Brouillard de pollution en Chine. Analyse sémantique différentielle de corpus institutionnels, médiatiques et de microblogues

Qinran Dang

INSTITUT NATIONAL DES LANGUES ET CIVILISATIONS ORIENTALES
École doctorale N°265 : *LANGUES, LITTÉRATURES ET SOCIÉTÉS DU MONDE*
Équipe de recherche Textes, Informatique, Multilinguisme

THÈSE

présentée par
DANG QINRAN

soutenue le 29 juin 2020
pour obtenir le grade de **Docteur de l'INALCO** en
Traitement automatique des Langues

Brouillard de pollution en Chine. Analyse sémantique différentielle de corpus institutionnels, médiatiques et de microblogues

Thèse dirigée par :

M. Mathieu Valette	Professeur des universités, INALCO
M. Nicolas Turenne	Maître de conférences, BNU - HKBU United International College

Rapporteurs :

M. Damon Mayaffre	Professeur, Université de Nice Sophia Antipolis
M. Hongmiao Wu	Professeur des universités, Université de Wuhan

---

Membres du jury :

M. Mathieu Valette	Professeur des universités, INALCO
M. Hongmiao Wu	Professeur des universités, Université de Wuhan
M. Damon Mayaffre	Professeur, Université de Nice Sophia Antipolis
M. Nicolas Turenne	Maître de conférences, BNU - HKBU United International College

Résumé

Au fur et à mesure de la dégradation de la qualité de l'air en Chine, de plus en plus d'articles journalistiques et de microblogues (*weibo* en chinois, équivalent de *tweet*), provenant de sites web gouvernementaux, médiatiques, de réseaux sociaux, de forums ou de blogs, traitent le problème du « 雾霾 » (*wumai* en chinois, pour désigner le brouillard de pollution) en Chine sous plusieurs angles : politique, écologique, économique, sociologique, sanitaire, etc. La sémantique des thèmes abordés dans ces textes diffère sensiblement en fonction de leur genre textuel. Dans cette thèse, nous avons pour objectif d'une part, de relever les différents thèmes d'un corpus numérique traitant du *wumai* et spécifiquement construit à cette fin, et d'autre part, d'interpréter de façon différentielle la sémantique de ces thèmes.

Dans un premier temps, nous collectons les données textuelles en langue chinoise relatives au *wumai*. Ces textes provenant de trois sites web chinois traditionnels et du réseau social sont divisés en quatre genres textuels. Après une série de traitements préparatoires : nettoyage, segmentation, normalisation, annotation, balisage et organisation, nous étudions les caractéristiques des quatre genres textuels du corpus à partir d'une série de variables discriminantes — hyperstructurelles, lexicales, sémiotiques, rhétoriques, modales et syntaxiques — réparties au niveau infratextuel et intratextuel. Ensuite, en nous basant sur les caractéristiques de chaque genre textuel, nous relevons les thèmes principaux exposés dans chaque genre de sous-corpus, et analysons de manière contrastive la sémantique de ces thèmes récupérés. Les résultats d'étude sont interprétés de manière quantitative et qualitative. Les analyses quantitatives s'effectuent à l'aide d'outils textométriques, les interprétations sémantiques s'inscrivent dans le cadre théorique de la sémantique interprétative (SI) proposée par Rastier (1987).

Mots-clés : Humanité numérique, fouille de texte, textométrie, analyses sémantique, genre textuel, analyses des réseaux sociaux, analyses du discours institutionnel, traitement automatique du chinois, corpus écologique

Abstract

Air pollution has increasingly become a serious problem in China, more and more journalistic articles and miniblogs (*weibo* in Chinese, equivalent to *tweet*), coming from governmental or media websites, social networks, blogs and forums, etc., discuss the issue of "雾霾" (*wumai* in Chinese, means smog) in China through several angles : political, ecological, economic, sociological, health, etc. The semantics of the themes addressed in these texts differ significantly from each other according to their textual genre. In the framework of our research, our objective is double-fold : on the one hand, to identify different themes of a digital propose-built corpus relating to *wumai* ; and on the other hand, to interpret differentially the semantics of these themes.

Firstly, we collect the textual data written in Chinese and related to *wumai*. These journalistic articles and *weibo* deriving from three traditional Chinese websites and the social network are divided into four genres of sub-corpus. Secondly, we constitute our corpus through a series of data processing : data cleaning, word segmentation, normalization, POS tagging, benchmarking and data organization. We study the characteristics of the four genres of sub-corpus through a series of discriminating variables — hyperstructural, lexical, semiotic, rhetorical, modal and syntactic — distributed at the infratextual and intratextual level. After that, based on the characteristics of each textual genre, we identify the main themes exposed in each genre of sub-corpus, and analyze the semantics of these identified themes in a contrastive way. Our analysis results are interpreted from two angles : quantitative and qualitative. All statistical analysis are assisted by textometric tools ; and the semantic interpretations are implemented on several fundamental concepts of SI (Sémantique interprétative) proposed by Rastier (1987).

Keywords : Digital humanity, text mining, textometrie, semantic analysis of corpus, textual genre, social network analysis, institutionnal discours analysis, chinese language processing, ecological corpus

Remerciement

On dirait souvent que « le parcours du doctorat est une Sādhana [1] solitaire en soi ». Mais, je dois dire je n'ai jamais été toute seule sur mon chemin de recherche, beaucoup de gens m'ont accompagnée, guidée, aidée, soutenue et encouragée tout au long de mon parcours doctoral à qui j'aimerais exprimer mes remerciements et ma gratitude.

Je tiens à remercier d'abord mon directeur de thèse, Mathieu Valette. Sans lui, ma thèse n'aurait jamais vu le jour. Sa rigueur scientifique, ses conseils éclairants, ses encouragements permanents et son humour inhérent ont été indispensables pour cette thèse. Grâce à lui, j'ai pu m'initier dans le domaine de la sémantique et de l'humanité numérique. Tout cela m'a guidé et me guidera dans mes futures recherches et dans ma vie personnelle.

Je présente également ma gratitude à mon co-encadrant, Nicolas Turenne, pour son intérêt infaillible, l'acuité de son regard sur le domaine de l'analyse du discours numérique, son soutien technique et la confiance qu'il a voulu m'apporter pour aboutir à ce travail. Les échanges ont été constructifs et fructueux qui ont considérablement fait progresser mon travail.

Je remercie les membres du jury, Monsieur le Professeur Damon Mayaffre et Monsieur le Professeur Hongmiao Wu, qui m'ont fait l'honneur d'avoir accepté de constituer le jury et de participer à évaluer mon travail.

Mes remerciements vont à tous mes collègues d'ERTIM pour le service de qualité et les conditions matérielles en or qu'ils apportent aux doctorantes et aux doctorants, et l'ambiance conviviale qu'ils créent. Plus spécialement à François Stuck, pour les petites expressions proverbiales drôles et les chansons françaises authentiques qu'il m'a recommandées ; à Damien Nouvelle, pour sa disponibilité et son soutien technique ; à Jean-Michel Daube pour l'appui logistique et les "spectacles" qu'il a joués de manière spontanée ; à Sophie Urbaniack, pour ses

1. Pratique spirituelle. Cf. Définiton de sadhana sur le site : https://sanskrit.inria.fr/DICO/69.html#saadhana. Consulté en février 2020.

aides administratives ; à Johanna Cordova et Amélie Martin, pour leurs relectures vigilantes, corrections minutieuses et conseils précis. Les collègues d'ERTIM sont de ceux avec qui le travail est un plaisir agréable. Sans leur aide, je n'aurais pas pu achever cette thèse.

Ma reconnaissance va également à mes amis qui m'ont accompagnée pendant mes années à Paris. Leurs pensées amicales, leur présence pétillante, leurs encouragements chaleureux m'ont donné la force nécessaire pour aller de l'avant. Je garde en tête les beaux moments que nous avons vécus ensemble dans les petits restaurants pour déguster (ou chez moi pour découvrir) les délices de la gastronomie chinoise ou européenne, dans les forêts pour ramasser les châtaignes, dans le parc de Saint-Cloud pour se promener, dans les TGV pour voyager, etc. À Xiaohua, Yewei, Yiyuan, Yasmine, Giorgio, Xiao Han, Saudade notre voisine, Liyun, Sha Ma.

Je dois mes remerciements à mes parents, ils m'ont toujours soutenue quelque soit ma décision. Ma mère, qui est très à l'écoute, est toujours là pour moi quand j'ai un besoin quelconque ; mon père, par son affection pour la littérature, m'a toujours inspirée sur les réflexions intellectuelles et philosophiques. Mes reconnaissances vont également à mes beaux-parents, qui, de près comme de loin, m'ont apporté inconditionnellement leurs soutiens moraux et matériaux. Merci à chaque repas qu'ils ont préparé soigneusement pendant la période de l'épidémie de Covid-19. Je souhaite remercier de tout mon cœur mon époux, sans qui je ne serais certainement pas arrivée au bout de cette aventure. Il est là, présent derrière chaque moment difficile pour m'aider, m'encourager, m'épauler, m'accompagner. Il est quelqu'un sur qui je peux toujours compter, que ce soit dans la recherche scientifique ou dans la vie quotidienne. Les mots ne suffisent pas pour lui exprimer toutes mes reconnaissances.

À ce moment spécial où le Covid-19 se propage dans le monde entier, j'aimerais adresser ma gratitude profonde à tous les personnels soignants, quelle que soit leur nationalité, qui travaillent de jour comme de nuit, pour l'humanisme, le professionnalisme et le dévouement dont ils font preuve face au danger. Ils sont les vrais héros de nos jours. Merci !

« 一个健康的社会不该只有一种声音 ».
— 李文亮

« Une société normale doit être celle qui tolère différentes voix ».
— Wenliang Li

Sommaire

Introduction

1 Le *wumai* en Chine

Depuis 2008, la dégradation de la situation environnementale en Chine est devenue un sujet brûlant. Le *wumai* — littéralement « brouillard », qui désigne le phénomène de pollution de l'air en Chine, suscite de vives discussions en ligne. Des institutions gouvernementales chinoises à la population civile, tout le monde s'inquiète de ce problème qui concerne des domaines très variés, de l'écologie à l'économie, de la politique à la vie quotidienne.

Suite à l'aggravation de la crise de la pollution de l'air, deux plateformes publient quotidiennement l'indice de qualité de l'air (AQI) de Chine. La première a été créée par le Centre national de surveillance de l'environnement de Chine ou AQISTUDY (中国空气质量在线监测分析平台)[1], rattaché au ministère de l'Écologie et de l'Environnement chinois. Les données de la qualité de l'air y ont été recensées et publiées de décembre 2013 à septembre 2018. La seconde a été lancée par l'Ambassade des États-Unis en Chine en 2008 en même temps qu'un programme de surveillance de la qualité de l'air en Chine. Le site internet a publié l'indice PM2,5[2] de l'air[3] heure par heure jusqu'en juin 2017, pour cinq grandes métropoles de Chine : Beijing, Shanghai, Chengdu, Guangzhou, Shenyang.

Du fait du modèle du développement économique dépendant essentiellement du charbon[4], les régions du nord de la Chine subissent davantage le brouillard de

1. Source d'information : https://www.aqistudy.cn/historydata/about.php.

2. La méthode de référence utilisée pour l'échantillonnage et la mesure des PM2,5 est celle décrite dans la norme EN 14907 (2005) : "Méthode de mesurage gravimétrique de référence pour la détermination de la fraction massique PM2,5 de matière particulaire en suspension." Source : Association Nationale pour la Prévention et l'Amélioration de la qualité de l'Air, site internet : http://www.respire-asso.org/particules-en-suspension-pm10-pm-25/. Consulté en janvier 2019.

3. Source d'information : http://stateair.net/. Consulté en octobre 2018. Après la comparaison des données offertes par les deux organismes (cf. Annexe 1), nous allons utiliser les graphes générés par AQISTUDY pour présenter la projection spatiale et évolution temporelle du *wumai* en Chine.

4. Source d'information : https://www.statista.com/statistics/265458/

pollution que celles du sud, dont Beijing et certaines provinces qui l'entourent, telles que Hebei[5], Shandong[6], Henan et Shanxi qui constituent les principales victimes de ce phénomène.

En plus de la distribution spatiale de la pollution atmosphérique, les mesures publiées sur les plateformes précédemment citées nous renseignent aussi sur l'évolution temporelle de la pollution de l'air dans l'ensemble de la Chine. Grâce à la mesure « de Charbon à gaz »[7], nous pouvons constater une amélioration évidente du *wumai* à partir de la fin de l'année 2017 dans l'ensemble de la Chine. Plus spécifiquement, pour Beijing et Shanghai, le niveau moyen de l'intensité des particules fines a commencé à baisser après avoir atteint le pic en 2015 (Shanghai) et en 2016 (le cas de Beijing) (voir Figure 1.2 Beijing vs Shanghai : L'évolution temporelle du PM2,5 de 2014 à 2018[8]).

Dans la mesure où le brouillard de pollution présente un taux d'humidité relativement élevé, des bactéries et des virus peuvent y proliférer. Respirer l'air pollué entraîne un important risque d'infection par ces bactéries et ces virus. Le brouillard de pollution contient de plus des particules de métaux lourds qui peuvent pénétrer facilement dans le système de circulation sanguine via les poumons. La population serait ainsi davantage exposées aux maladies respiratoires, pulmonaires, cardiovasculaires, voire au risque de cancer. L'Organisation Mondiale de la Santé (OMS) estime qu'environ 58% des décès prématurés liés à la pollution de l'air extérieur résultaient en 2016 de cardiopathies ischémiques et d'accidents vasculaires cérébraux, 18% de bronchopneumopathies chroniques obstructives ou d'infections aiguës des voies respiratoires inférieures, tandis que les 8% restants sont imputables au cancer du poumon[9], etc. Selon les résultats d'analyse (Dang, Turenne et Valette, 2018), en Chine, les maladies causées par le

---

chinese-coal-production-in-oil-\equivalent/. Consulté en octobre 2018.

5. Hebei a produit 245,51 millions tonnes d'acier brut en une seule année, ce qui représente environ un quart de la production globale en Chine – 1046,42 millions tonnes d'acier brut. http://data.stats.gov.cn.

6. La capacité du traitement du pétrole brut de la province du Shandong s'établit à 210 millions tonnes par an, soit 28% de la production totale de Chine, selon les chiffres de 2018 fournis par le gouvernement de Shandong. Source d'information : https://www.sohu.com/a/235862337_617351. Consulté en octobre 2018.

7. Source d'information : http://www.gov.cn/gzdt/2008-06/20/content_1022215.htm,ethttps://www.francetvinfo.fr/monde/asie/pollution-en-chine/chine-la-fin-du-charbon_2512109.html. Consulté en octobre 2018.

9. Source d'information : https://www.afro.who.int/fr/node/3828. Consulté en octobre 2019.

brouillard de pollution se concentrent sur les régions les plus touchées par ce problème environnemental, notamment, celles du Beijing, Hebei, Tianjin, Shanghai, etc. En outre, les régions, en fonction de leur position géographique et de leur modèle de développement économique, sont touchées par des types de problèmes de santé distincts. Par exemple, Beijing présente un fort taux de cancer (癌症 (cancer)), et des maladies et symptômes liés au système respiratoire, tels que la 支气管炎 (bronchite), la 咳嗽 (toux), la 上呼吸道感染 (inflammation des voies respiratoires), la 打喷嚏 (sternutation), la 刺鼻 (irritation du nez) ; alors que les maladies typiques de la province du Hebei est le 癌症 (cancer), le 肺癌 (cancer du poumon), et la 心脏病 (cardiopathie) ; pour Tianjin, nous récupérons les maladies ou les symptômes quotidiens comme représentatives, par exemple le 过敏 (allergie), le 头痛 (mal de tête), la 鼻炎 (rhinite), et la 刺鼻 (irritation du nez) ; la ville de Shanghai est caractérisée par les maladies maladies respiratoires : le 哮喘病 (asthme), la 咳嗽 (toux) et la 肺炎 (pneumonie).

