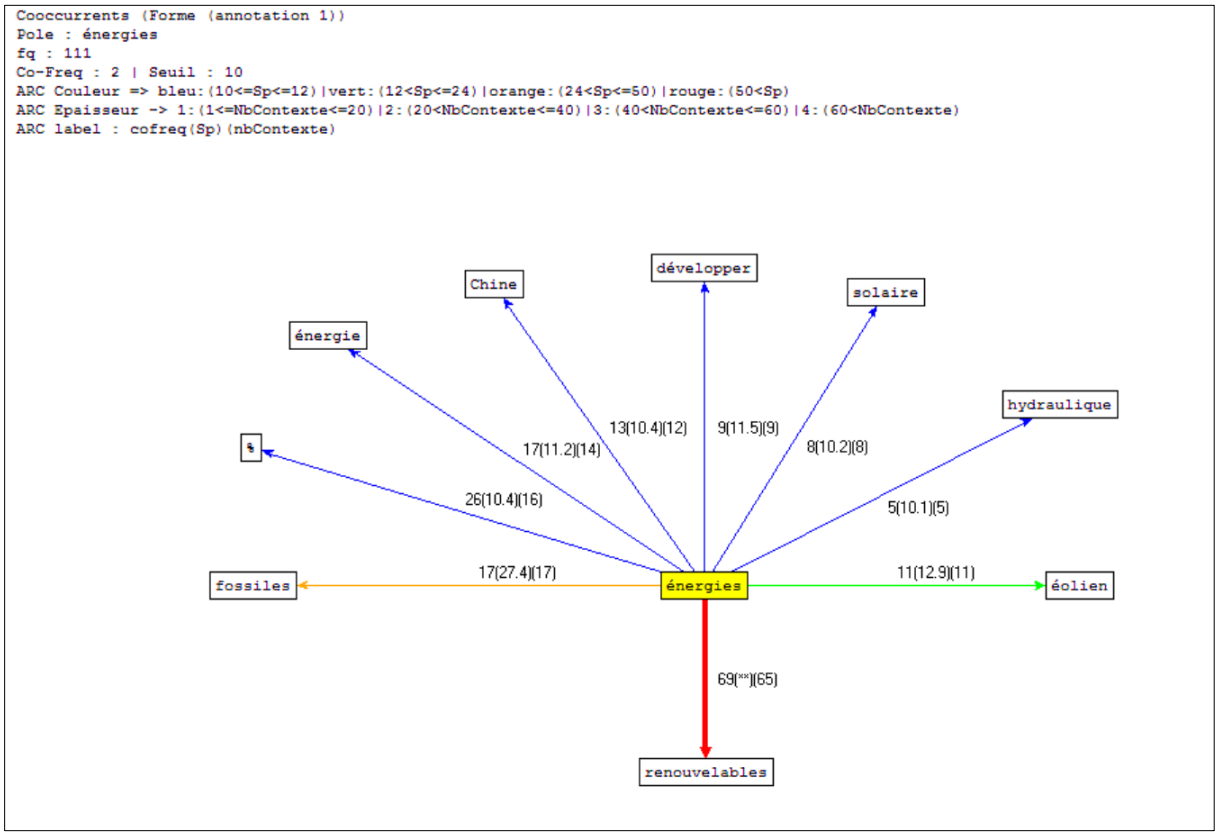


Figure N.4

ENRG_FR en 2010 : réseau cooccurrentiel autour de la forme-pôle *énergies*

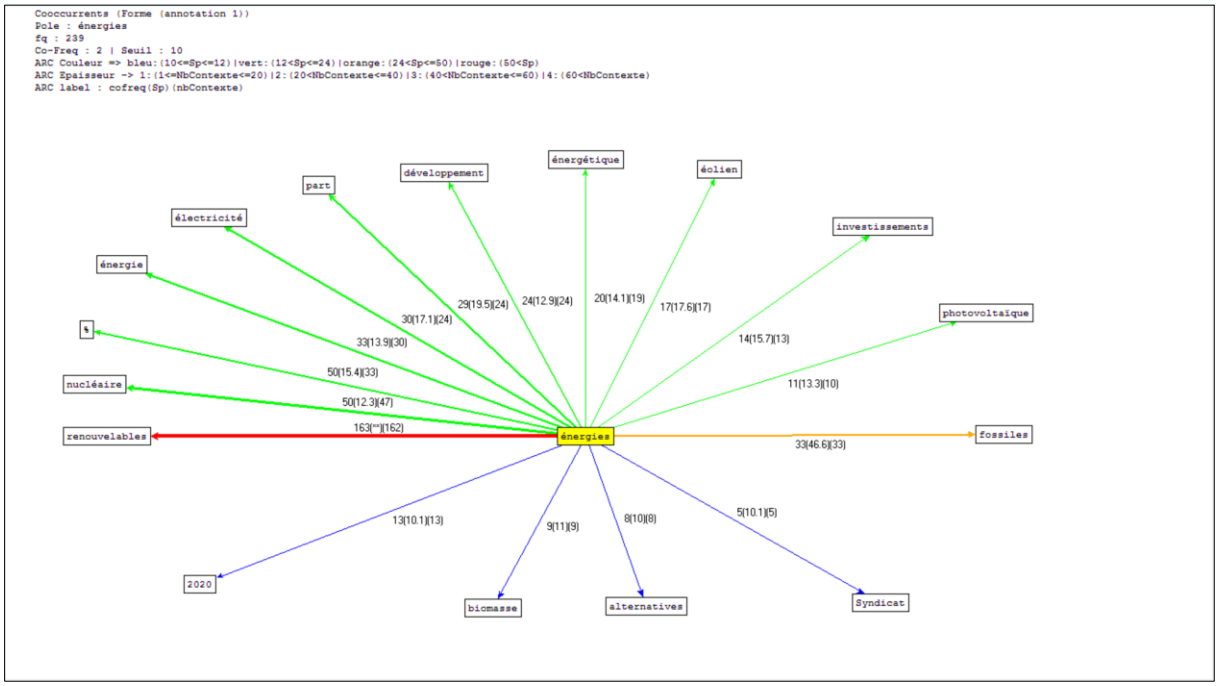


Figure N.5

ENRG_FR en 2011 : réseau cooccurrentiel autour de la forme-pôle *énergies*

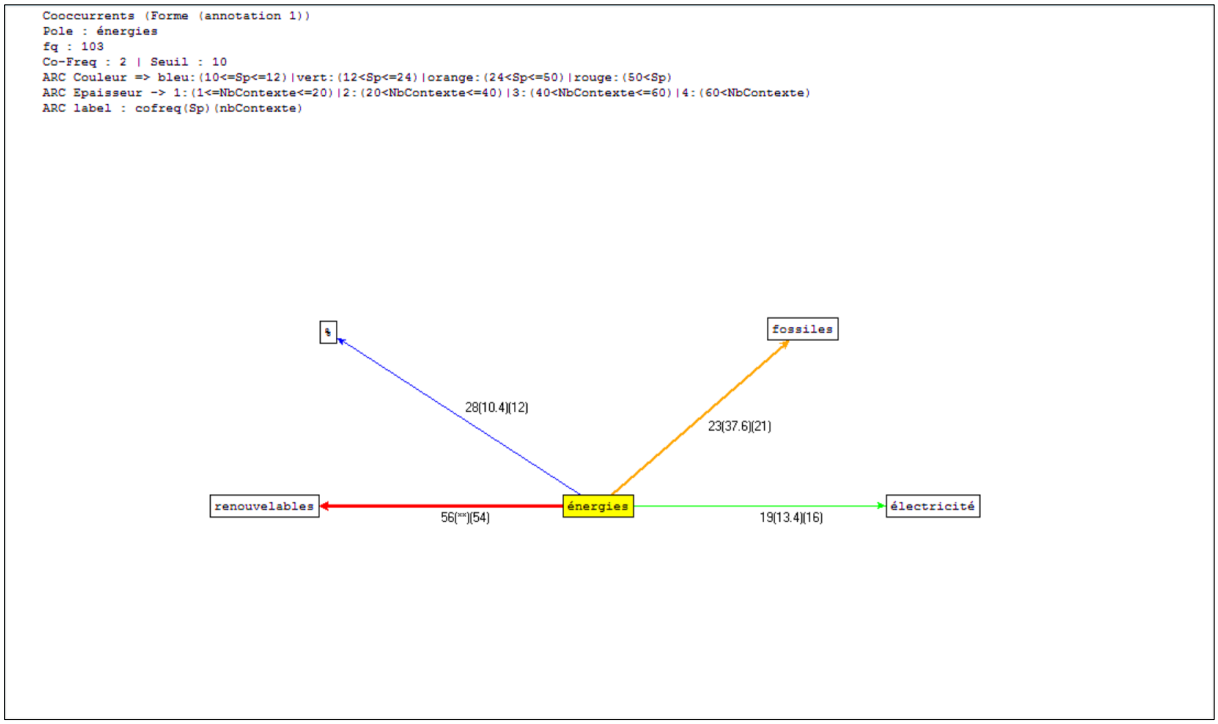


Figure N.6

ENRG_FR en 2012 : réseau cooccurrentiel autour de la forme-pôle *énergies*