Dans le but de limiter les effets nocifs du brouillard de pollution, de multiples mesures et efforts sont proposés et émis par le gouvernement et la population chinoise afin de réduire la pollution atmosphérique et de se protéger du *wumai*. La population se base sur l'expérience pratique issue de la vie quotidienne, elle recommande « de porter le masque, d'équiper l'appartement ou la maison de purificateurs d'air, de manger plus léger et des nourritures qui permettent d'humecter les poumons ou de déduire l'infection, comme les poires et les champignons noirs, etc.». Les jours où le brouillard de pollution atteint un niveau très élevé, il faut « réduire autant que possible la sortie à l'extérieur » [10]. Le gouvernement quant à lui considère qu'il faut mettre en place des mesures et politiques sur le long terme pour véritablement résoudre le problème de *wumai*. Nous pouvons citer comme exemple la transition énergétique — la mesure de « de charbon à gaz » évoquée plus en haut en fait partie — , qui vise à diminuer la consommation des énergies fossiles (le charbon, le pétrole), utiliser davantage les énergies vertes et renouvelables (énergies solaire, éolienne et hydraulique), et à réduire les émissions des usines. La mise en œuvre des dispositifs administratifs et judiciaires est aussi proposée par le gouvernement afin d'assurer la bonne application des mesures mentionnées ci-dessus.

10. Source d'information : https://wenku.baidu.com/view/8ce0460d43323968011c9256.html. Consulté en octobre 2019.

2 Les quatre genres de sous-corpus

Les sites web jouent le rôle de caisse de résonance de la société. Ils fournissent de multiples informations dans divers domaines : politique, économie, écologie, éducation, sport, santé, divertissement, etc. En tant que plateforme de communication, ils créent des liens sociaux entre le monde réel et le monde virtuel. À travers la transmission, l'appel, la participation, l'échange et le débat, ces textes sont présentés soit sous forme de publications d'articles ou de *weibo*[11] ; soit sous forme de commentaires entre les producteurs et les utilisateurs/lecteurs, ou entre les utilisateurs/lecteurs mêmes. Ces textes publiés nous rendent compte de ce qui s'est passé/se passe/se passera dans le monde. Ils permettent également de susciter l'attention des individus sur des sujets délicats à discuter, à convaincre, ou encore à simplement transmettre des idées des décideurs. Tous ces procédés de production des informations issus des Web et des réseaux sociaux donnent naissance aux données textuelles numériques. Depuis une vingtaine d'années, de plus en plus de corpus dérivés de sites web se relèvent des sources précieuses de matériaux pour mener des recherches en sciences humaines et linguistiques. Ces textes nous ouvrent accès aux opinions du grand public, mais aussi à celles des autorités.

Ces dernières années, dans un contexte de dégradation environnementale et de sensibilisation croissante du grand public à ce problème, le brouillard de pollution est devenu le sujet délicat de nombreuses données textuelles numériques. Les textes en ligne portant sur le brouillard de pollution se sont développés très rapidement et relèvent de différents champs d'expression (politique, sciences naturelles, sciences sociales, etc.). Par ailleurs, les utilisateurs ou les producteurs d'horizons différents nous offrent l'accès à presque tous les points de vue possibles. Ces textes proviennent aussi bien des messages laissés par les utilisateurs sur les plateformes d'expression libre (blogs ou réseaux sociaux), que des articles publiés sur des sites informationnels ou institutionnels. Ils appartiennent respectivement à quatre genres de sous-corpus définis par nous-mêmes : le sous-corpus institutionnel (Ins), le sous-corpus institutionnel-médiatique (InsM), le sous-corpus informel-médiatique (InfM) et le sous-corpus profane (Profane).

11. Équivalent de tweets, utilisé pour désigner les messages publiés sur le site WEIBO.

2.1 *Sous-corpus Ins*

Le sous-corpus institutionnel est produit par un « énonciateur singulier ou collectif qui occupe une position juridiquement inscrite dans l'appareil d'État » (Oger et Ollivier-Yaniv, 2003). Dans ce sens, les articles publiés sur le site du gouvernement chinois par les éditeurs officiels peuvent être tous définis comme des textes institutionnels. Les textes institutionnels ont pour l'objectif principal soit de transmettre des décisions des hiérarchies, soit de rendre compte publiquement des grandes affaires internes nationales ou externes internationales. Ils diffèrent de ceux publiés sur les réseaux sociaux à plusieurs égards. Ces distinctions résident dans le mode de transmission, le temps de mise à jour des informations, les contenus thématiques présentés ainsi que le langage discursif adopté.

D'abord, les sites traditionnels, ici, en l'occurrence l'Ins, l'InsM et l'InfM, ont des éditeurs professionnels qui s'occupent de la rédaction jusqu'à la publication des articles, en passant par la mise en forme des textes. Ils doivent respecter les normes et règles établies par les institutions à chaque chaînon. De plus, les procédures de la rédaction sont plus compliquées et délicates, qui prennent plus longtemps que ceux produits par les utilisateurs eux-mêmes de manière spontanée. De ce fait, le temps de production et de mise à jour des informations est plus long que celui pour le *weibo*. Ensuite, leur mode de transmission des informations suit l'ordre univoque « de haut en bas », c'est-à-dire des décideurs vers les lecteurs/utilisateurs, et non inverse. Ce qui se traduit aussi par l'absence ou la déficience du cadre réservé au commentaire. Et puis, contraint par sa vocation principale et son organisation du site, les contenus thématiques des textes sont uniformisés, pour la plupart politiques ou diplomatiques. Ce qui fait que le site adopte un langage « uniformisé » afin de garder sa forte identité et son autorité institutionnelle. En ce sens, les textes institutionnels en ligne construisent un **système d'« exportation »** (SE) statique et distant pour transmettre des informations homogènes du haut en bas dans une sphère unitaire (cf. Figure 1 Système unidirectionnel d'« exportation » des textes institutionnels).

2.2 *Sous-corpus InsM et sous-corpus InfM*

Le sous-corpus institutionnel-médiatique (InsM) partage, d'une part, la caractéristique « institutionnelle » dans un sens plus large, qui « comprend l'ensemble des discours que l'on peut considérer à des degrés divers comme des discours auto-

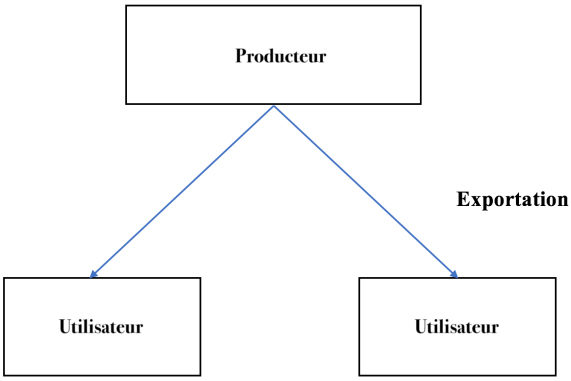


FIG. 1 – Système unidirectionnel d'« exportation » des textes institutionnels

risés dans un milieu donné (en ou hors contexte officiel), sans référence nécessaire à l'État (production des syndicats, des états — majors des partis politiques, etc.)» (OGER et OLLIVIER-YANIV, 2003) ; d'autre part, il fait partie du discours médiatique, qui est produit par des journalistes, des rédacteurs d'institutions pour un groupe d' « audiences cibles » à l'aide d'une mise en forme médiatique et un mode de transmission sémiotique.

Par analogie, le sous-corpus informel-médiatique (InfM) fait aussi partie du discours médiatique, mais de nature informelle, il est produit dans un cadre socio-économique et sémiologique (CHARAUDEAU, 2003). Comparée au sous-corpus InsM, d'une part, l'exigence professionnelle du sous-corpus InfM s'avère d'abord dans la sélection, l'évaluation et la problématisation de l'information de masse afin de « maintenir [une] position compétitive et de capter le plus grand nombre possible d'audiences » ; d'autre part, de divers outils et techniques médiatiques ont mis en avant et en valeur, par les pratiques professionnelles, le centre d'intérêt, les logiques et la représentation du monde des producteurs de l'information médiatique.

En apparence, le mode de transmission d'informations de ces deux types de sous-corpus est similaire que celui de l'Ins — unidirectionnel « de haut en bas ». Toutefois, la présence de supports d'expression libre (blogs, forum) dans les lieux des sous-corpus InfM et InsM favorise la communication interactive entre les émetteurs (producteurs) et les récepteurs (lecteurs). Il faut noter que ce type de communication ne peut pas équivaloir à celle du sous-corpus Profane (nous verra

la différence majeure dans la partie 2.3 Sous-corpus Profane). Les messages laissés sur les forums ou les blogs par les lecteurs-récepteurs ne peuvent être considérés comme des informations additionnelles importées. Par rapport à la production principale du contenu de ces deux types de sous-corpus, ces contributions textuelles ne jouent qu'un rôle de complément. Ainsi, nous désignons ce mode de transmission d'informations de l'InsM et de l'InfM dans un **système mixte** (SM) d'« exportation » et d'« importation » additionnel optionnel, comme montré par la figure suivante :

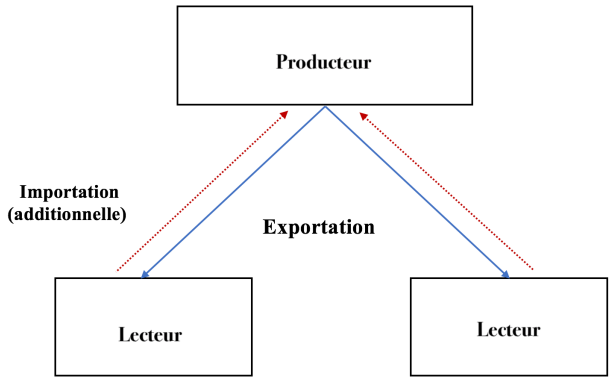


FIG. 2 – Système mixte des sous-corpus InsM et InfM

2.3 *Sous-corpus Profane*

Le sous-corpus Profane, appelé par GOMEZ-MEJIA *discours vernaculaire*, est un sous-corpus plutôt « polymorphe, fluctuant » et « polyphonique » que les dispositifs du Web contemporain sollicitent aux internautes (GOMEZ-MEJIA, 2010). La polymorphie du sous-corpus Profane se manifeste d'abord par une large gamme de formes interactives de présentation : les blogs, les chats, les réseaux sociaux, etc. ; ensuite par la multidimensionnalité du contenu : du faits divers à la politique, du sport à l'économie, de la partage de la vie personnelle à la publication des publicités, etc. En créant « une réalité mixte » par un mélange d'énonciateurs (CALABRESE, 2014), le discours englobe à la fois la voix citoyenne, experte, politicienne, et journalistique. Le sous-corpus Profane, caractérisé par le dispositif *user-generated-content*, met en valeur la place centrale des utilisateurs. Dans le sous-corpus Profane, d'un côté, les utilisateurs sont eux-

mêmes des émetteurs-producteurs qui produisent en temps réel des informations spontanées ; de l'autre, ils jouent le rôle de récepteurs-rénovateurs, car lorsqu'ils consomment les contenus publiés par d'autres utilisateurs, ils réagissent, parfois rectifient, complètent ou voire mettent en cause les informations par leurs commentaires. Dans ce sens, ces récepteurs deviennent de nouveaux « producteurs ». Ce dispositif du sous-corpus Profane crée ainsi un **système circulaire** (SC) interactif d' « exportation — importation — exportation » de production dynamique des informations hétérogènes dans une sphère multidimentionnel.

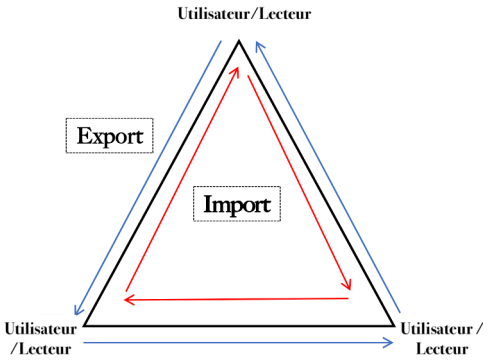


FIG. 3 – Système circulaire de production « importation — exportation — importation » du sous-corpus Profane

3 Hypothèse et problématique du travail

En nous basant sur ces trois modes de circulation des informations (SE, SM et SC), nous exploiterons notre corpus numérique constitué de textes appartenant aux quatre genres de sous-corpus dans le cadre théorique de la sémantique interprétative (SI) proposé par Rastier en 1987.

Nous avons établi l'hypothèse principale suivante : les quatre genres textuels du corpus partagent les thèmes communs au palier mésosémantique, mais se différencient voire s'opposent l'un à l'autre au palier macrosémantique par leurs caractéristiques distinctes du genre textuel. Il en résulte que les thèmes sont interprétés de manière différente au palier microsémantique.

Ainsi, dans le cadre de notre recherche, nous étudierons, de manière contras-

tive avec la méthodologie statistique quantitative de la textométrie, notre corpus écologique relatif au sujet « brouillard de pollution en Chine » sur trois questions principales : 1) les caractéristiques singulières de chaque genre textuel de sous-corpus à travers une série de variables linguistiques et sémiotiques ; 2) les thèmes centraux de chaque sous-corpus à travers les sèmes et l'isotopie ; 3) la sémantique de chaque thème selon la co-interaction des composantes sémantiques.

4 Cadre théorique et méthode technique

4.1 *Cadre théorique : Sémantique interprétative*

Chaque texte procède d'un genre, et le genre (global) détermine la sémantique textuelle (le local) du corpus. Autrement dit, au niveau macroscopique, la sémantique textuelle provient du genre textuel, qui est lié aux pratiques et aux normes sociales. Au niveau microscopique, sur les formes sémantiques, elle émane des termes lexicalisés ; sur le fond sémantique, elle provient des termes non-lexicalisés mais récurrents comme les isotopies (Biber (1992), Denise et François (2001) et Rastier (1996)).

L'analyse du genre textuel et l'étude de la sémantique sont liées de manière très étroite. Au niveau macroscopique, le genre exerce une influence sur les lexiques et les valeurs sémantiques au niveau mésoscopique et microscopique. Différents genres de textes se distinguent l'un de l'autre par les différents comportements linguistiques au niveau lexical, sémiotique, rhétorique, syntaxique, morphologique, ainsi que sur l'organisation et la construction des énoncés des textes. De même, l'interprétation de la sémantique textuelle change en fonction des modes de co-interaction des termes lexicalisés ou non-lexicalisés qui sont actualisés et véhiculés dans les genres différents de sous-corpus.

Basée sur ce principe de base dans le cadre théorique de la sémantique interprétative de Rastier (2001), notre analyse sémantique contrastive repose sur quelques propositions rastériennes que nous préciseront dans le Chapitre 2 Cadre théorique et méthodologique. Ces concepts théoriques sont adéquats et pertinents pour nos analyses sémantiques du sujet du « brouillard de pollution en Chine » sur quatre genres de sous-corpus.