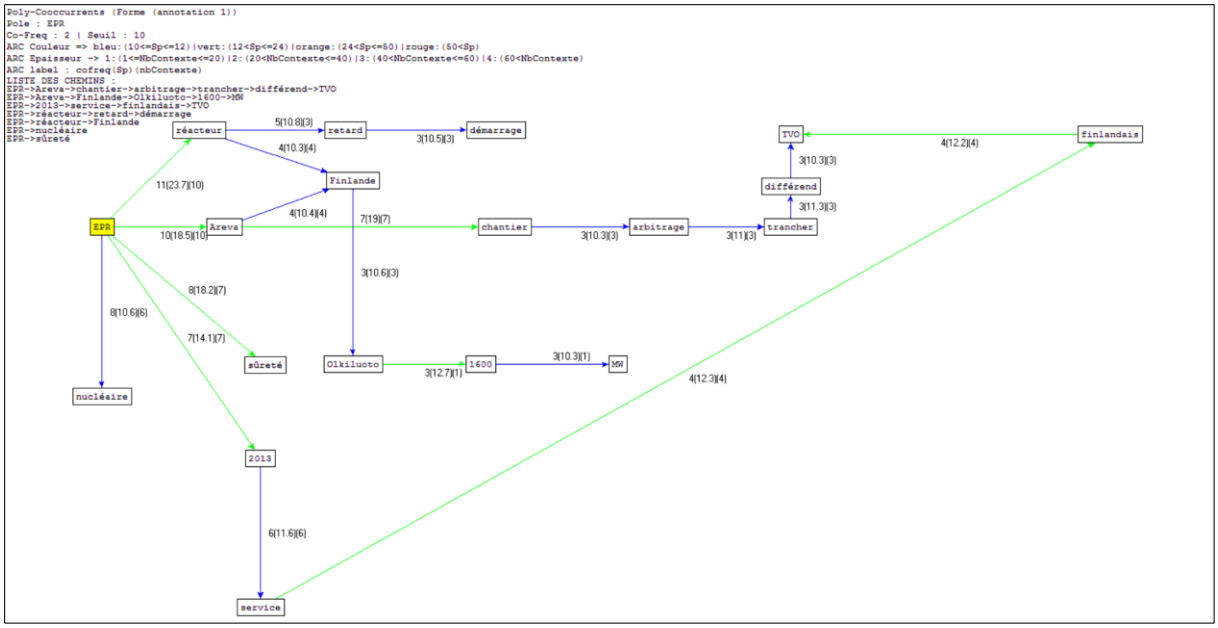


Figure N.7

ENRG_FR en 2010 : réseau poly-cooccurrentiel autour de la forme-pôle *EPR*

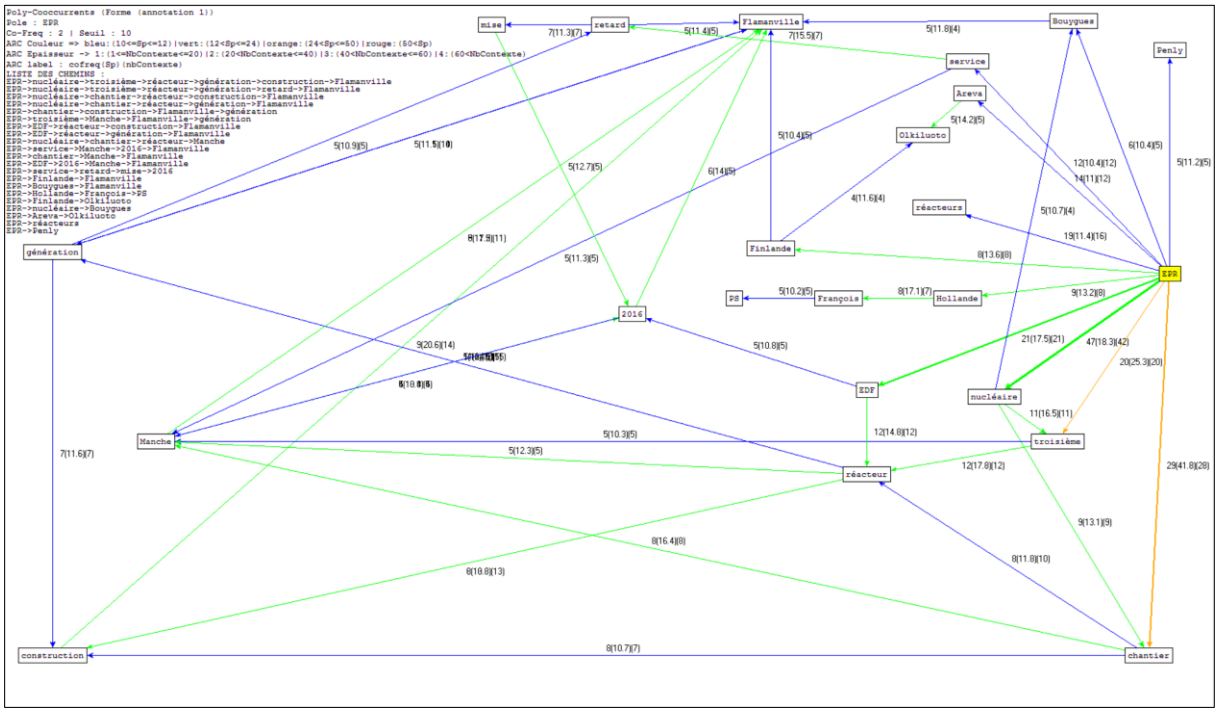


Figure N.8

ENRG_FR en 2011 : réseau poly-cooccurrentiel autour de la forme-pôle *EPR*

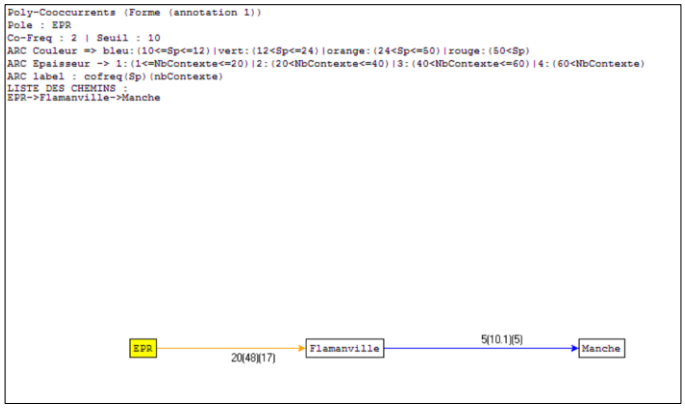


Figure N.9

ENRG_FR en 2012 : réseau cooccurrentiel autour de la forme-pôle *EPR*

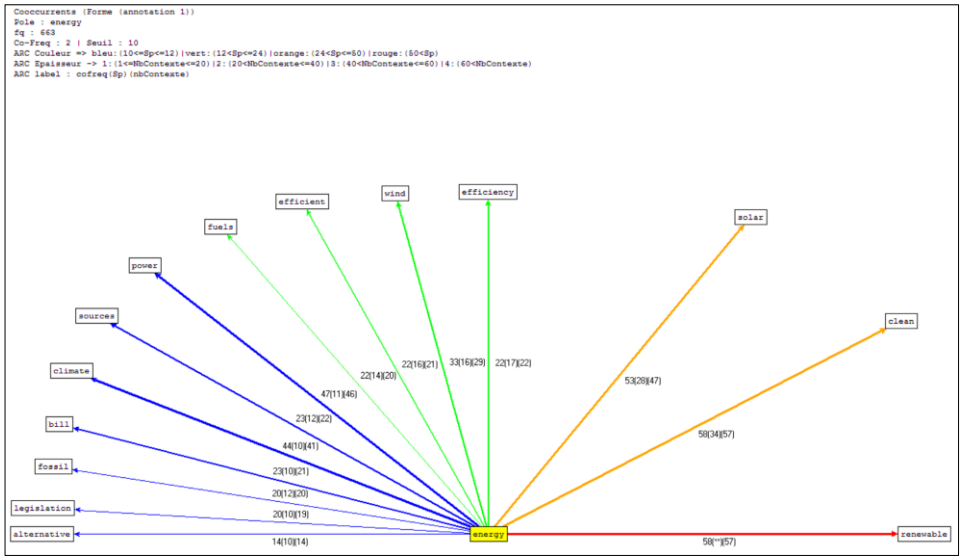


Figure N.10

ENRG_US en 2010 : réseau cooccurrentiel autour de la forme-pôle *énergy*

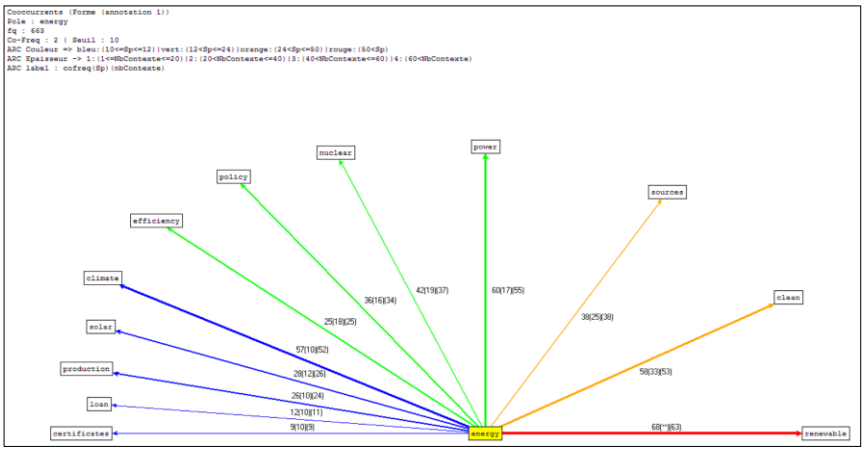


Figure N.11

ENRG_US en 2011 : réseau cooccurrentiel autour de la forme-pôle *énergy*

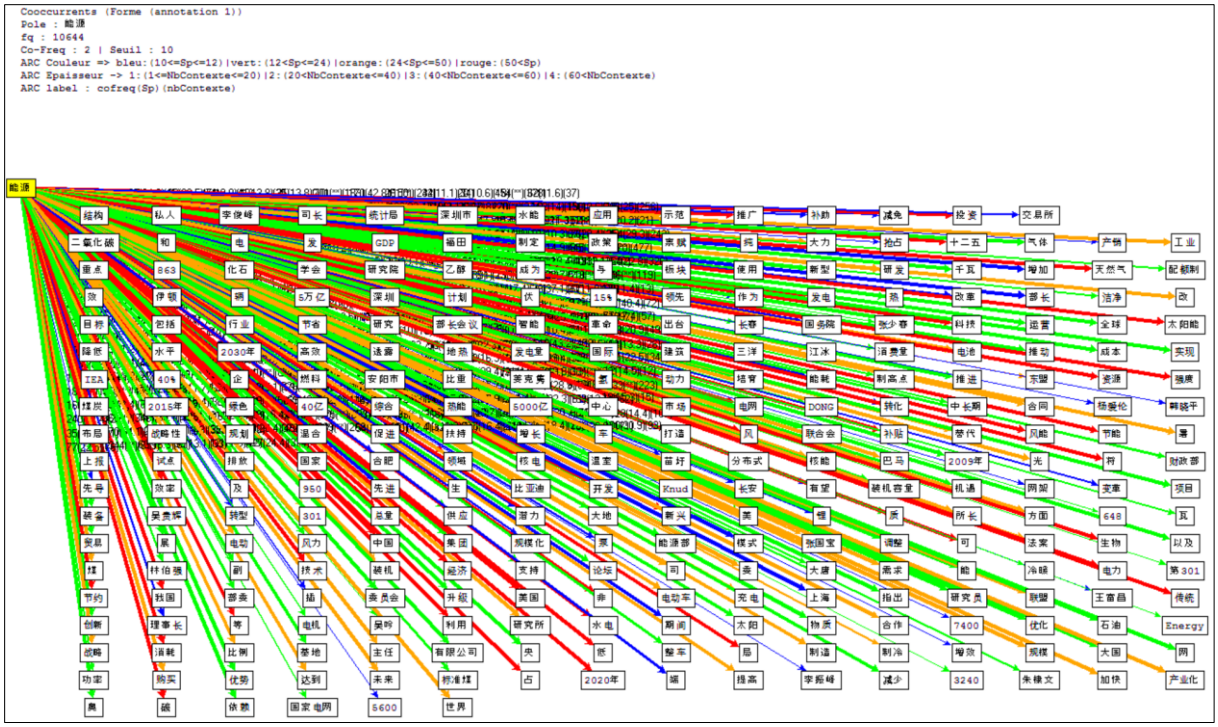


Figure N.12

ENRG_CN en 2010 : réseau cooccurrentiel autour de la forme-pôle 能源/néng yuán/énergies