L'étude et l'interprétation de la sémantique des quatre genres de sous-corpus se déploient sur trois paliers sémantiques et quatre composantes sémantiques (ci-

infra). Ces éléments nous permettent à répondre à notre problématique qui s'est posée plus haut.

- **Trois paliers sémantiques** :
  - **macrosémantique** : nous relèverons les caractéristiques de chaque genre de sous-corpus à partir d'une série de paramètres discriminants (lexicaux, sémiotiques, modaux, rhétoriques et syntaxiques) au niveau infratextuel et intratextuel ;
  - **mésosémantique** : nous étudierons les isotopies (les mêmes sèmes récurrents) et les thèmes (la récurrence de deux sèmes distincts) à travers les termes lexicalisés ;
  - **microsémantique** : nous nous intéresserons aux lexies et aux mots.
- **Quatre composantes sémantiques** :
  - **dialogique** : fonde la typologie des énonciateurs représentés, et rend compte des modalités énonciatives et évaluatives.
  - **dialectique** : met en jeu des acteurs, des rôles et des fonctions du récit. Elle assume son rôle d'argumentation dédiée aux éléments descriptifs ou argumentatifs.
  - **thématique** : autrement dit le sujet d'un texte. Elle est déterminée par des contenus investis par des unités récurrentes sémantiques structurées (Rastier, 2001).
  - **tactique** : permet de définir des rythmes sémantiques.

4.2 *Méthode technique : Textométrie*

La théorie de la SI constitue le cadre théorique auquel s'attachent nos analyses qualitatives, l'ingénierie de la textométrie forme notre appui technique, qui assiste notre interprétation sémantique à travers des calculs quantitatifs.

La textométrie propose des méthodes qui s'inspirent des principes fondamentaux de la théorie de la SI. En travaillant sur des corpus de textes intégraux, elle cherche avant tout à décrire et à interpréter des phénomènes linguistiques qui pourraient être négligés par l'humain. De plus, la mise à profit des procédures de tris et des calculs statistiques outillés par des logiciels TAL (Traitement Automatique des Langues) permet d'assister et outiller la lecture humaine dans le respect des données recueillies, de lui suggérer des points d'appui, des pistes d'investigation, et de mettre en évidence des régularités et des spécificités, etc.

5 Plan de la thèse

Dans le Chapitre 1, d'abord, nous présenterons le mot 雾霾 (brouillard de pollution) à deux niveaux : l'explication étymologique du *wumai* à partir de deux caractères qui le composent — le 雾 wu et le 霾 mai , et la description du phénomène du *wumai* dans le domaine écologique à travers ses composantes principales. Ensuite, nous dépeindrons le contexte général du problème écologique à propos du *wumai* : 1) la situation actuelle à travers la projection spatio-temporelle du brouillard de pollution en Chine de 2013 à 2018 ; 2) les trois topics les plus discutés dans les textes numériques en ligne à propos du *wumai* : les origines principales du *wumai*, les impacts du *wumai* sur la santé et les mesures préventives contre le *wumai*. Et à la fin du chapitre, nous aborderons une brève présentation de la question de la *censure d'Internet en Chine*.

Le Chapitre 2 est consacré à la présentation du cadre conceptuel théorique et la méthode technique du travail. Il est composé de trois sections principales : premièrement, la méthodologie qualitative. Nous rappellerons des notions et des concepts fondamentaux de la SI : 1) les trois paliers sémantiques sur lesquels axent respectivement a) l'étude du genre textuel, b) l'analyse des thèmes sur le fond sémantique qui est actualisée par c) les formes sémantiques lexicalisées en mots ; 2) les quatre composantes sémantiques : thématique, dialectique, dialogique et tactique ; deuxièmement, la méthode quantitative. Nous présenterons la textométrie et les outils techniques utilisés, y compris les fonctionnalités principales utilisées ; troisièmement, les travaux précédents. Nous listerons et commenterons les travaux de recherches effectués par d'autres chercheurs dans le domaine de l'exploitation du genre textuel, de la sémantique textuelle et de la textométrie.

Le Chapitre 3 sera consacré à la constitution du corpus. Après un bref exposé du choix des sources, nous expliquerons dans les détails, étape par étape, les procédés et les outils techniques qui nous ont permis de récupérer les données, mais aussi de normaliser, nettoyer, segmenter, annoter et organiser le corpus. Nous aborderons également les spécificités de la langue chinoise, notamment au niveau de l'encodage et de la segmentation, qui engendrent des difficultés supplémentaires quant à son traitement informatique. Enfin, nous effectuerons une présentation générale des informations quantitatives du corpus.

Afin de répondre aux questions que nous nous sommes posées dans le Chapitre

4 et le Chapitre 5, qui sont liées à la problématique du travail dans notre corpus, nous explorerons de manière constrastive et méthodologique le corpus sur deux axes : 1) il s'agit d'abord d'étudier, à travers une série de variables discriminantes, les caractéristiques des quatre genres textuels auxquels sont associés nos quatre sous-corpus ; et puis, 2) d'identifier les thèmes principaux exposés dans chaque genre de sous-corpus à travers les sèmes et l'isotopie, et puis d'exploiter encore la sémantique de chaque thème dans chaque sous-corpus. Toutes les analyses statistiques sont assistées par l'ingénierie et les outils de la Textométrie, toutes les interprétations seront effectuées dans le cadre théorique de la SI.

Dans le Chapitre 6, nous présenterons de manière synthétique les résultats obtenus après toutes les études quantitatives et qualitatives. Pour chaque genre de sous-corpus, nous résumons ses caractéristiques du genre textuel ainsi que les résultats d'analyses sémantiques des thèmes identifiés.

Chapitre 1

Contexte du domaine de la pollution atmosphérique en Chine

1.1 Introduction

Selon les résultats du Rapport du Développement durable de la Chine 2019 (« 2019 中国可持续发展报告 ») mené par l'Académie des Sciences sociales de Chine[1], et des recherches sur les problèmes majeurs de la Chine effectuées par d'autres chercheurs (Yang et al., 2011 ; Huchet, 2016 ; Gunson et Veiga, 2004 ; Larssen et al., 2006 ; Liu et Diamond, 2005 ; Wong et al., 2006), dix troubles sociaux font obstacle au progrès continu de la Chine : l'éducation et la médecine, le prix de l'immobilier, la corruption, le vieillissement de la population, la persistance des difficultés dans l'emploi, l'élargissement de l'inégalité des revenus, les problèmes de pensions, les problèmes affectant les régions rurales et la pollution de l'environnement (en particulier la pollution de l'air, la pollution de l'eau et la désertification). Si durant la dernière décennie, la Chine a connu une période de croissance rapide avec un fort développement économique, les problèmes environnementaux font à présent sérieusement obstacle au développement durable du pays. La pollution de l'air, en particulier, a déjà atteint un degré alarmant et s'aggrave chaque jour davantage. Elle a déjà provoqué 1,6 million de décès prématurés par an, selon une étude de Robert A. Rhode et de Richard A. Muller de l'Université de Berkeley. Dans cette partie, nous nous donnons comme objectif de présenter de manière générale ce que l'on désigne par le 雾霾, qui est la manifestation la plus visible de la pollution de l'air : nous verrons quelle est sa source, quels sont ses effets négatifs ainsi que les mesures proposées ou adoptées par le gouvernement et la population chinoise pour contenir ces effets nocifs et

1. Source d'information : https://wenku.baidu.com/view/21f4f780bdd126fff705cc1755270722182e594c.html.Consulté en octobre 2019

pour se protéger.

1.2 Brève description du *wumai*

雾霾 est une pollution de l'air constituée de particules ultrafines (ou PM, *Particulate Matter*, <10 μm pour les PM10 et <2,51μm pour les PM2,5) qui sont en suspension dans l'air. Le diamètre de ces particules peut être inférieur à 2,5 micromètres, ce qui fait qu'elles sont facilement portées par l'eau ou par l'air, et leur forte concentration dans l'air réduit considérablement la visibilité. La méthode de référence utilisée pour l'échantillonnage et la mesure des PM10 est celle décrite dans la norme EN 12341 (1999) [2].

雾霾 est un mot composé [3] de deux caractères chinois 雾 (brume) et 霾 (smog), désignant deux phénomènes différents. Commençons par envisager le premier mot, 雾 (brume). Il est lui-même constitué de deux caractères : la partie supérieure 雨, signifie « pluie », ou dans un sens plus large « humidité » ; la partie inférieure, 务, désigne la « vue floue ». De cela, nous pouvons déduire que le mot 雾 exprime l'idée d'« avoir la vue floue à cause de la pluie (ou de l'humidité)». Quant au mot 霾 (smog), il est composé de trois parties. On retrouve 雨 (pluie) dans la partie supérieure ; dans la partie inférieure gauche, 豺 représente les « animaux ». Dans la partie inférieure droite, 里 constitue la forme elliptique du verbe 埋 signifiant lui-même « enterrer ». La combinaison de ces trois éléments nous donnent la traduction littérale suivante : « la pluie et le vent font voler les poussières et les sables qui enterrent les animaux ». On remarque que le mot 雾 (brume) désigne la vue réduite, sans pour autant porter une connotation négative, tandis que 霾 (smog) insiste non seulement sur le fait qu'« on ne voit pas », mais aussi sur le caractère délétère de ce phénomène. Ainsi, quand ces deux mots se combinent pour former une expression figée 雾霾 — le brouillard de pollution, ils co-signifient « une brume brunâtre épaisse, provenant d'un mélange de polluants atmosphériques, qui limite la visibilité dans l'atmosphère » [4].

2. Qualité de l'air — détermination de la fraction PM10 de matière particulaire en suspension — méthode de référence et procédure d'essai in *situ* pour démontrer l'équivalence à la référence de méthodes de mesurage

3. Source d'information : http://www.vividict.com/WordInfo.aspx?id=2056. Consulté en janvier 2019.

4. Source d'information : https://www.linguee.fr/francais-anglais/search?source=auto&query=smog. Consulté en mai 2018.

En plus du mot 雾霾, d'autres dénominations sont aussi utilisées pour désigner la pollution de l'air liée aux particules fines : nous avons vu plus haut 霾 (le smog) ; on trouve également 空气污染 (la pollution de l'air), 大气污染 (la pollution atmosphérique), PM2,5, PM10. Ce sont ces mots qui constitueront la liste des différentes dénominations de 雾霾 utilisée pour récupérer les données textuelles.

1.3 La situation actuelle du *wumai* en Chine

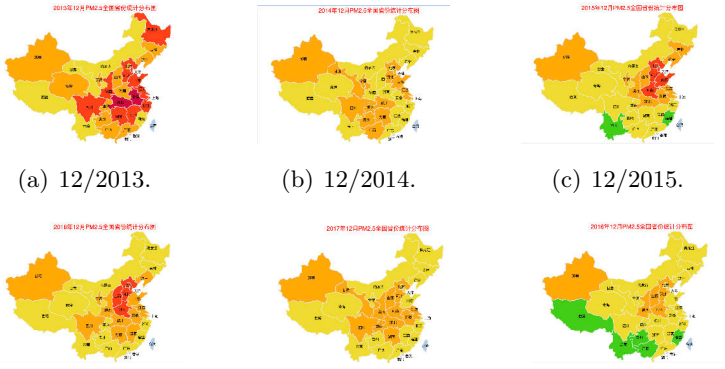
Avant de présenter l'évolution spatio-temporelle du brouillard de pollution en Chine, nous allons expliquer quelles sont nos sources pour les données de l'indice de qualité de l'air en Chine. Puis, en nous appuyant sur les données sélectionnées, nous allons montrer la distribution régionale et l'évolution temporelle du *wumai* en Chine.

1.3.1 *Projection spatiale et évolution temporelle du wumai en Chine*

L'intensité des particules fines dans une ville ou une région donnée ne reste pas toujours la même durant toute l'année. De manière générale, la qualité de l'air se dégrade au fur et à mesure que la température baisse. Ainsi, la qualité de l'air est sensiblement plus mauvaise en hiver qu'en été, et le pic de pollution atmosphérique apparaît souvent dans la période hivernale [5]. Afin de visualiser l'évolution temporelle du *wumai*, nous avons sélectionné six graphiques (cf. figure 1.1 Distribution régionale du brouillard de pollution en Chine de 2013 à 2018 [6]), obtenus depuis AQISTUDY [7], qui montrent l'intensité de la pollution dans l'ensemble de la Chine, du mois de décembre de 2013 à 2018.

5. Source d'information : https://e-rse.net/pollution-air-hiver-ete-saisons-25414/. Consulté en novembre 2018.

7. Source d'information : https://www.aqistudy.cn/historydata/about.php. Consulté en novembre 2018.



(a) 12/2013. (b) 12/2014. (c) 12/2015.

(d) 12/2016. (e) 12/2017. (f) 12/2018.

FIG. 1.1 – Distribution régionale du brouillard de pollution en Chine de 2013 à 2018 [8]

Ces graphiques mettent en évidence 1) une concentration des polluants dans la région de Beijing ainsi que dans certaines provinces qui l'entourent telles que Hebei, Shandong, Henan et Shanxi ; 2) une diminution progressive de la pollution atmosphérique à deux niveaux : au niveau de la position géographique - du Nord au Sud (Beijing par rapport à Yunnan), et de l'Ouest à l'Est (Xinjiang par rapport à Fujian) ; au niveau temporel, du moins récent au plus récent (l'année 2013 par rapport à l'année 2018). Il faut noter que cette situation de la projection et de l'évolution est liée principalement aux activités humaines : le modèle du développement économique (la production industrielle, la combustion des charbons) et le mode de vie (l'émission des automobiles) (cf. sections 1.4.1.1, 1.4.1.3 et 1.4.1.2). La baisse de la concentration du *wumai* permet d'un côté de confirmer l'effet de la mesure politique « Charbon à gaz » (cf. section 1.4.3.2) proposée en début 2017 : à partir de décembre 2017, une amélioration évidente du *wumai* se forme dans l'ensemble de la Chine. Plus spécifiquement, pour Beijing et Shanghai, le niveau moyen de l'intensité des particules fines a commencé à baisser, après avoir atteint son pic en 2015 (le cas de Shanghai) ou en 2016 (le cas de Beijing) (voir figure 1.2 Beijing vs Shanghai : L'évolution temporelle du PM2,5 de 2014 à 2018 [9]). La synthèse des résultats exposée dans ces graphiques pourra nous permettre d'établir d'éventuelles corrélations entre la pollution de l'air et

8. Plus la couleur tend vers le rouge-violet, plus la concentration en polluants est élevée, Plus la couleur tend vers le vert, plus la concentration en polluants est élevée. Source d'information : https://www.aqistudy.cn/historydata/about.php.

la répartition des maladies affectant la population chinoise.

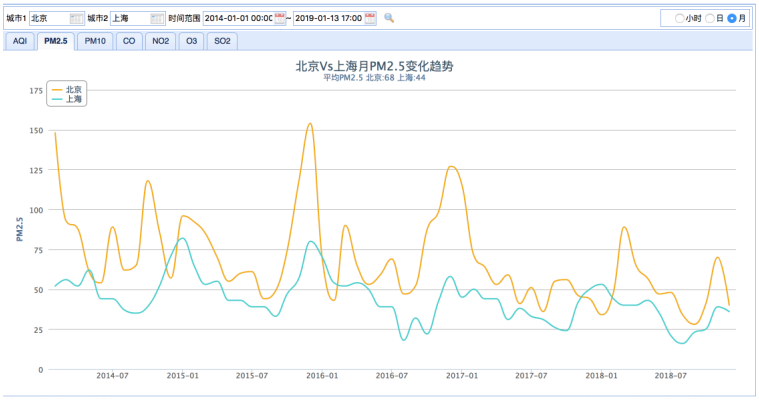


FIG. 1.2 – Beijing vs Shanghai : L'évolution temporelle du PM2,5 de 2014 à 2018 [10]

1.4 À propos du *wumai* en Chine

1.4.1 *Causes principales du wumai*

Le *wumai* est composé d'éléments hétérogènes. Ainsi, dans le brouillard de pollution, on trouve des éléments différents tels que le monoxyde de carbone (CO), le dioxyde de soufre (SO2), l'ozone (O3), ou les oxydes d'azote (NOx), mais aussi des composés organiques volatils (COV), des métaux lourds (plomb (pb), mercure (hg), arsenic (as), cadmium (cd), nickel (ni), etc.), et des PM (particules en suspension) [11]. Ces éléments proviennent de sources diverses. Des études montrent que le brouillard de pollution est principalement (à 90%) engendré par les activités humaines (HUCHET, 2016), par exemple l'industrie lourde, la consommation du charbon, les chantiers de construction, les émissions des véhicules, etc. Le reste des composés sont des matériaux naturels : sables, sels, cendres issues de feux ou cendres volcaniques, etc. Plus les particules sont fines, plus elles sont dangereuses pour la santé [12].

10. Source : https://www.aqistudy.cn/historydata/about.php

11. Source d'information : « La réglementation technique de l'indice de la Qualité de l'Air » fourni par le ministre de la Protection de l'environnement en Chine. Source d'information : http://kjs.mee.gov.cn/hjbhbz/bzwb/jcffbz/201203/W020120410332725219541.pdf. Consulté en janvier 2019.

12. Voir l'intervention de Bruno Housset au 22e Congrès de pneumologie de la langue française, qui a lieu du 26 au 28 janvier au

Dans ce qui suit, nous détaillerons les sources principales du brouillard de pollution en Chine.

1.4.1.1 Consommation du charbon

La pollution atmosphérique en résultant est en grande partie due à la dépendance de la Chine au charbon. Des statistiques publiées sur le site Statista [13] montrent que la Chine a été et reste toujours grande consommatrice de charbon, et ce, même avant la mise en œuvre de la politique de réforme et d'ouverture en 1978, c'est-à-dire avant son essor économique. En 2016, par exemple, la Chine a consommé 4 milliards de tonnes de charbon [14]. En général, nous constatons une baisse continue de la part du charbon dans la consommation énergétique du pays, mais cette baisse n'est pas significative et la part du charbon se maintient à 62 % environ (chiffre de 2017 [15]).

1.4.1.2 Émission industrielle

Une autre cause de la grave pollution atmosphérique en Chine réside dans le fait que l'industrie lourde prend une place importante dans l'économie chinoise. Dans la province du Hebei (au sud de Beijing), la métallurgie constitue un pilier de l'économie de la région. En 2018, Hebei a produit 245,51 millions de tonnes d'acier brut en une seule année, ce qui représente environ un quart de la production globale en Chine — 1046,42 millions tonnes d'acier brut [16], ce qui fait de la Chine le premier producteur d'acier du monde. Cette concentration de la capacité de production d'acier dans la seule province du Hebei explique en grande partie la dégradation de l'environnement (surtout de la qualité de l'air) observée dans cette région. Ainsi, en 2017, parmi les dix villes ayant enregistré un taux d'intensité de particules fines (PM2,5) le plus élevé, cinq relèvent de cette région : au podium se

Centre de Congrès de Lyon. Source : http://sante.lefigaro.fr/article/quel-est-l-impact-de-la-pollution-atmospherique-sur-notre-sante-/. Consulté en janvier 2019.

13. Source d'information : https://www.statista.com/statistics/265458/chinese-coal-production-in-oil-equivalent/. Consulté en janvier 2019.

14. Selon les statistiques de Planétoscope : https://www.planetoscope.com/Source-d-energie/1036-consommation-de-charbon-en-chine.html. Consulté en février 2019.

15. Source d'information : http://www.chem99.com/news/26704416.html. Consulté en janvier 2019.

16. Source d'information : National Bureau of Statistics of China. http://data.stats.gov.cn. Consulté en février 2019.

trouvent Shijiazhuang, Handan, et Xingtai, ces villes sont talonnées par Baoding et Tangshan qui occupent respectivement la quatrième et la cinquième place [17].

En plus de la métallurgie, l'industrie de la pétrochimie est aussi responsable de la pollution atmosphérique de certaines régions de Chine. Comme Hebei, la province du Shandong est parmi les régions les plus touchées par la dégradation de la qualité de l'air. Pour cause, sa dépendance de l'industrie pétrochimique et sa gigantesque capacité de production. En effet, la capacité du traitement du pétrole brut de la province du Shandong s'établit à 210 millions tonnes par an, soit 28% de la production totale de la Chine, selon les chiffres de 2018 fournis par le gouvernement du Shandong [18]. Rien d'étonnant donc à ce que la ville de Jinan, le chef-lieu du Shandong, figure parmi les dix villes dont la qualité de l'air est la plus mauvaise [19].

1.4.1.3 Émission des automobiles

La pollution de l'air dans les grandes villes chinoises est aussi étroitement liée aux émissions de CO2 et de divers polluants des véhicules à moteur thermique. La pollution des véhicules est un problème mondial, mais il est d'autant plus épineux en Chine que ce pays est devenu depuis peu de temps le premier marché des véhicules du monde avec 28,08 millions d'immatriculations enregistrées en cours de la seule année 2018 [20]. Certaines métropoles chinoises, à l'instar de Beijing et de Shanghai, sont en proie à la pollution de l'air et à l'embouteillage. En 2018, il y a au total 6,08 millions de voitures immatriculées à Beijing [21].

Malgré la forte croissance des véhicules électriques, celles-ci ne représentent que 2,6% de la vente totale des véhicules neufs en 2017 (28,87 millions) [22]. Par ailleurs, étant donné que 58% de la production d'électricité est assurée par le charbon en Chine en 2017 [23], les véhicules électriques s'avèrent moins « verts »

17. Source d'information : Ministre de l'Écologie et l'Environnent de la Chine. http://www.xinhuanet.com/politics/2018-01/19/c_1122281477.htm. Consulté en février 2019.
18. Source d'information : https://www.sohu.com/a/235862337_617351. Consulté en janvier 2019.
19. Source d'information : Ministre de l'Écologie et l'Environnent de Chine.http://www.xinhuanet.com/politics/2018-01/19/c_1122281477.htm. Consulté en février 2019.
20. Chiffre de l'Association chinoise des constructeurs automobiles.
21. Source d'information : https://news.sina.com.cn/c/2019-03-21/doc-ihsxncvh4264005.shtml. Consulté en février 2019.
22. Source d'information : http://www.sohu.com/a/216250800_335308. Consulté en février 2019.
23. Source d'information : http://www.chyxx.com/industry/201804/627882.html. Consulté

que ce qu'on pourrait penser. De ce point de vue, l'émission des véhicules en Chine reste pour longtemps un obstacle à l'amélioration de la qualité de l'air, surtout dans les grandes villes.

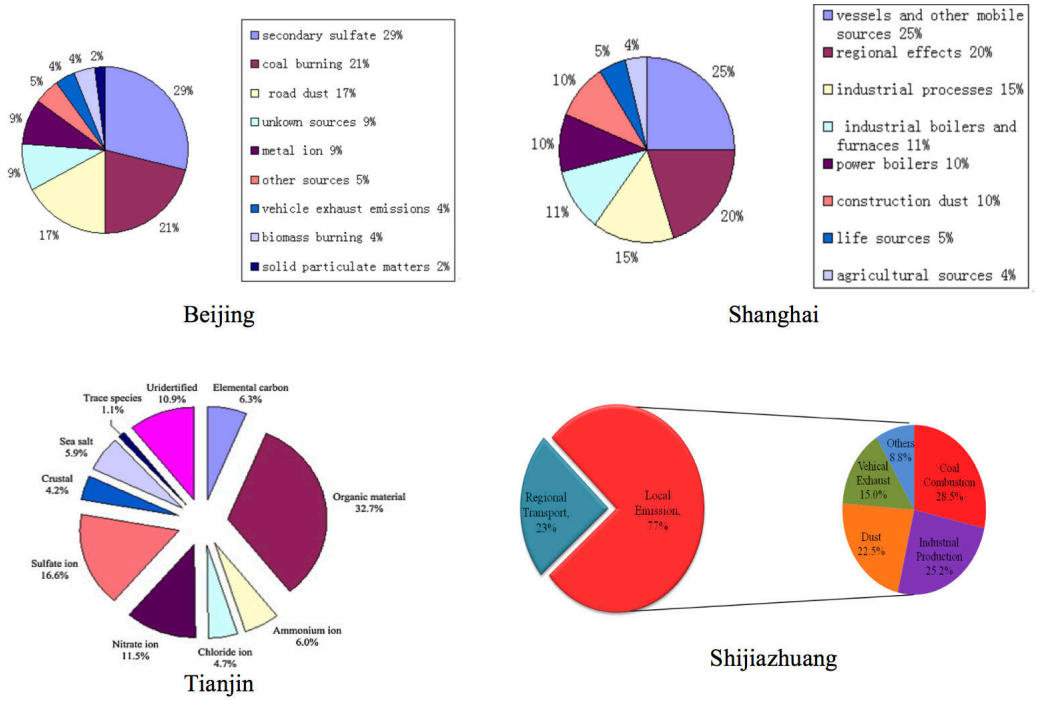


FIG. 1.3 – Composantes principales de PM2,5 dans quatre villes chinoises (Beijing, Tianjin, Shanghai et Shijiazhuang) [24]

1.4.2 *Impacts du wumai sur la santé*

Le brouillard de pollution a des impacts négatifs considérables pour l'environnement comme pour la santé publique. Avec le brouillard de pollution, on risque de subir plus de phénomènes climatiques extrêmes, car le *wumai* peut faire baisser la température de la surface de la Terre, tout en faisant augmenter celle de l'atmosphère [25]. En outre, dans la mesure où le brouillard de pollution présente

en février 2019.

24. Source des graphes : Les quatre graphes viennent tous de l'article « Characteristics of PM2.5 speciation in representative megacities and across China» (YANG et al., 2011).

25. Informations fournies par l'OMS. Source : https://www.who.int/fr/news-room/fact-sheets/detail/household-air-pollution-and-health. Consulté en février 2019.

un taux d'humidité relativement élevé, des bactéries et des virus peuvent y trouver un abri. En conséquence, quand on respire l'air pollué, on risque d'être infecté par ces bactéries et ces virus. Il faut encore remarquer que le brouillard de pollution contient des particules de métaux lourds. Ces particules sont tellement fines qu'elles pénètrent facilement dans le système de circulation sanguine via les poumons. La population sera ainsi davantage exposée aux maladies respiratoires, pulmonaires, cardiovasculaires, voire au risque de cancer [26], etc.

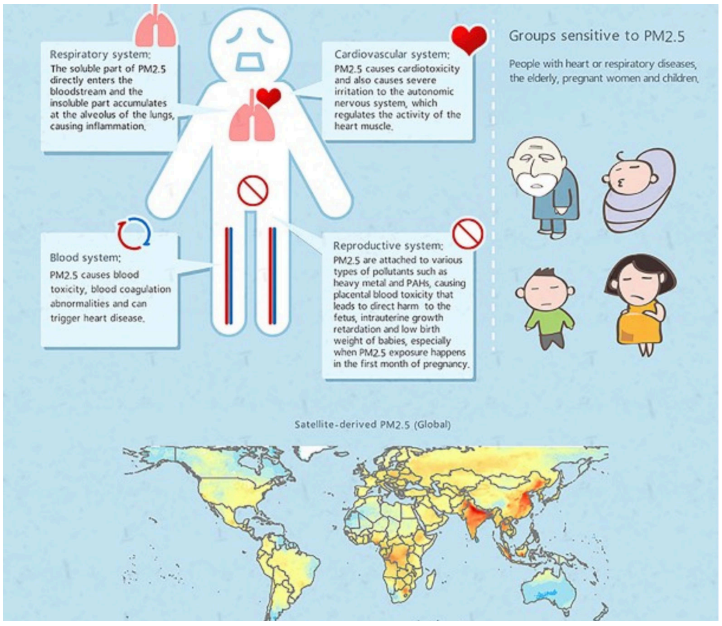


Fig. 1.4 – Maladies causées par le *wumai* et les populations les plus sensibles [27]

La pollution atmosphérique cause chaque année des millions de morts dans le monde entier. En 2013, on compte 5,5 millions de décès prématurés attribuables à la pollution atmosphérique ainsi qu'aux diverses maladies engendrées par celle-ci. La Chine et l'Inde sont les pays les plus touchés par ce problème environnemental,

26. « L'Organisation mondiale de la Santé (OMS) estime qu'environ 58% des décès prématurés liés à la pollution de l'air extérieur résultaient en 2016 de cardiopathies ischémiques et d'accidents vasculaires cérébraux, 18% de bronchopneumopathies chroniques obstructives ou d'infections aiguës des voies respiratoires inférieures, tandis que les 6% restants sont imputables au cancer du poumon. fg. Extrait d'un article publié par l'Organisation mondiale de la santé. Source : https://www.who.int/fr/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.. Consulté en février 2019.

27. Source : http://www.greenpeace.org/eastasia/campaigns/air-pollution/problems/coal-hard-truth-air-pollution/.

comme le montre la figure 1.5 Décès dus à la pollution de l'air en 2013 [28] :

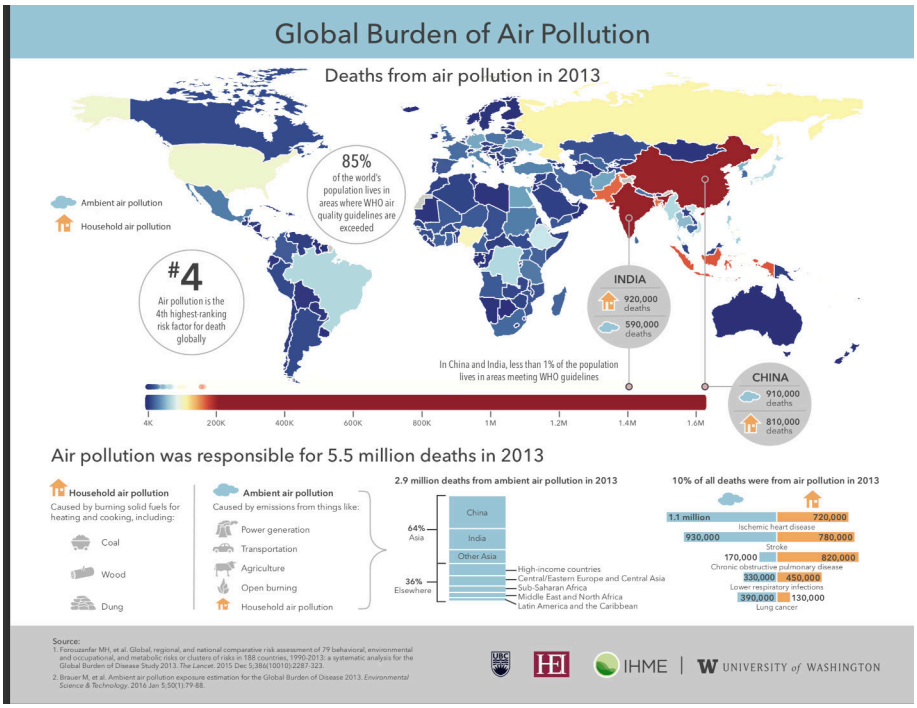


FIG. 1.5 – Décès dus à la pollution de l'air en 2013 [29]

1.4.2.1 Types de problèmes de santé exposés dans le corpus

À l'issue des analyses menées sur le sous-corpus WEIBO et détaillées dans notre article intitulé «« Using smog-related data of Chinese Sina Weibo to explore correlation between health issues and relevant regions»» (DANG, TURENNE et VALETTE, 2018), nous avons relevé huit types de problèmes de santé présents dans les grandes métropoles chinoises.