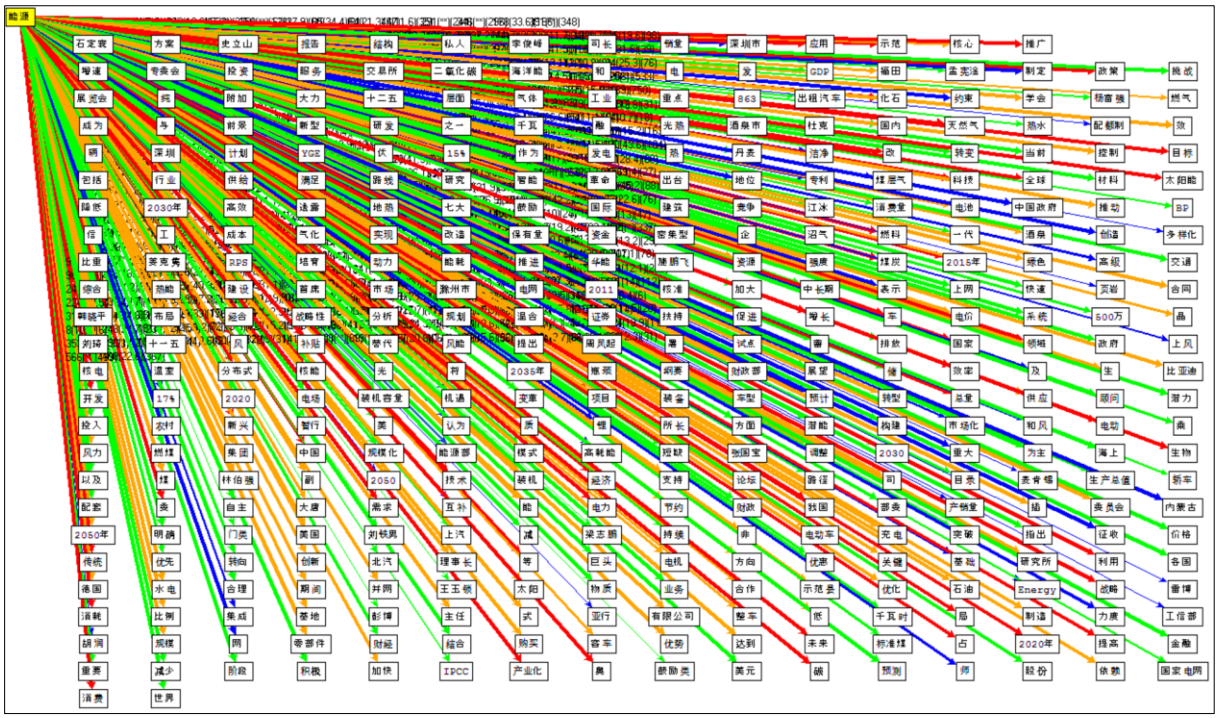


Figure N.13

ENRG_CN en 2011 : réseau cooccurrentiel autour de la forme-pôle 能源/néng yuán/énergies

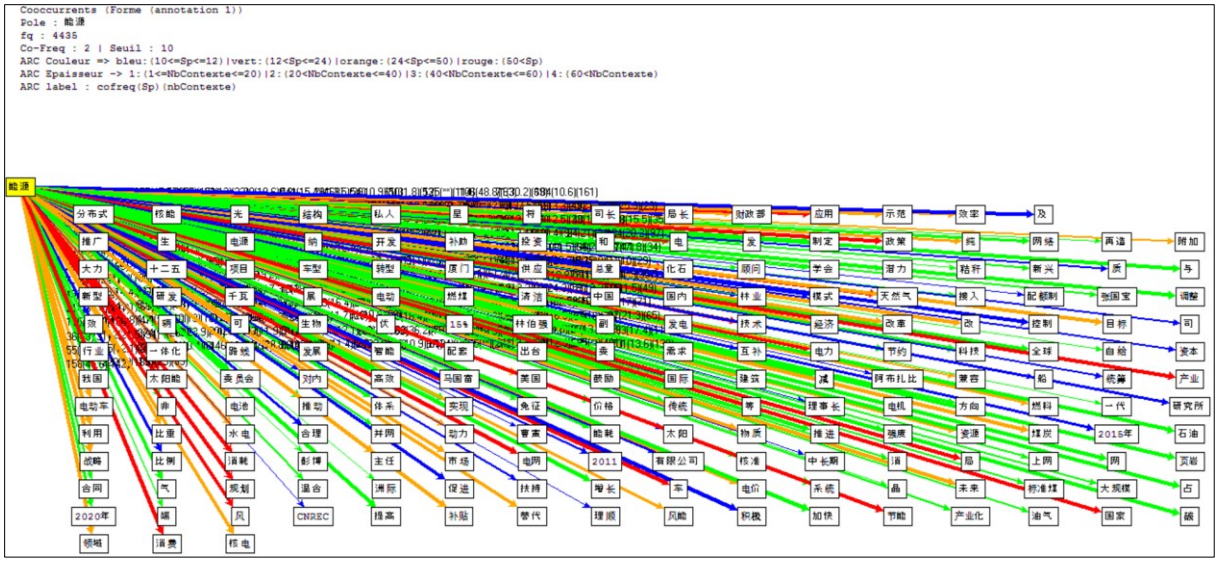


Figure N.14

ENRG_CN en 2012 : réseau cooccurrentiel autour de la forme-pôle 能源/néng yuán/énergies

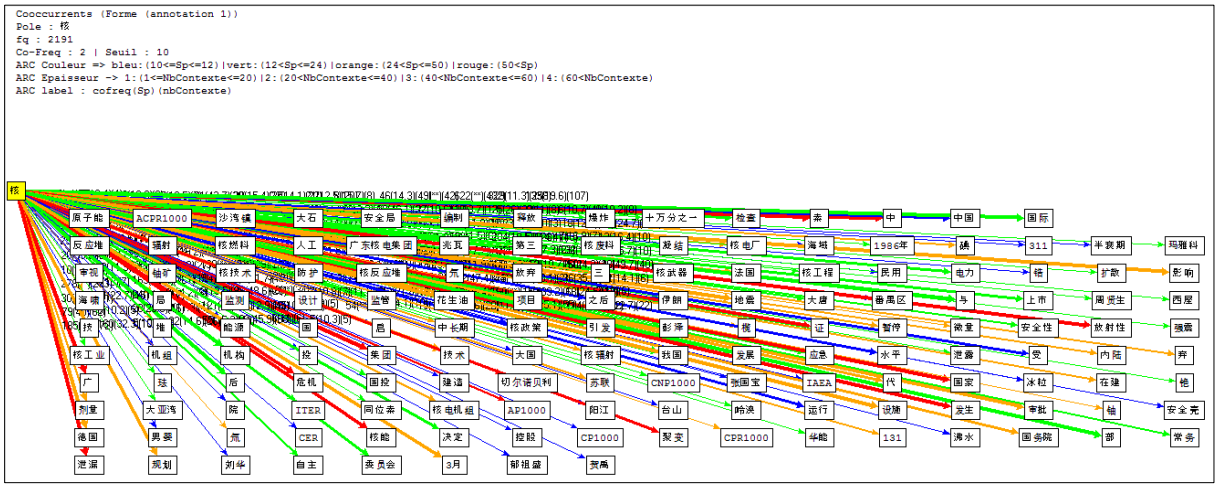


Figure N.15

ENRG_CN 2010, 2011 et 2012 : réseau cooccurrentiel autour de la forme-pôle 核/hé/nucléaire en un seul caractère chinois