1. **Maladies pulmonaires** : 肺病 (maladie pulmonaire), 肺炎 (pneumonie), 尘肺病 (pneumoconiosis), 肺癌 cancer pulmonaire, 伤肺 (lésion des poumons), 肺心病 (cardiopathie pulmonaire) ;
2. **Maladies respiratoires** : 呼吸困难 (l'essoufflement), 支气管炎 (la bronchite), 气管炎 (la bronchite), 哮喘病 (l'asthme), 刺鼻 (irritation du nez) ;
3. **Cancer** : 癌症 (cancer), 肺癌 (cancer pulmonaire) ;

29. Source d'information : www.ncbi.nlm.nih.gov/pmc/articles/PMC4043337/.

4. **Inflammation** : 鼻炎 (rhinite), 炎症 (inflammation), 皮炎 (dermite), 咽喉炎 (inflammation de la gorge), 咽炎 (pharyngite), 发炎 (inflammation) ;
5. **Maladies cardio-cérébro-vasculaires** : 心脏病 (cardiopathie), 脑溢血 (hémorragie cérébrale), 脑梗塞 (Accident vasculaire cérébral), 脑血栓 (thrombus), 心脑血管病 (maladies cardio-cérébro-vasculaires) ;
6. **Maladies quotidiennes** : 感冒 (rhume), 流感 (grippe), 头痛 (mal à la tête), 发烧 (fièvre), 咳嗽 (toux), 咽痛 (pharyngalgie), 咳痰 (toux grasse), 鼻塞 (nez bouché), 打喷嚏 (sternutation) ;
7. **Maladies psychologiques** : 心理健康 (santé mentale), 心理压力 (pressions psychologiques), 心理障碍 (problèmes psychologiques) ;
8. **Autres maladies** : 佝偻病 (rachitisme), 眼病 (ophtalmo-pathie), 皮肤病 (dermatose), 贫血 (anémie).

Les principaux polluants primaires d'origine automobile

Nom	Symbole chimique ou acronyme	Origine	Dommages
Monoxyde de carbone	CO	Combustion incomplète des carburants	Troubles respiratoires et cardiovasculaires (réduction de la concentration d'oxygène fournie à l'organisme).
Oxydes d'azote	NO_x	Combustion de carburants	Dioxyde d'azote : troubles respiratoires, désagréments oculaires.
Particules fines (de taille inférieure à 10 Conseil général des Ponts et Chaussées)	PM 10	Véhicules (en particulier équipés d'un moteur diesel)	Troubles respiratoires et cardiovasculaires. Les personnes âgées, les enfants et les personnes souffrant de pathologies pulmonaires ou cardiovasculaires chroniques sont particulièrement sensibles aux particules. Des travaux ont montré qu'il existe des interactions entre particules d'origine diesel (PD) et pneumallergènes.
Composés organiques volatils (dont les hydrocarbures)	COV (hydrocarbures : HC)	Evaporation de l'essence et combustion incomplète	Certains sont nocifs (le benzène serait cancérigène)
Dioxyde de soufre	SO_2	Combustion de carburants soufrés	Troubles respiratoires et cardiovasculaires Pluies acides (acide sulfurique, H_2SO_4)

FIG. 1.6 – Relation entre les agents nuisibles et les types de maladies causés [30]

30. Source d'information : https://www.senat.fr/rap/r01-113/r01-1132.html. Consulté

1.4.2.2 Distribution régionale de problèmes de santé en Chine

À travers le graphe 1.7 Distribution des maladies par région en Chine dans le sous-corpus WEIBO, nous pouvons observer que les problèmes de santé se concentrent soit dans les régions qui subissent le plus le *wumai* : Beijing, Hebei, Tianjin, Shandong, Shaanxi, Heilongjiang ; soit dans les régions qui occupent une place importante au niveau de l'économie et de la population, telles que la ville de Shanghai et la province du Jiangsu et du Zhejiang. Et ces régions sont caractérisées par des types de maladies distincts :

- **Beijing** : La population de Beijing est particulièrement affectée par : 1) le cancer 癌症 et 2) les maladies respiratoires (bronchite 支气管炎, toux 咳嗽, inflammation des voies respiratoires 上呼吸道感染, sternutation 打喷嚏, irritation du nez 刺鼻) ;
- **Hebei** : La population de Hebei est significativement touchée par le cancer (癌症 (cancer), 肺癌 (cancer du poumon)) et par la cardiopathie 心脏病 (cardiopathie) , et la cardiopathie ;
- **Tianjin** : Les maladies les plus saillantes de Tianjin sont les maladies ou les symptômes fréquents, comme l'allergie : 过敏, mal de tête : 头痛 et la rhinite : 耳鼻喉 (ENT), 刺鼻 (irritation du nez) ;
- **Shanghai** : Les maladies respiratoires : 哮喘病 (asthme), 咳嗽 (toux), la cardiopathie : 心脏病 et les maladies pulmonaires : 肺炎 (pneumonie) constituent les maladies spécifiques de Shanghai.

en octobre 2019.

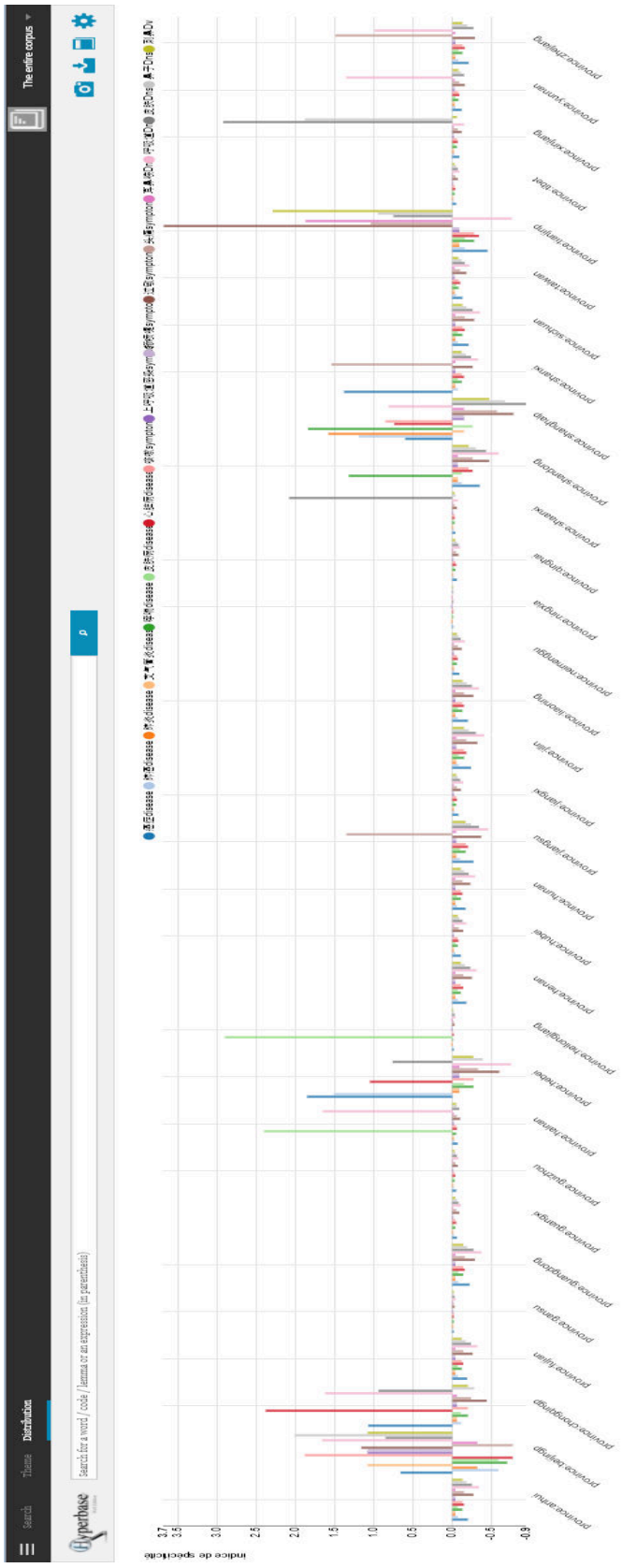


FIG. 1.7 – Distribution des maladies par région en Chine dans le sous-corpus WEIBO

1.4.3 *Mesures préventives contre le wumai*

Dans les articles médiatiques et microblogs des réseaux sociaux, de multiples mesures préventives contre le brouillard de pollution sont proposées par le gouvernement et le peuple chinois. Nous allons présenter dans les parties suivantes les mesures préventives les plus discutées dans les articles de presse institutionnels et dans les messages publiés sur les réseaux sociaux.

1.4.3.1 Mesures préventives proposées par la population chinoise pour se protéger contre le *wumai*

Le brouillard de pollution préoccupe non seulement le gouvernement, mais aussi le grand public. Il suffit de lire les messages postés sur le réseau social chinois WEIBO pour s'en rendre compte. Après avoir parcouru le rapport annuel des utilisateurs de WEIBO [31], nous avons constaté que déjà en 2013, le mot 雾霾 (le brouillard de pollution) est apparu pour la première fois dans la liste des dix mots les plus cherchés ou discutés, quatre ans avant son apparition dans le sous-corpus institutionnel-médiatique.

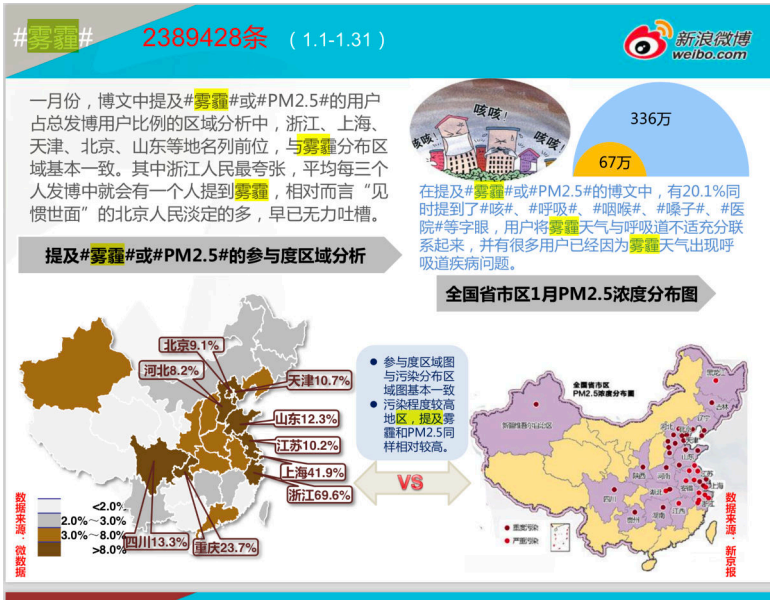


FIG. 1.8 – 雾霾 apparu dans la liste des mots les plus cherchés ou discutés sur WEIBO [32]

31. Il faut noter que WEIBO n'a commencé qu'à partir de 2012 à mettre en ligne son rapport annuel des utilisateurs.

En général, les mesures préventives les plus saillantes proposées ou adoptées par les utilisateurs dans les réseaux sociaux sont qualifiés « efficace » et visent à apprendre aux gens à se protéger en cas de *wumai*. Il s'agit soit de « porter un masque », soit d'« équiper la maison d'un purificateur d'air », ou encore de « recommander des recettes détox » pour humecter les poumons. Et les jours où le brouillard de pollution atteint un niveau très élevé, on propose de « 减少外出 »(réduire autant que possible les sorties à l'extérieur) ou « 减少开窗痛风 » (ne pas ouvrir les fenêtres). Ces mesures sont diffusées soit par les comptes personnels dans leurs propres zones, soit par les comptes des associations ou des institutions dans les informations complémentaires de leurs messages sur la qualité de l'air (cf. Chapitre 5 Tableau 5.5.4 Études sémantiques du thème 1 dans le sous-corpus Profane).

1.4.3.2 Réforme politique « à long terme » proposée par le gouvernement chinois pour régulariser le *wumai*

Au cours du XIII[e] Plan quinquennal de développement économique et social, 雾霾 (le brouillard de pollution) est apparu en 2017 dans le « 2017 政府工作报告 » (*Rapport du travail du gouvernement chinois en 2017*) en tant que mot-clé parmi les 14 mots de hautes fréquences sélectionnées par le gouvernement [33]. La même année, le sujet « régulariser le brouillard de pollution » a été sélectionné comme *top topic* des médias pendant les Deux Sessions (La conférence consultative politique chinoise et du Congrès national du peuple) [34]. Nous avons remarqué que, dans le rapport du gouvernement ainsi que dans d'autres articles institutionnels, les mesures proposées et discutées par les institutions sont incorporées dans des réformes politiques et économiques à long terme. L'attention est portée sur la source de la pollution de l'air beaucoup plus que sur des recommandations pratiques pour se protéger en cas de *wumai*. Ces mesures politiquent visent à : accélérer la transition énergétique (加快能源结构调整) : [D], utiliser

32. Source : « 2013 年微博用户发展报告 » (Le rapport annuel des utilisateurs de WEIBO en 2013).

33. Source d'information : http://www.gov.cn/xinwen/2017-03/06/content_5173852.htm. Consulté en février 2019.

34. Le terme « Deux Sessions » est la forme courte du mot chinois 两会, qui désigne les sessions plénières annuelles de l'Assemblée populaire nationale (APN) ou locale, et du comté national ou local de la Conférence consultative politique du Peuple chinois (CCPPC). Source d'inforamation : http://www.chinanews.com/gn/z/lh2017/index.shtml. Consulté en février 2019.

davantage les énergies vertes et renouvelables (使用清洁/再生能源) : [D] ; économiser l'énergie et réduire les émissions de CO2 (节能减排) : [A], [B] ; mener des réformes structurelles, et changer de mode de développement (加快产业结构调整, 产业模式升级) : [D], mettre en œuvre des dispositifs administratifs et judiciaires (健全行政和司法系统) : [C], etc.

A. La mise en place de la politique de circulation alternée. Par exemple, dans la ville de Beijing, cette politique est entrée en vigueur le 20 juillet 2008[35]. Ainsi, en cas de grave pollution, seule une partie des voitures peuvent circuler dans les rues de Beijing, selon que le dernier chiffre de l'immatriculation est pair ou impair. Depuis, cette politique s'est généralisée dans les principales villes chinoises, telles que Guangzhou et Shanghai ;

B. L'encouragement de l'usage des transports en commun ou des véhicules électriques. Dans certaines grandes villes chinoises, les véhicules électriques ne sont pas soumis aux contraintes de la circulation alternée ; les acheteurs de ces véhicules peuvent bénéficier d'avantages fiscaux. De plus, le gouvernement (central et local) subventionne considérablement l'achat d'un véhicule électrique. Avec ces mesures de soutien, la vente des véhicules électriques a fait un bond en avant pour atteindre les 770 000 unités en 2017[36]. Ceci fait de la Chine le premier marché des véhicules électriques.

C. Réforme de la loi environnementale. Auparavant, il suffisait aux usines polluantes de payer une certaine somme d'argent pour pouvoir continuer à tourner à plein régime. Maintenant, elles sont obligées de prendre des mesures réelles et rigoureuses pour réduire leurs émissions et être en règle. Dans ce contexte, les usines qui n'arrivent pas à satisfaire aux exigences environnementales sont menacées de fermeture. Ainsi, la ville de Dongguan, située au sud-est de Canton et connue pour son industrie manufacturière légère, a fermé plus de 4000 usines polluantes en sept mois au cours de 2018[37].

35. Source d'information : http://www.gov.cn/gzdt/2008-06/20/content_1022215.htm. Consulté en février 2019.
36. Agence Xinhua : http://www.xinhuanet.com/auto/2018-04/17/c_1122692994.htm. Consulté en février 2019.
37. Source d'information : http://www.chinadevelopment.com.cn/news/zj/2018/07/1320872.shtml. Consulté en février 2019.

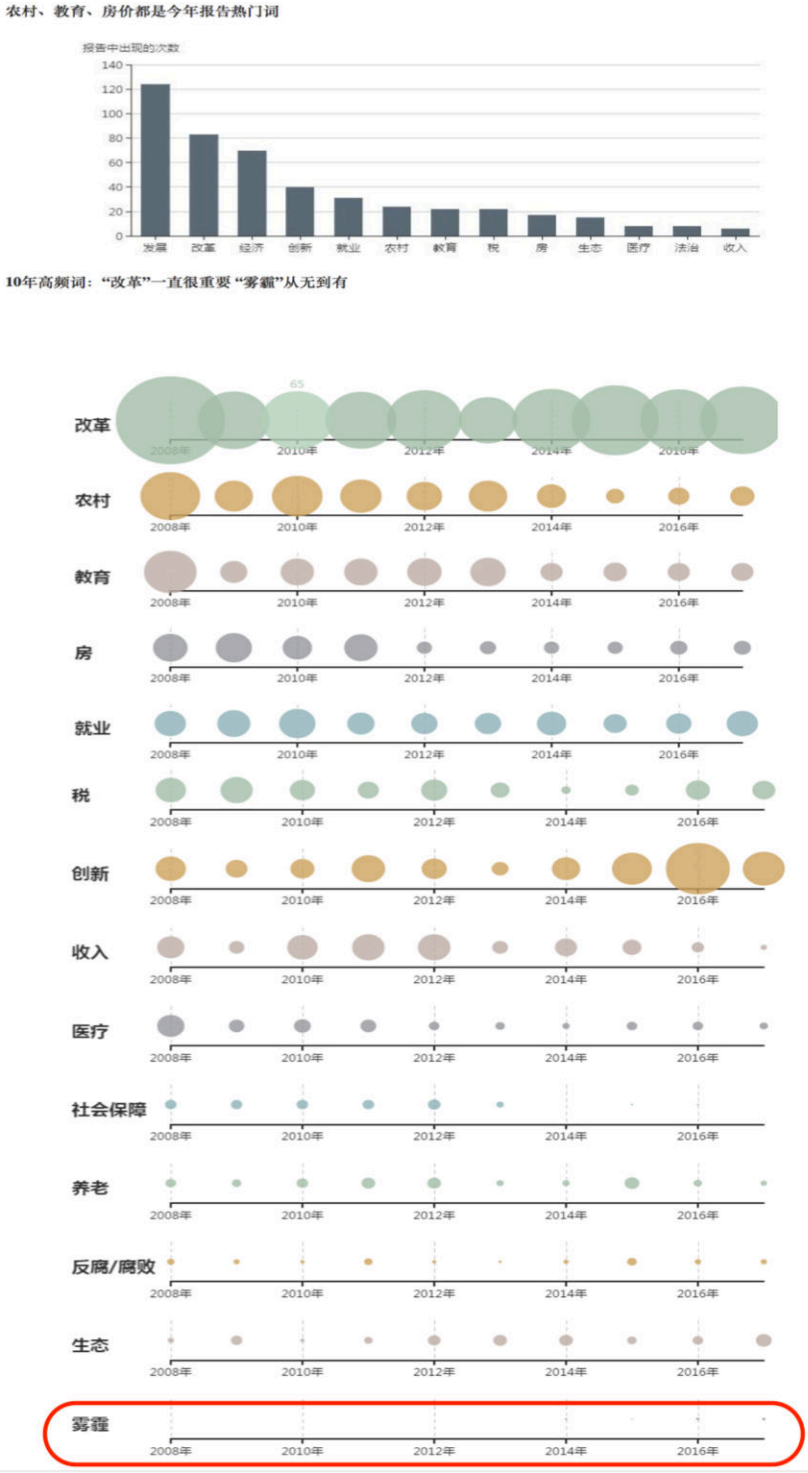


FIG. 1.9 – 雾霾 est parmi la liste des mots-clés du Rapport annuel du travail du gouvernement chinois [38]

D. La politique de « réduction de l'utilisation du charbon au profit du gaz naturel » (煤改气 en chinois)[39]. On incite les habitants à abandonner le charbon au profit du gaz naturel pour se chauffer ou faire la cuisine[40]. Cette politique s'inscrit dans une stratégie plus générale de modification la structure de la consommation d'énergie, qui consiste à augmenter la part des énergies renouvelables (éolienne, solaire, hydraulique), de l'énergie nucléaire, tout en baissant celle des énergies fossiles (surtout l'utilisation du charbon). Le gaz naturel, même s'il fait partie des énergies fossiles, est en revanche soutenu par le gouvernement chinois qui a décidé d'importer davantage de cette énergie et de moderniser l'infrastructure pour se préparer à cette augmentation de l'importation.

1.5 La censure d'Internet en Chine

Quand on étudie les textes publiés sur internet en Chine, la question de la censure est non négligeable. En Chine, tout texte publié dans le journal, tout livre, tout film ou feuilleton, est supposé être soumis à la censure ; ce processus aboutit à l'autorisation ou l'interdiction de leur diffusion totale ou partielle. Les messages postés sur les réseaux sociaux, les textes publiés sur internet ne font naturellement pas exception à cette règle. Il y a même une liste des mots-clés établie par les autorités permettant de bloquer la recherche en ligne sur internet. La censure y est une réalité quotidienne, avec une suppression de messages estimée à environ 16 %, allant jusqu'à plus de 50 % dans certaines provinces instables politiquement comme Ningxia ou le Tibet, contre seulement 12 % à Beijing (Bamman, O'Connor et Smith, 2012). En général, les sujets de la vie quotidienne ne rentrent pas dans la sphère de contrôle dans la mesure où on reste dans des généralités. Toutefois, le 28 février 2015, une présentatrice chinoise - 柴静 Chai Jing - a diffusé sur internet un documentaire « 穹顶之下 (sous le

38. Source du graphe : « 2017 政府工作报告 » (Le Rapport annuel du travail du gouvernement 2017).

39. La Banque asiatique d'Investissement dans les Infrastructures a investi 250 millions de dollars pour soutenir le programme de « réduction de l'utilisation du charbon au profit du gaz naturel » mené dans la région de Beijing. Source d'information : http://news.sina.com.cn/o/2017-12-12/doc-ifypnyqi4239575.shtml. Consulté en février 2019.

40. Source d'information : http://www.bjepb.gov.cn/bjhrb/xxgk/fgwj/qtwj/hbjfw/815239/index.html. Consulté en février 2019.

ciel)» dont l'enquête a été réalisée par elle-même[41]. Ce documentaire porte sur la pollution atmosphérique en Chine et a été vu plus de cent millions de fois en 24 heures après sa mise en ligne[42], preuve que ce documentaire a bien réussi à sensibiliser le grand public aux problèmes de pollution d'air. Mais cette réussite médiatique a vite tourné à la censure. Ainsi, une semaine après la diffusion, le documentaire s'est vu supprimé des sites internet, et ce, malgré le soutien du ministre de l'Environnement chinois à l'époque — 陈吉宁 Chen Jining — au début de la diffusion. De ce changement radical dans l'attitude du gouvernement chinois vis-à-vis de ce documentaire, nous pouvons apprendre deux choses : d'une part, la protection de l'environnement est un sujet moins sensible par rapport aux autres sujets typiquement tabous, sinon, ce documentaire n'aurait pas été autorisé à être diffusé ; d'autre part, si la protection environnementale n'est pas un sujet tabou, le gouvernement chinois exige toutefois que les discussions qui y sont relatives soient encadrées, de manière à ce qu'il n'y ait pas d'atteinte à l'image du gouvernement. Dans le cas du documentaire « Sous le ciel », peu de temps après la diffusion, les discussions commençaient à se concentrer sur le modèle économique du pays (caractérisé par l'omniprésence de l'État dans les activités économiques), puis sur le mode de gouvernance. S'il n'existe pas de critères précis pour déterminer si un texte ou une vidéo est « inapproprié ou non-sécurisé », il semble cependant que les articles publiés sur le sujet du *wumai* bénéficient d'une permissivité variable.

Il est difficile de mesurer l'impact réel de la censure sur la construction du corpus du présent travail. Mais une chose est sûre, c'est que depuis la Chine, nous n'arrivons pas à accéder à tous les textes ou messages relatif au sujet *wumai* et publiés sur internet. Cela ne signifie pas pour autant que notre corpus n'est pas représentatif de ce qui est produits par les internautes mais de ce qui persiste après la censure. Ce que nous voulons souligner en évoquant la question de la censure, c'est qu'il faut toujours en tenir compte quand on interprète les résultats, qui ne reflètent sans doute que partiellement les opinions du public ou du gouvernement sur le brouillard de pollution.

41. Source d'information : http://video.sina.com.cn/view/249334232.html. Consulté en février 2019.

42. Source d'information : http://news.youth.cn/kj/201503/t20150301_6497831.htm. Consulté en février 2019.

Chapitre 2

Cadre théorique et méthodologique

2.1 Introduction

Dans la théorie de la sémantique interprétative (SI), nous distinguerons trois objectifs correspondant à nos besoins : 1) au palier macrosémantique, la catégorisation du genre textuel par des caractéristiques infratextuelles et intratextuelles du corpus ; 2) au palier mésosémantique, l'identification des thèmes à partir du sème et de l'isotopie ; 3) au palier microsémantique, l'étude de l'interprétation sémantique en tenant compte de la pratique sociale. Notre travail s'inscrit dans le cadre de l'analyse sémantique de corpus qui s'appuie sur la SI et des procédés outillés textométriques.

Dans ce chapitre, nous faisons un usage opportuniste de la SI en exposant certaines propositions essentielles qui sont adéquates à notre recherche. Nous en présenterons les concepts fondamentaux en suivant les trois paliers sémantiques (cf. section 2.3 Trois Paliers sémantiques) : macrosémantique, mésosémantique et microsémantique. Nous commencerons par expliquer globalement nos objets d'étude pour chaque palier ; puis, nous aborderons d'autres concepts élémentaires relatifs tels que 1) la conceptualisation du genre textuel (cf. section 2.3.1 Palier macrosémantique : analyse du genre textuel) et sa caractérisation à partir de la co-interaction des quatre composantes sémantiques : dialogique, dialectique, tactique et thématique (cf. section 2.3.1.1 Quatre composantes sémantiques) ; 2) les notions théoriques de sème, d'isotopie (cf. section 2.3.2.1 Sème et Isotopie) et de thème (cf. section 2.3.2.2 Thème) ; 3) la démarche analytique en vue d'identifier une isotopie ou un thème dans le corpus. La présentation de ces concepts élémentaires est suivie de quelques travaux intéressants effectués par d'autres chercheurs qui sont inspirés de la SI (cf. sections 2.3.1.2 Travaux sur l'étude du genre textuel et 2.3.2.4 Travaux d'étude de l'isotopie ou du thème). À la fin de

ce chapitre, nous introduirons, de manière conjointe, l'approche quantitative de la textométrie (cf. section 2.4.1 Textométrie) ainsi que les outils techniques (cf. section 2.4.3 Outils textométriques) qui permettent de réaliser et illustrer nos analyses qualitatives.

2.2 La Sémantique Interprétative

Fondée par François Rastier dans les années 80, la SI a pour objectif principal d'interpréter la sémantique dans les textes (corpus). Développée dans le sillage de la sémantique structurale, la SI décompose le sens d'un texte sous forme de diverses unités sémantiques : du « local » sur le *sème* jusqu'au « global » sur le *genre textuel*, en passant par un « zone intermédiaire » — l'isotopie et le thème. Selon Rastier, ces trois catégories de l'unité sémantique sont axées respectivement sur trois paliers de description, qui sont lexicalisés et incarnés sous forme de mot, de syntagmes/phrases et des textes. La SI prend les textes (corpus) comme objet d'études, et définit le concept fondamental suivant : *le global détermine le local*, c'est-à-dire que l'interprétation sémantique des textes est contextuelle, le sens des textes étant déterminé d'abord par son genre textuel au niveau global, puis par les sèmes, les isotopies et les thèmes. Par ailleurs, selon Rastier, la sémantique du texte est constitué de quatre facteurs majeurs : des acteurs, un but, un sujet et la linéarité (Rastier, 2002). La sémantique textuelle procède ainsi de la co-interaction des quatre composantes sémantiques relatives : la dialogique, la dialectique, le thématique et la tactique, dont relève aussi le genre textuel.

L'objectif principal de la théorie est d'interpréter la sémantique dans les textes (corpus), et d'extraire la cohésion entre les éléments locaux et globaux. Le schéma 2.1 Structure de la Sémantique Interprétative [1] suivant montre les concepts fondamentaux dans la théorie de la SI et présente les relations générales qu'entretiennent les unités sémantiques.

2.3 Trois Paliers sémantiques

« Un texte peut être analysé à trois paliers : micro-, méso-, et macrosémantique, qui correspondent au mot, au syntagme/phrase et au texte » (Rastier, 1995). En suivant le principe fondamental «*le global détermine le local*», les trois paliers

2. Note : COOC signifie les cooccurrents.

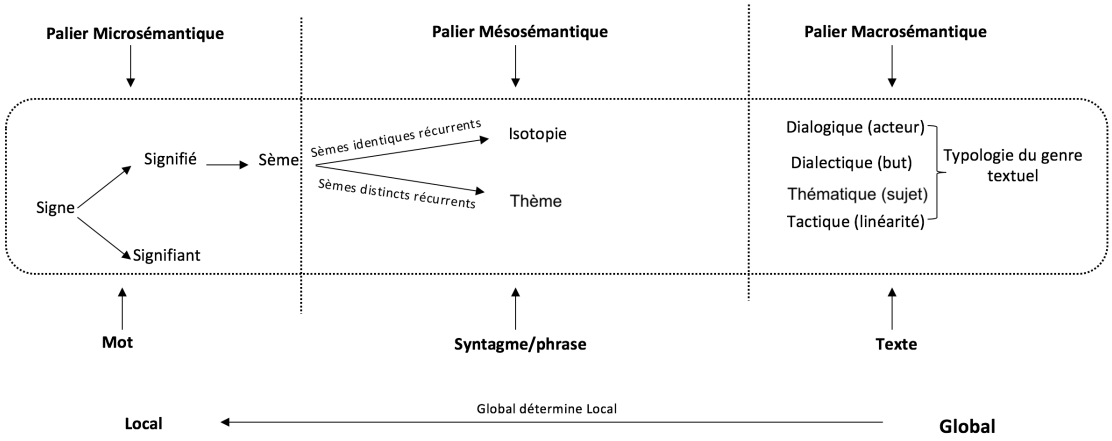


FIG. 2.1 – Structure de la Sémantique Interprétative [2]

sémantiques relient l'étude du sème à celle du thème, le mot au texte, le genre textuel à la sémantique textuelle. Nous allons détailler dans les parties suivantes 1) ce que nous étudions au palier macro- et méso- et microsémantique 2) le fonctionnement coordonné des trois paliers et des quatre composantes sémantiques mentionnées plus haut (dialogique, dialectique, tactique et thématique). Cette interaction permet de caractériser le genre textuel et d'interpréter la sémantique du thème. Chaque présentation des concepts théoriques est suivie et enrichie soit des exemples concrets de notre propre travail, soit des travaux effectués par d'autres chercheurs dans le domaine.