Annexe N

Tableau N.1 :
CLRG 2006-2014 principales caractéristiques textométriques des deux volets en 99 mois

Volet chinois						Volet anglais					
aaaa/mm	occurrences	formes	hapax	Fréq. Max	Forme	Aaaa/mm	occurrences	formes	hapax	Fréq. Max	Forme
200606	10481	2742	1619	1002	的	200606	10288	2647	1577	587	the
200607	10769	2900	1706	928	的	200607	11376	3000	1746	641	the
200608	8119	2281	1325	618	的	200608	8357	2081	1166	452	the
200609	7055	2107	1240	574	的	200609	6987	2069	1236	409	the
200610	3682	1310	811	290	的	200610	4115	1402	899	202	the
200611	19144	3563	1878	1583	的	200611	20645	3729	1889	1061	the
200612	3912	1439	903	321	的	200612	4557	1496	917	219	the
200701	6512	2168	1367	532	的	200701	6709	2039	1268	340	the
200702	9443	2937	1799	729	的	200702	10162	2705	1588	606	the
200703	7258	2476	1517	605	的	200703	8218	2425	1462	464	the
200704	9530	2832	1712	800	的	200704	11115	2986	1797	700	the
200705	4328	1610	1017	354	的	200705	4736	1591	994	210	the
200706	13559	3720	2226	1107	的	200706	15450	3573	2028	862	the
200707	9192	3022	1890	754	的	200707	9551	2667	1541	484	the
200708	8892	2657	1565	734	的	200708	9554	2571	1497	571	the
200709	21026	4788	2767	1718	的	200709	22397	4721	2598	1246	the
200710	19818	4273	2375	1590	的	200710	22161	4285	2270	1223	the
200711	21708	4682	2633	1769	的	200711	23326	4462	2315	1277	the
200712	18679	3749	2017	1536	的	200712	20375	3754	1928	1240	the
200801	14723	3664	2066	1135	的	200801	15673	3634	2031	894	the
200802	11752	3180	1811	837	的	200802	12440	3157	1792	619	the
200803	11505	3155	1867	982	的	200803	12497	3159	1769	676	the
200804	8952	2699	1628	735	的	200804	10282	2722	1609	624	the
200805	11951	3238	1912	949	的	200805	12728	3116	1733	704	the
200806	12334	3424	2007	995	的	200806	13257	3386	1958	770	the
200807	11650	3018	1700	918	的	200807	12207	2876	1591	712	the
200808	10240	2943	1716	829	的	200808	11260	2763	1579	641	the
200809	9425	2730	1582	720	的	200809	10226	2691	1536	557	the
200810	12743	3553	2115	947	的	200810	13698	3372	1949	826	the
200811	7874	2678	1662	649	的	200811	8516	2659	1693	428	the
200812	14326	3473	1906	1145	的	200812	16097	3579	1916	816	the
200901	13126	3632	2122	994	的	200901	14226	3543	2022	787	the
200902	18735	4531	2537	1515	的	200902	20869	4385	2319	1266	the
200903	23468	5088	2863	2018	的	200903	25058	4661	2420	1411	the
200904	21362	4594	2542	1818	的	200904	21948	4577	2444	1131	the
200905	15662	3729	2078	1325	的	200905	16722	3650	1976	897	the
200906	11293	2767	1500	883	的	200906	12100	2697	1449	591	the
200907	15081	3695	2090	1275	的	200907	15718	3662	2058	828	the
200908	16593	3985	2274	1325	的	200908	16905	3639	1929	904	the
200909	17792	4282	2434	1506	的	200909	18497	3896	2117	1054	the
200910	12200	3052	1704	927	的	200910	12453	2892	1578	562	the

200911	21137	4978	2865	1754	的	200911	22731	4908	2749	1291	the
200912	13367	3418	1989	1015	的	200912	14102	3282	1839	767	the
201001	10022	3047	1805	836	的	201001	10725	2895	1665	607	the
201002	17328	4060	2243	1534	的	201002	18273	3972	2095	1018	the
201003	17108	4143	2384	1438	的	201003	18219	4122	2275	937	the
201004	10790	3019	1779	849	的	201004	12638	3158	1769	745	the
201005	11394	3279	1907	856	的	201005	13287	3105	1686	795	the
201006	14818	3754	2166	1194	的	201006	15625	3708	2076	844	the
201007	14807	3571	2031	1303	的	201007	16289	3494	1874	911	the
201008	9010	2839	1717	652	的	201008	9997	2682	1557	459	the
201009	15418	4073	2383	1129	的	201009	17147	3924	2133	908	the
201010	13817	3350	1818	1068	的	201010	15802	3294	1766	784	the
201011	17939	3774	1986	1413	的	201011	20177	3867	1994	1065	the
201012	17130	3833	2113	1380	的	201012	18309	3617	1907	1086	the
201101	9746	3155	1949	759	的	201101	10682	3108	1887	659	the
201102	9301	2693	1594	776	的	201102	9997	2657	1552	573	the
201103	17862	4128	2332	1376	的	201103	19700	3898	2069	1145	the
201104	12778	3822	2377	1089	的	201104	13703	3599	2146	795	the
201105	14766	3800	2120	1129	的	201105	16121	3621	1976	891	the
201106	12384	3475	2034	1009	的	201106	13593	3329	1895	753	the
201107	11626	3120	1823	980	的	201107	11499	3211	1864	541	the
201108	10441	3226	1960	838	的	201108	11013	3074	1859	716	the
201109	7798	2629	1652	623	的	201109	8210	2446	1497	459	the
201110	14734	3976	2296	1115	的	201110	16105	3679	2002	842	the
201111	11417	3100	1800	866	的	201111	12040	2889	1622	703	the
201112	10760	2907	1716	869	的	201112	11380	2860	1627	607	the
201201	7387	2624	1663	545	的	201201	8234	2532	1614	410	the
201202	1183	594	424	95	的	201202	1240	559	394	50	the
201203	9212	2803	1644	667	的	201203	9874	2665	1573	655	the
201204	15106	4244	2539	1156	的	201204	16101	3921	2254	783	the
201205	12131	3482	2090	935	的	201205	12817	3345	1985	622	the
201206	14466	3427	1903	1167	的	201206	15329	3354	1762	811	the
201207	10956	3266	1940	840	的	201207	12431	3322	1983	577	the
201208	9680	3038	1764	623	的	201208	11861	2882	1647	656	the
201209	9434	2938	1749	645	的	201209	11208	2810	1578	588	the
201210	6676	2223	1336	516	的	201210	7312	2195	1327	404	the
201211	14392	3721	2158	1159	的	201211	15893	3639	2021	832	the
201212	5834	2074	1241	477	的	201212	6178	2038	1282	376	the
201301	3520	1422	894	228	的	201301	3900	1300	775	183	the
201302	6092	2043	1233	472	的	201302	6561	1942	1162	277	the
201303	4508	1610	970	328	的	201303	4465	1536	977	212	the
201304	7823	2405	1442	604	的	201304	8373	2392	1432	428	the
201305	6424	2005	1175	489	的	201305	7200	2093	1243	322	the
201306	2963	1260	820	211	的	201306	3300	1279	884	191	the
201307	11574	3183	1782	788	的	201307	11595	2871	1631	601	the
201308	3979	1532	975	316	的	201308	4039	1455	975	228	the