2.3.1 *Palier macrosémantique : analyse du genre textuel*

Il n'existe pas de langue générale, mais des pratiques sociales variées (la science, la politique, le journalisme, etc.) (RASTIER, 2001). Quatre niveaux hiérarchiques supérieurs au texte ont été définis par RASTIER : « les discours, les champs génériques (roman, récit, nouvelle, etc.), les genres et les sous-genres » (cf. figure 2.2 Niveaux de classification (RASTIER, 2001)). A cause de la structure hétérogène du discours, c'est le genre qui caractérise la textualité. Le genre constitue en fait une médiation qui relie les normes sociales et les actions individuelles. Ainsi, avant de procéder à l'interprétation, il faut préalablement prendre en compte le genre textuel auquel s'attache chaque sous-corpus.

Au niveau macrosémantique, nous étudions l'ensemble des indices linguistiques

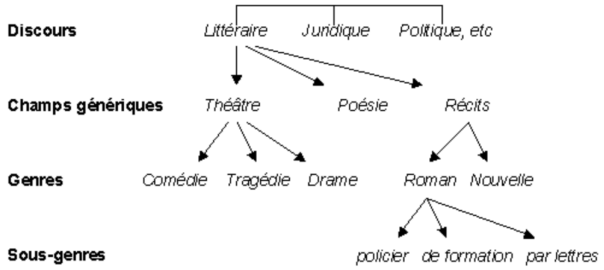


FIG. 2.2 – Niveaux de classification (RASTIER, 2001)

qui permettent de relever les caractéristiques de la typologie de quatre genres de sous-corpus (Ins, InsM, InfM et Profane). Ces caractéristiques se manifestent à deux niveaux — infratextuel et intratexuel — que nous détaillerons dans le Chapitre 4 section 4.3 Choix des variables.

Selon le concept principal de la théorie rastiérienne « Chaque texte procède d'un genre, et chaque genre est relatif aux discours déterminés (politique, religieux, scientifique, littéraire, etc.)»(RASTIER, 1996). L'étude de la sémantique des genres se caractérise par « les modes de co-variation entre un faisceau de critères des composantes sémantiques rattachées aux pratiques et cultures sociales » (RASTIER, 2001). Ces composantes sémantiques se manifestent par le biais de variables lexicales, sémiotiques, rhétoriques, modales, syntaxiques, etc[3].

> « Pour une sémantique des genres, ce qu'on cherche à faire, c'est de caractériser les modes des co-variation entre un faisceau de critères des composantes sémantiques : thématique ouverte/fermée, concentrée/diffuse ; dialectique ordonnée, désordonnées, impertinente, orientée positivement ou négativement ; dialogique variant ou non les foyers de l'énonciation et de l'interprétation représentées ; tactique pertinente ou non pertinente, etc. Un genre est défini comme un mode d'interaction normé entre composantes » (RASTIER, 2001).

3. « Dans l'analyse sémantique, on doit être restreint à bon escient pour caractériser la spécificité des discours et des genres, les thèmes du roman ne sont pas ceux de l'essai ni du poème » (RASTIER, 1995).

2.3.1.1 Quatre composantes sémantiques

Puisqu'un texte est un alignement de lexiques suivant des règles de construction syntaxique, il consiste en une « suite linguistique autonome (oral ou écrite) constituant une unité empirique, et produite par un ou plusieurs énonciateurs dans la pratique sociale attestée[4] ». Les unités sémantiques peuvent faire l'objet de diverses descriptions selon les composantes. À chaque composante correspondent des types d'opérations productives et interprétatives. Nous détaillons *infra* les proposition de Rastier (2001, 1987) relatives aux quatre composantes et à leur relation corrélative, et les illustrons grâce à des exemples concrets extraits de notre corpus.

La **thématique** : Ensemble, système organisé des thèmes[5]. Autrement dit le sujet d'un texte. La thématique est déterminée par des contenus investis par des unités récurrentes sémantiques structurées (Rastier, 2001). Selon Rastier (1995), un des buts de la thématique est de repérer les regroupements de thèmes par ses récurrences. En ce sens, l'étude thématique rejoint donc l'étude des thèmes que ce soit au niveau de la méthode ou de l'objectif (cf. section 2.3.2.2 Thème).

La **dialectique** : met en jeu des acteurs, des rôles et des fonctions du récit. Elle assume son rôle d'argumentation dédiée aux éléments descriptifs ou argumentatifs. Les structures dialectiques des textes dénotent le type de l'évolution temporelle, le déroulement aspectuel et l'évolution modale mise en œuvre dans le texte(Rastier, 2001). « Le vocabulaire caractéristique dialectique est assez varié[6], il peut s'agir des marqueurs de structuration (enfin, donc, cependant), des verbes modaux (devoir, falloir, vouloir) et des indicateurs rhétoriques (emphase, point d'interrogation, mots interrogatifs)» (Eensoo et Valette,2015). Dans notre cas, pour rendre compte de comment sont organisés les intervalles temporels et le déroulement aspectuel des thèmes majeurs dans chaque genre de sous-corpus, nous avons repéré dans notre corpus institutionnel des vocabulaires caractéristiques de la composante dialectique. Nous avons par exemple relevé

4. Définition de texte, cf. http://www.cnrtl.fr/definition/texte

5. cf. La définition de thématique dans le dictionnaire TLFi http://stella.atilf.fr/Dendien/scripts/tlfiv5/advanced.exe?8;s=3175227615. Thématique d'un auteur, d'un genre, etc.

6. Eensoo et Valette interprètent librement le concept de dialectique dans leur article « Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité».

les verbes de modalité : 应该 (falloir), 必需 (devoir), 需要 (vouloir) et des indicateurs emphatiques : 加强 (renforcer), 强调 (accentuer), des occurrences de parallélisme[7] pour des expressions de l'ordre (voir Chapitre 4 section 4.6 Verbes et adverbes modaux dans les quatre genres textuels selon l'indice de spécificité et 4.6.5.4 Parallélisme), ainsi que des points d'interrogation, mots interrogatifs (cf. section 4.6.3 Variables sémiotiques du Chapitre 4) et des connecteurs logiques, tels que 终于 (enfin), 因此 (donc), 但是 (mais), 然而 (cependant) (cf. section 4.6.2.2 Analyse des conjonctions du Chapitre 4).

La **dialogique** : La dialogique fonde la typologie des énonciateurs représentés, et rend compte des modalités énonciatives et évaluatives. Rapportées à la thématique, et à la dialectique, les variations dialogiques introduisent des différences au niveau des acteurs et de leurs positionnements énonciatifs avec l'actualisation des pronoms personnels et de certaines entités nommées (nom de personne, nom d'établissement, titre professionnel, nom d'État, etc.). Par exemple, les textes émis par les institutions contiennent de nombreuses phrases impératives et nous y observons un usage plus important du pronom « on » (cf. section 4.6.2.5 Analyse des pronoms personnels). En comparaison, les textes médiatiques et profanes multiplient les énonciateurs, souvent introduits par des noms de personnes (cf. section 4.6.2.7 Analyse des noms du Chapitre 4) et repris à travers de nombreux pronoms : « elle », « ils », « vous », etc. (voir section 4.6.2.5 Analyse des pronoms personnels du Chapitre 4). Ces textes possèdent donc un foyer énonciatif et un foyer interprétatif non nommés.

La **tactique** : La tactique permet de définir des rythmes sémantiques. Ces rythmes, (non) séquentiels ou (non) linéaires, marquent l'ordre de la production et l'interprétation des unités sémantiques. Nous nous servons de ce concept pour comparer et relever, au niveau infratextuel, les caractéristiques de la disposition des pages des sites web contenant les textes qui composent notre corpus, pour chacun des genres étudiés.

Jusqu'ici, nous avons exposé ce que l'on étudie (le genre textuel) au palier macrosémantique et sur quels critères (les quatre composantes sémantiques) sont recherchées les caractéristiques du genre textuel. Dans la partie suivante, nous

7. En chinois, le parallélisme est une figure de rhétorique, qui permet de mettre en parallèle une paire ou une série de mots similaires, ou encore des expressions ou des phrases apparentées. Par exemple : l'expression 低碳生活, 绿色出行 (vie à charbon bas, voyage vert) contient deux locutions nominales, dont chacune est constituée de quatre caractères. Ces quatre caractères forment deux mots : un adjectif plus un verbe nominalisé.

allons présenter les travaux effectués par d'autres chercheurs qui ont mis en pratique directement (ou indirectement) la théorie de Rastier sur l'analyse du genre textuel, pour la caractérisation et la comparaison du genre textuel des corpus numériques. À partir des jeux de variables lexicaux/sémiotiques reposant sur les quatre composantes sémantiques, ces travaux ici recensés sont pour la plupart réalisés sur des corpus discursivement hétérogènes.

2.3.1.2 Travaux sur l'étude du genre textuel

Dans le cadre de la linguistique, les textes constituent l'objet d'étude, et les traits pertinents, qui viennent des textes eux-mêmes, permettent de différencier des groupements de textes. Les travaux de Biber (1992) sont intéressants pour étudier des genres. En se basant sur le principe que les genres préexistent aux textes, la démarche proposée par Biber permet d'examiner les caractéristiques de chaque genre, autrement dit les variables qui rapprochent ou qui éloignent les textes d'un corpus. Pour ce faire, Biber définit, en s'appuyant sur 481 textes écrits annotés, seize catégories de traits linguistiques discriminants, dans lesquelles il répartit 67 traits linguistiques, qui se rapportent à la dialectique et à la dialogique. Ces traits sont par exemple «*les marqueurs de temps et d'aspect, les adverbes et locutions adverbiales de temps et de lieu, les pronoms et proverbes, questions, passifs, modaux, coordination, négation*» , etc.

L'étude particulière de Beauvisage (2001) menée sur le roman policier nous permet de répondre partiellement aux problèmes méthodologiques soulevés par l'étude des genres textuels. Elle valide pleinement un travail fondé sur les variables morphosyntaxiques et sur la ponctuation. Son travail montre que ces éléments sont à même de donner, par contraste d'un autre genre dans le corpus,une représentation de la spécificité des genres.

L'interprétation se fait au niveau de la contextualisation, qui se décline en général sur deux niveaux principaux : l'intratextuel et l'intertextuel. POUDAT (2006) a mobilisé les calculs textométriques et certains concepts de la SI afin d'étudier le genre textuel des articles scientifiques. Dans ses expériences, elle combine l'analyse lexicale et la description morphosyntaxique sur les paliers infratextuels (les sections) et supratextuel (le style, le domaine, etc.).

Kessler, Nunberg et Schutze (1997) ont étudié la détection automatique des genres des documents issus du Web, et montré qu'un système automatique

est capable de reconnaître les genres. La reconnaissance des genres est réalisée en analysant la cooccurrence d'éléments de nature différente. Ils ont mis en question les variables hétérogènes associées à la structure de surface et profond, qui sont en mesure de détecter le genre textuel de ces types de documents. En définissant le genre comme « un principe de classement hétérogène au texte », à chaque classement sont attachés des traits discriminants spécifiques. Dans l'étude de Kessler, quatre types de traits discriminants relatifs à la dialectique et à la dialogique sont examinés de manière automatique :

- traits structuraux : passif, normalisation, la fréquence de chaque type de POS-tagging prédéfini ;
- traits lexicaux : abréviation de titre de civilité, par exemple *Mr.*, *Ms.* est prédominant dans le journal *New York Times* ; expressions de dates pour les histoires des actualités, par exemple ;
- traits des caractères : les ponctuations (valeur rythmique et syntaxique), les séparateurs, et des marqueurs de délimiteurs ;
- traits dérivatifs : des ratios et mesures dérivées des deux traits précédents.

Selon Bonhomme (2015), les genres permettent d'élaborer, de planifier et de repérer les activités verbales proposées sur Internet en fournissant des normes : comment produire un texte, comment assurer une fonction de représentation et comment interpréter un texte. L'étude des genres permet d'une part, de percevoir la singularité de l'organisation et de construction des énoncés des textes (rapporté à la tactique) relevée par les paramètres discriminants lexicaux et sémiotiques ; d'autre part, le genre est un moyen d'établir une liaison entre la linguistique et le social, en fonctionnant comme « modèle de production et d'interprétation » (Gonçalves, 2014).

Dans les travaux de Gonçalves (2014), l'auteur étudie les similitudes et différences textuelles dans les genres numériques. Ces textes numériques qui proviennent des blogs et des sites web sur un même sujet — le tourisme — sont mis en comparaison à partir des traits discriminants, qui sont en mesure de décrire et analyser les caractéristiques singulières et récurrentes de différents genres textuels. Les particularités relevées par Gonçalves dénotent que 1) sur le site web, au niveau de la tactique, la configuration globale fragmentée en différentes sections du genre numérique est séquentielle et ramifiée ; par rapport au site web, la structure du blog, qui semble aussi fragmentée en bloc, possède en fait une

structure linéaire et antéchronologique (du plus récent au moins récent) ; 2) les analyses discursives des deux types de genres numériques permettent de mettre en lumière la temporalité (rapportée à la dialectique) et la présence ou l'absence des marqueurs de personne (rapportée à la dialogique). Ayant pour objectif la communication et l'interaction avec les lecteurs/utilisateurs, le site web convoque l'ordre du *raconter*, en témoignent l'emplacement important du cadre de navigation, la structure concise de la page d'accueil, la présence de la 3ème personne du singulier, la conjonction entre la date de publication et la date de production des textes, etc. Le blog, quant à lui, est essentiellement de l'ordre de l'*exposer* (discours interactif). Cet aspect est mis en évidence par la disposition du cadre réservé aux commentaires et l'utilisation du présent de l'indicatif et de la 1ère personne. Le blog est également de l'ordre du *raconter* (discours narratif) où nous trouvons les indices suivants : a) verbes aux temps passés ; b) disjonction spatio-temporelle de la situation d'énonciation ; c) absence de la 1ère personne.

En tenant compte des caractéristiques des documents numériques, l'auteur a dans cette étude mis en avant cinq traits qui sont propres au genre numérique :

1. la non-linéarité : la flexibilité de l'organisation des textes numériques permise par l'insertion des liens hypertextes -> ramification ;
2. la volatilité : la possibilité de reformuler et de modifier ce qui est déjà sur le site Web ;
3. la plurisémiose : plusieurs variables sémiotiques (images, animations flash, schémas, vidéos, etc.) ;
4. l'utilisabilité : les textes numériques doivent être créés de manière à être efficaces et efficients, à satisfaire, à attirer les lecteurs ou les utilisateurs, voire à les fidéliser ;
5. l'interaction physique ou corporelle : au niveau de la construction du texte et de sa réception (cadre réservé aux commentaires des utilisateurs) ;
6. la multidimensionnalité : les sites web, notamment le blog, offrent des dispositifs plus ou moins sophistiqués et d'hypertextualité, par exemple le cadre de commentaire réservé aux lecteurs. Le blog constitue lui-même à la fois un réseau social et un espace d'expression personnelle.