201309	2418	1132	768	165	的	201309	2602	1063	712	136	the
201310	1228	606	415	57	的	201310	1353	614	431	64	the
201311	9191	2738	1593	665	的	201311	9866	2465	1364	479	the
201312	3767	1445	909	196	的	201312	4291	1365	843	210	the
201401	6688	2206	1353	494	的	201401	7299	2219	1400	367	the
201402	733	412	300	42	的	201402	834	410	292	28	the
201403	5821	2088	1285	428	的	201403	6273	2012	1235	322	the
201404	2279	1054	694	174	的	201404	2443	1052	733	95	the
201405	729	368	260	36	的	201405	1009	466	336	48	the
201406	9781	2567	1417	743	的	201406	10534	2709	1541	580	the
201407	2923	1179	766	212	的	201407	3356	1218	780	169	the
201408	5296	1883	1158	378	的	201408	5576	1770	1093	244	the

Annexe N

Accroissement de vocabulaire bilingue sur le corpus parallèle CLRG

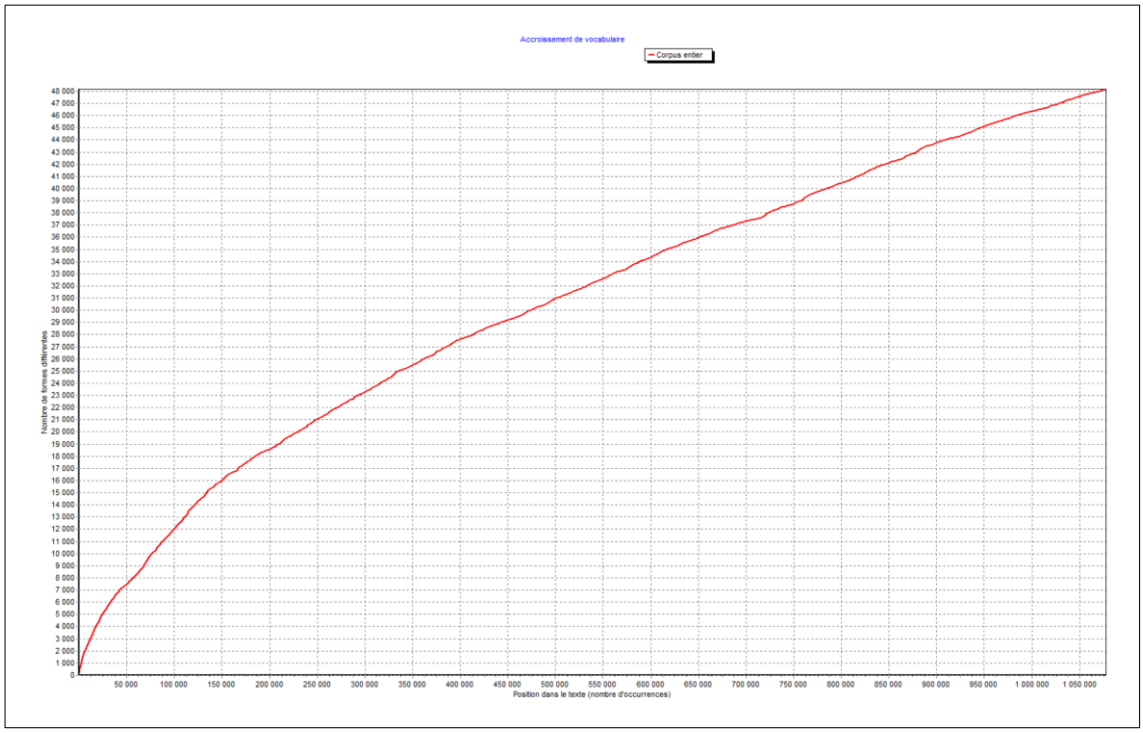


Figure N.16
CLRG_CN : accroissement de vocabulaire
de 2006 à avril 2014

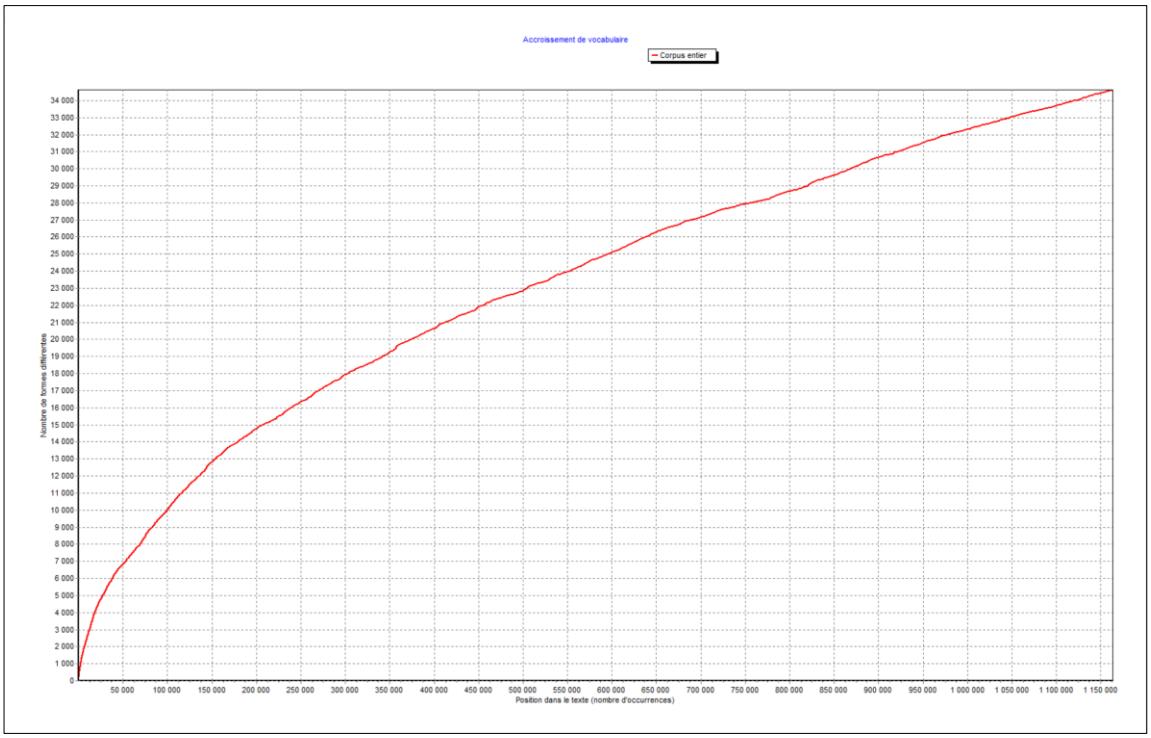


Figure N.17
CLRG_EN : accroissement de vocabulaire
de 2006 à avril 2014

Annexe N

Tableau N.2 :

CLRG de 2006 à 2014 : réseaux cooccurrentiels parallèles des formes 能源/néng yuán/énergies et *energy*

Cooccurrents (source) : 能源 (fq:3015) Co-Freq : 2 | Seuil : 10 Cooccurrents (cible) : *energy* (fq:4475) | Co-Freq : 2 | Seuil : 10

Cooccurrents : (source) 能源

- *Co-Freq : 2*
- *Seuil : 10*

Forme	Équivalent français	Fq	co-Fq	specif	context
消耗	consommer	208	93	25.7	77
合作社	coopérative	38	24	12.0	17
太阳能	énergie solaire	832	269	38.5	169
封存	captage et stockage	94	39	10.7	27
2030	2030	141	65	19.4	48
煤炭	charbon / houille	630	284	75.3	157
比例	proportion	180	71	16.5	59
美	États-Unis	170	75	20.7	50
能耗	consommation d'énergie	246	87	16.4	57
能源供应	approvisionnement en énergie	58	36	16.8	33
政策	politique	1519	390	30.1	300
能效	consommation d'énergie	313	143	39.8	103
加强	renforcer	251	75	10.4	66
分布式	de façon distribuée	54	41	24.1	22
效率	efficacité	265	140	48.3	111
为主	être majoritairement à	74	34	11.0	26
交通	trafic ou circulation	134	49	10.6	45
排放	émission	2904	606	20.6	357
天然气	gaz naturel	475	172	31.8	119
部落	tribu	92	39	11.0	16
强度	intensité ou force	197	104	36.4	72
中美	Chine et les États-Unis	144	59	14.9	44
火电	électricité issue du charbon	90	42	13.4	22
署	agence	49	31	15.1	30
结构	structure	206	96	28.2	81
需求	demande ou exigence	528	141	13.6	120
目标	objectif	1457	378	30.2	238
系统	système	600	153	12.8	103
密集型	intensive	41	27	14.0	23
燃煤	charbon (combustible)	248	101	23.9	73
单位	unité	182	60	10.5	47
进口	importer	224	74	12.5	62
一次能源	énergie primaire	31	22	12.7	20
欧盟	UE	693	171	12.8	90
成本	coût	902	212	13.2	146
风能	énergie éolienne	247	96	21.1	84

Cooccurrents : (cible) energy

- *Co-Freq : 2*
- *Seuil : 10*

Forme	Équivalent français	Fq	co-Fq	specif	context
potential	potentiel	480	161	14.0	149
both	les deux	864	251	12.6	220
bulbs	ampoules	60	41	16.7	31
appliances	appareils	59	45	21.3	39
dependence	dépendance	55	33	11.5	30
15%	15%	78	39	10.0	37
green	vert	650	210	15.7	170
lighting	éclairage	51	38	17.7	33
target	cible	332	129	17.0	109
intensive	intensif	143	88	28.7	79
emissions	émissions	2793	712	16.5	490
power	puissance	2309	936	122.6	599
China	Chine	6988	2005	80.9	1089
coal	charbon	1443	591	80.2	344
bulb	ampoule	25	21	12.3	14
its	son	3008	752	15.4	544
transition	transition	122	55	11.1	55
sector	secteur	495	184	20.9	159
reduction	réduction	520	172	14.2	149
supply	alimentation	445	192	30.9	170
Plan	Plan	224	106	21.6	77
20%	20%	237	120	27.5	100
sources	sources	369	227	71.0	204
Five	cinq	194	100	24.0	70
standards	normes	378	123	10.2	112
security	sécurité	642	269	39.8	214
panels	panneaux	101	48	10.9	39
Chu	Chu	21	18	11.0	15
technology	technologie	1010	354	32.4	284
and	et	32318	7176	40.9	2378
sectors	secteurs	160	68	11.9	61
source	source	270	98	11.4	94
economic	économique	1325	358	12.5	293
transport	transport	188	80	13.7	74
efficiency	efficacité	481	374	168.9	311
gas	gaz	1326	409	24.3	297
policies	politiques	428	137	10.7	119
11th	11ème	64	39	13.5	36
wind	vent	918	412	69.8	291
production	production	668	217	16.5	187
demand	demande	498	225	39.6	174
buildings	bâtiments	362	245	89.4	147

Annexe N

再生能源	énergies renouvelables	823	385	107.5	276
能源需求	la demande d'énergie	104	71	35.5	59
化石	fossile	400	199	62.1	164
加州	Californie	209	67	10.9	40
增长	augmenter	942	264	26.9	177
比重	proportion	59	30	11.2	25
提高	augmenter	681	205	25.4	171
依赖	dépendance ou dépendre	282	100	18.6	89
消费结构	structure de consommation	11	11	10.2	8
捕集	captage	88	38	11.0	17
电力	puissance électrique	578	175	22.3	136
公司	société	1428	316	15.0	183
发展	développement	3269	795	49.3	542
市场	marché	1108	241	11.2	166
低碳	densité carbonique faible	501	127	10.8	98
建筑物	bâtiments	80	38	12.6	26
两国	les deux pays	186	65	12.5	50
燃料	carburant	959	324	50.5	223
能源消耗	consommation d'énergie	112	51	15.3	43
气体	gaz	1098	271	19.1	178
2020	2020	392	129	20.3	103
能源安全	sécurité énergétique	144	69	21.6	61
减少	réduction	1273	273	11.7	210
等	etc.	1888	414	18.3	331
使用	usage	717	175	12.7	140
石油	huile	1174	262	13.2	171
实现	atteindre	1012	278	26.7	215
方面	aspect	1082	232	10.1	193
占	considération	575	151	13.8	123
CCS	CCS	191	83	22.2	28
亿吨	cent millions de tonnes	206	69	12.1	43
发电	générer de l'électricité	884	293	44.0	179
资源	ressources	857	208	14.4	159
人均	par habitant	275	86	12.8	47
领域	zone	566	144	12.1	114
绿色	couleur verte	891	202	11.2	167
降低	inférieur	602	151	12.1	132
中国	chinois	8453	1791	61.0	808
温室	effet de serre	1044	258	18.4	167
低	bas	667	156	10.0	134
煤	houille	196	79	18.8	65
节能	économie d'énergie	517	131	11.1	96
IEA	AIE	72	35	12.1	23

fossil	fossile	389	216	56.7	189
reduce	réduire	698	233	19.1	210
targets	cibles	571	198	18.5	153
costs	frais	678	213	14.5	172
FYP	plan quinquennal	43	29	12.1	17
geothermal	géothermique	33	24	11.4	23
renewable	renouvelable	601	525	285.7	420
generation	génération	380	154	21.9	129
unit	unité	112	71	24.7	64
intensity	intensité	222	132	39.8	97
fuel	carburant	633	221	20.8	182
investments	investissements	183	71	10.1	56
technologies	technologies	556	256	46.6	211
conditioning	conditionnement	81	55	21.6	32
fuels	carburants	303	160	38.9	147
generated	générée	152	69	13.6	61
economy	économie	791	268	22.6	230
Year	An	197	99	22.7	69
2020	2,020	387	171	29.2	137
systems	systèmes	426	149	14.7	110
nuclear	nucléaire	1091	402	42.1	221
consumption	consommation	606	392	131.7	275
policy	politique	943	304	21.7	258
new	nouveau	2064	528	13.1	442
EU	UE	493	169	15.6	96
carbon	carbone	3163	878	31.4	572
cost	coût	648	209	15.6	159
centralised	centralisée	23	19	11.0	13
cleaner	nettoyeur	55	34	12.3	33
reducing	réduire	317	109	10.8	106
electricity	électricité	895	374	54.1	278
biomass	biomasse	103	68	25.2	49
per	par	799	238	13.3	165
clean	nettoyer	494	304	94.4	255
renewables	énergies renouvelables	192	122	41.2	106
reliance	dépendance	65	36	11.0	36
building	bâtiment	652	229	21.8	155
development	développement	1583	470	23.8	370
deployment	déploiement	78	40	10.6	36
US	États-Unis	2500	608	10.8	412
measures	mesures	316	112	12.0	97
Germany	Allemagne	275	127	24.3	95
solar	solaire	837	446	106.5	282
needs	besoins	542	170	12.0	154
primary	primaire	87	50	15.4	47
efficient	efficace	272	177	61.2	159
use	utilisation	1193	467	57.0	382
growth	croissance	868	273	18.1	211
subsidies	subventions	163	65	10.0	49
12th	12ème (quinquennat)	92	52	15.5	42

Annexe N

电厂	centrale/*Power Plant*	309	89	11.1	64
转型	transition	191	71	15.0	58
利用	usage	648	196	24.7	161
德国	Allemagne	417	153	29.3	100
美国	United States	2693	509	10.5	340
局	bureau (du ministère)	53	39	22.2	36
电网	puissance	221	69	10.6	53
技术	technologie	1987	477	29.2	322
可	pouvoir	1990	616	77.1	435
生物质能	biomasse (énergie)	37	24	12.4	16
CO2	CO2	92	41	12.3	28
消费	consommation	389	196	62.4	139
合作	coopération	897	254	26.7	157
清洁	propre	734	383	126.2	291
替代	substituer ou suppléant	200	102	34.0	91
总量	quantité globale	328	134	31.2	87
革命	révolution	71	32	10.1	30
节约	économiser	89	41	12.9	31
利用效率	efficacité d'utilisation	26	23	17.0	21
投资	investissement	1095	284	23.2	192
产业	Industrie	557	153	15.7	114
建筑	construction (bâtiment)	540	136	11.3	78
碳	charbon	2388	512	20.1	317
生产	fabrication ou production	789	183	11.2	149
核能	énergie nucléaire	371	152	35.3	95
GDP	PIB	293	93	14.1	61
经济	économie	2388	568	33.2	399
Renewable	renouvelables	57	36	13.3	36
low	faible	1187	391	29.3	313
grid	réseau ou grille	271	151	40.5	109
saving	économie	271	204	88.4	153
increase	augmentation	629	186	10.5	165
investment	investissement	653	227	21.0	195
industry	industrie	815	260	18.2	221
mix	mélanger ou mixte	54	32	11.0	28
oil	gaz	1363	361	11.5	238
heating	chauffage	149	71	15.3	46
GDP	PIB	279	110	15.3	83
industries	industries	210	102	21.9	87
increasing	croissant	345	118	11.4	109
savings	économie ou épargnes	105	75	31.1	58
Energy	énergie	587	246	36.6	207
30%	30%	134	58	10.8	53
goals	buts	164	71	12.8	64

Table des matières

Table des matières

Table des matières

Table des matières

METHODES DE VEILLE TEXTOMETRIQUE MULTILINGUE APPLIQUEES A DES CORPUS DE L'ENVIRONNEMENT ET DE L'ENERGIE
« Restitution, prévision et anticipation d'événements par poly-résonances croisées »

Résumé

Cette thèse présente une série de méthodes de veille textométrique multilingue appliquées à des corpus thématiques. Pour constituer ce travail, deux types de corpus sont mobilisés : un corpus comparable et un corpus parallèle, composés de données textuelles extraites des discours de presse, ainsi que ceux des ONG. Les informations récupérées proviennent de trois mondes en trois langues différentes : français, anglais et chinois. La construction de ces deux corpus s'effectue autour de deux thèmes d'actualité ayant pour objet, l'environnement et l'énergie, avec une attention particulière sur trois notions : les énergies, le nucléaire et l'EPR. Après un bref rappel de l'état de l'art en intelligence économique, veille et textométrie, nous avons exposé les deux sujets retenus, les technicités morphosyntaxiques des trois langues dans les contextes nationaux et internationaux. Successivement, les caractéristiques globales, les convergences et les particularités de ces corpus ont été mises en évidence. Les dépouillements et les analyses qualitatives et quantitatives des résultats obtenus sont réalisés à l'aide des outils de la textométrie, notamment grâce aux analyses factorielles des correspondances, réseaux cooccurrentiels et poly-cooccurrentiels, spécificités du modèle hypergéométrique, segments répétés ou encore à la carte des sections. Ensuite, la veille bi-textuelle bilingue a été appliquée sur les trois mêmes concepts dans l'objectif de mettre en évidence les modes selon lesquels les corpus multilingues à caractère comparé et parallèle se complètent dans un processus de veille plurilingue, de restitution, de prévision et d'anticipation. Nous concluons notre recherche en proposant une méthode analytique par Objets-Traits-Entrées (OTE).

Mots-clés : textométrie, veille multilingue, opinions, corpus comparable, corpus parallèle, discours de presse, discours des ONG, fouille textuelle, cooccurrences, poly-cooccurrences, nucléaire, EPR, énergies, environnement

Textometric Multilingual Information Monitoring Methods Applied to Energy & Environment Corpora: "Restitution, Forecasting and Anticipation of Events by Cross Poly-resonance"

Abstract

This thesis presents a series of textometric multilingual information monitoring methods applied to thematic corpora (textometry is also called textual statistics or text data analysis). Two types of corpora are mobilized to create this work: a comparable corpus and a parallel corpus in which the textual data are extracted from the press and discourse of NGOs. The information source was retrieved from three countries in three different languages: English, French and Chinese. The two corpora were constructed on two topical issues concerning the environment and energy, with a focus on three concepts: energy, nuclear power and the EPR (European Pressurized Reactor or Evolutionary Power Reactor). After a brief review of the state of the art on business intelligence, information monitoring and textometry, we first set out the two chosen subjects – the environment and energy – and then the morphosyntactic features of the three languages in national and international contexts. The overall characteristics, similarities and peculiarities of these corpora are highlighted successively. The recounts and qualitative and quantitative analyses of the results were carried out using textometric tools, including factor analysis of correspondences, co-occurrences and polyco-occurrential networks, specificities of the hypergeometric model and repeated segments or map sections. Thereafter, bilingual bitextual information monitoring was applied to the same three concepts with the aim of elucidating how the comparable corpus and the parallel corpus can mutually help each other in a process of multilingual information monitoring, by restitution, forecasting and anticipation. We conclude our research by offering an analytical method called Objects-Features-Opening (OFO).

Keywords: textometry, multilingual information monitoring, opinions, comparable corpus, parallel corpus, media discourse, discourse of NGOs, text mining, co-occurrences, poly-cooccurrences, nuclear, EPR, energy, environment

Discipline : Sciences du langage

UNIVERSITE SORBONNE NOUVELLE - PARIS 3
Ecole Doctorale 268 Langage et Langues
4, rue des Irlandais 75005 Paris