Fløttum et al. (2014) reprennent certaines caractéristiques du genre numérique exposées par Gonçalves : la *multimodalité*, qui intègre le son, le texte, des images et des clips vidéos, etc. ; l'*hypertextualité*, qui permet la circulation entre

les pages et les sites web grâce aux hyperliens ; la *non-linéarité*, qui permet de naviguer entre des pages et des sites ; l'*interactivité*, qui permet la communication entre le producteur et les utilisateurs/lecteurs ; l'*hétérogénéité* des thèmes : sports, politique, technologie, économie,etc.et *l'ancrage dans un contexte socio-culturel*. Le blog est inventé dans les années 90, à une période caractérisée par la mise en question des frontières entre le domaine privé et le domaine public. Aujourd'hui, deux aspects du blog sont mis en avant : 1) la zone privée , personnelle, intime et interactive, actualisée par le dispositif « user-generated-content » [8] ; 2) le rôle politique et social dans la blogosphère contemporaine. Par ailleurs, l'auteur présente deux caractéristiques spécifiques au blog qui ne sont pas mentionnées par Gonçalves : la *rapidité* de la mise à jour des textes et le caractère *identitaire*, qui sert d'indices de communautés discursives et marque les sphères d'emploi privilégiées d'un groupe avec son point de vue, ses normes et son style.

Pour contraster le genre numérique du blog et celui du site d'institutions, Bonhomme (2015) a analysé les pages d'accueil des sites politiques. Les résultats obtenus mettent en évidence les caractéristiques du genre des sites politiques par le biais de diverses variables sémiotiques, lexicales et modales rapportées à la tactique (variables systématiques), à la dialogique et à la dialectique (variables communicationnelles) :

- **variables systématiques** :
  1. barre de navigation thématisée : position/votation/parti ;
  2. moyens de contacts avec les citoyens : Abonnez-vous au flux RSS ; Rejoignez-nous ; Devenir membre ; Gérer mes données ;
  3. l'actualité politique : la une, les nouvelles politiques, etc. ;
  4. titres et images en hyperlien : animation des images centrales.
- **variables compositionnelles** :
  1. rubriques, titres et chapeaux ;
  2. symboles visuels : logo (forte identité), dessins emblématiques de l'activité politique ;
  3. portrait dirigeant ;
  4. colorisation du background : couleur représentative des partis.
- **variables communicationnelles** :
  1. mots d'ordre [9] ;

8. Ce qui fait du texte la création non seulement de l'auteur mais aussi du lecteur
9. Cf. https://www.fileane.com/docpartie5/quelques_slogans.htm.

2. slogans ;
3. titres idéologiques : par exemple : Pour une Europe Sociale ;
4. hyperliens externes vers Facebook ou Tweeter : orientation moderniste axée sur les médias à la mode ;
5. impératif : Participez !
6. questions : Veux-tu faire bouger les choses ?
7. infinitifs incitatifs : Faire un don !

Après avoir étudié le genre numérique du blog et celui des sites traditionnels, l'auteur a constaté que « le blog offre une grande diversité architecturale, thématique et procédurale », et les genres du web traditionnel sont des hybrides réaménagés à partir des genres anciens.

Dans cette partie, nous avons défini le palier macrosémantique et le genre textuel ; nous avons présenté des travaux de chercheurs qui exposent des méthodes pour l'analyse du genre textuel, et des travaux spécifiques sur les genres du web, en rapport avec notre sujet. Dans la partie suivante, nous allons aborder les paliers micro- et mésosémantique et présenter les notions de sème, d'isotopie et de thème.

2.3.2 *Palier microsémantique et palier mésosémantique*

Le palier microsémantique vise à étudier la sémantique du palier inférieur du texte. L'étude s'appuie sur les unités minimales d'un signe — les sèmes. Rastier (2005) définit le palier mésosémantique comme suit : « La mésosémantique rend compte du palier intermédiaire entre la lexie et le texte. Elle traite donc de la phrase, ou plus précisément de l'espace qui s'étend du syntagme pourvu d'une fonction syntaxique jusqu'à la phrase complexe et à ses connexions immédiates.». Plus concrètement, au palier mésosémantique, on explore dans un texte son isotopie (récurrence d'un même sème) et ses thèmes (lexicalisation des faisceaux des sèmes différents).

L'interprétation sémantique du texte au niveau du palier micro- et mésosémantique repose ainsi sur le repérage et l'identification des fonds et des formes sémantiques. Les fonds sémantiques sont présentés par les groupes de sèmes récurrents qui construisent en fait une/plusieurs isotopies à l'intérieur. Et les formes sémantiques sont actualisées par les thèmes formés par un petit réseau sémantique, qui est établi de manière régulière, par des sèmes hétérogènes et des mots

isotopants, autrement dit des *cooccurrents*[10]. Ainsi les trois éléments, le sème, l'isotopie et le thème, sont liés de manière très étroits. C'est à partir d'un sème qu'on peut obtenir l'isotopie et le thème. Nous allons présenter ces trois concepts dans les sections suivantes.

2.3.2.1 Sème et Isotopie

Pottier (1985) définit le *sème* comme « la plus petite unité de signification définie par l'analyse ». À partir de la matrice des sèmes (cf. figure 2.3 Matrice des sèmes et sémèmes proposée par Bernard Pottier (1985) [11]), nous pouvons observer que la signification d'un mot est constituée par différents sèmes. Par exemple dans la matrice suivante, le mot « chaise » contient au total quatre sèmes : S1 /pour s'asseoir/, S2 /pour une personne/, S3 /avec dossier/ et S4 /avec bras/. Parmi les quatre sèmes, le S1 /pour s'asseoir/ [12] constitue le sème commun des quatre mots : « chaise », « fauteuil », « tabouret », « canapé ». Dans un texte, le phénomène que le même sème (en l'occurrence /pour s'asseoir/) apparaît de manière récurrente est considéré comme l'*isotopie* dans la théorie rastérienne de SI.

SÈME		pour s'asseoir S_1	pour une personne S_2	avec dossier S_3	avec bras S_4
SÉMÈME	chaise	+	+	+	-
	fauteuil	+	+	+	+
	tabouret	+	+	-	-
	canapé	+	-	+	ø

Fig. 2.3 – Matrice des sèmes et sémèmes proposée par Bernard Pottier (1985) [13]

Le concept de l'isotopie a été inventé par Greimas dans les années 70. Il était défini alors comme « un ensemble redondant de catégories sémantiques qui rend possible la lecture uniforme du récit, telle qu'elle résulte des lectures partielles

10. 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 formes données. (cf. André Salem, 1994)

12. Les conventions typographiques de la présentation des unités linguistiques en SI : « signe » pour **signe** ; /sème/ pour **sème** ; /isotopie/ pour **isotopie** 'sémème' pour **sémème** ; //classe sémantique// pour **classe sémantique**.

13. Source d'information : `http://www.linguistes.com/mots/lexique.html`.

des énoncés et de la résolution de leurs ambiguïtés qui est guidée par la recherche de la lecture unique ». Elle a ensuite été développée par Rastier comme un effet de la combinaison des sèmes récurrents dans le contexte textuel (Rastier, 1987). Ainsi Rastier (2009) définit une **isotopie** comme la répétition d'un sème donné qui apparaît de manière récurrente en contexte d'un texte. Par conséquent, l'isotopie « apparaît comme un principe régulateur fondamental ». Étant la particularité récurrente des sèmes homogènes, l'isotopie permet d'introduire, dans la suite de chaînes linguistiques (phrase, texte, corpus), une cohésion entre les mots. Et cette cohésion constitue le fond sémantique du texte (Rastier,1987). De manière générale, les analyses sémantiques d'un corpus textuel partent d'une hypothèse précaire pour identifier les isotopies du corpus. « L'identification des isotopies permet d'extraire le sens véhiculé dans le corpus dans un moment donné ou dans un certain zonage d'un texte : d'une part, le repérage des isotopies fait comprendre que la cohésion textuelle se manifeste par les points communs sémantiques à partir des indices isotopants lexicalisés par les occurrences des mots et des phrases ; de l'autre, nous pouvons constater les irrégularités isotopantes en fonction de l'évolution temporelle ou spatiale d'un texte ou d'un corpus » (Pincemin, 1999). Par exemple, malgré la variété du lexique employé —雾霾 (*brouillard de pollution*), 霾 (*smog*), 空气污染 (pollution de l'air), 大气污染 (pollution atmosphérique), PM2,5 (matière particulaire 2,5), PM10 (matière particulaire 10), ils partagent le même sème la /pollution de l'air/, qui constitue l'isotopie, autrement dit le fond sémantique dans l'ensemble de notre corpus.

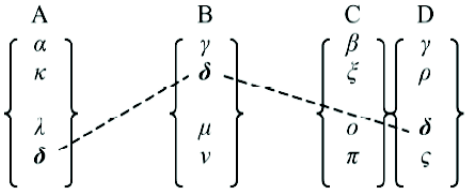


Fig. 2.4 – Exemple d'Isotopie dans un texte (Valette, 2009)

2.3.2.2 Thème

Le thème est défini par RASTIER(1995) comme « une structure stable de traits sémantiques, c'est à dire les sèmes, récurrents dans un corpus, et susceptibles de lexicalisations diverses ». VALETTE (2009) le présente quant à lui comme des « groupements instanciés par l'ensemble des sèmes distincts avec une certaine régularité ». Nous pouvons ainsi dire que le *thème* est formé par un petit réseau sémantique, qui est établi de manière récurrente, par des sèmes hétérogènes lexicalisés dans la suite de chaîne linguistique (phrase, texte, corpus). Pour identifier les thèmes principaux de notre corpus écologique relatif au brouillard de pollution, nous nous inspirons du procédé proposé par VALETTE (2009) dans sa recherche « Approche textuelle du lexique». Sa démarche analytique nous renseigne, avec des schémas (cf. figures 2.5 Repérer une isotopie dans un texte (graphe modifié à partir de son original proposé par VALETTE (2009) et 2.6 Repérer un thème dans un texte (graphe modifié à partir de son original proposé par VALETTE (2009))), comment identifier l'isotopie et le thème dans le corpus textuel. Plus concrètement, pour trouver d'autres sèmes distincts qui apparaissent de manière simultanée et récurrente avec l'isotopie /pollution de l'air/, nous procédons d'abord au calcul des cooccurrents des unités lexicales isotopantes (tels que 雾霾 (brouillard de pollution), 霾 (smog), 空气污染 (pollution de l'air), etc.) dans notre corpus ; puis, à partir des cooccurrents du mot-pivot, nous pouvons en déduire les sèmes qui sont correspondants. Prenons le mot isotopant 雾霾 (brouillard de pollution) comme exemple, dans la liste de ses cooccurrents calculés, nous obtenons finalement trois faisceaux de sèmes cooccurrents distincts : 1) sème /cause/(sc) présenté par les mots : « 原因 » (raison), « 来源 » (origine) et « 成因 » (cause) ; 2) sème /maladie/ (sm) sous forme des lexiques : « 肺炎 » (pneumonie), « 呼吸系统疾病 » (maladie respiratoire) ; 3) sème /mesure préventive/ (sp) lexicalisé par les mots : « 口罩 » (masque), « 治理 » (régulariser) et « 预防 » (prévenir). De ces trois faisceaux résultent les trois thèmes principaux de notre corpus : « les causes de la pollution de l'air » (sc+isotopie), « les impacts de la pollution de l'air sur la santé » (sm+isotopie) et « les mesures préventives de la pollution de l'air » (sp+isotopie) (voir les détails dans les sections 5.5 Études sémantiques du thème 1 : Causes de la pollution de l'air, 5.6 Études sémantiques du thème 2 : Impacts de la pollution de l'air sur la santé, et 5.7 Études sémantiques du thème 3 : Mesures préventives contre la pollution de

l'air du Chapitre 5).

2.3.2.3 Démarche de recherche des thèmes

En plus des concepts théoriques du thème, RASTIER propose aussi la démarche de recherche des thèmes assistée par l'outil technique. Cette démarche a été schématisée par VALETTE (2009) (voir les graphes 2.5 et 2.6) et adoptée par les linguistes dans leur travail de recherche (voir *infra* section 2.3.2.4 Travaux d'étude de l'isotopie ou du thème) :

Voici les grandes étapes de la démarche selon laquelle se sont déroulées nos études :

1. Choix des hypothèse : en fonction de l'objectif général de la recherche (une préanalyse statistique peut guider la recherche d'hypothèse, par exemple, la fréquence d'occurrences dans le corpus reste indispensable pour guider les intuitions) ;
2. Recherche de cooccurrents par la méthode statistique des écarts réduits ou hypergéométrique. Les cooccurrents sont employés en guise des indices potentiels d'isotopies afin de contraster sémantiquement les lexèmes *a priori* proches ;
3. Transformation interprétative des cooccurrents en corrélats, et constitution des réseaux thématiques (cette étape est facilitée si l'on pratique une interrogation simultanée sur plusieurs cooccurrents) ;
4. Validation des résultats, par croisement de l'analyse thématique avec l'analyse d'autres composants du même corpus, par test sur un corpus de contrôle, ou par confrontation avec d'autres recherches thématiques.

2.3.2.4 Travaux d'étude de l'isotopie ou du thème

Dans cette section, nous présentons quelques travaux menés dans des domaines variés mais s'appuyant sur les concepts théoriques de l'isotopie et du thème. Nos études sur l'analyse de thème s'inspirent de ces expériences.

Dans les travaux de VALETTE (2009), plusieurs exemples sont donnés pour explorer l'isotopie et le thème. Par exemple, l'isotopie /ville/ est récupérée par divers vocabulaires qui partagent le même sème /ville/, tels que « urbaine »,

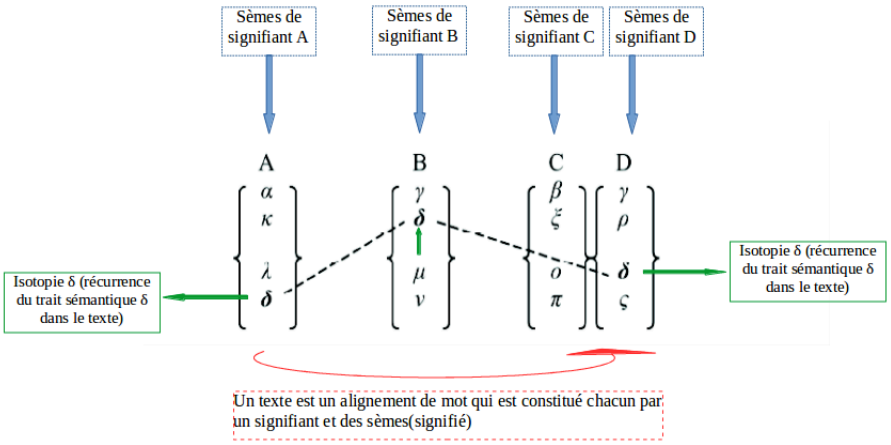


FIG. 2.5 – Repérer une isotopie dans un texte (graphe modifié à partir de son original proposé par VALETTE (2009)

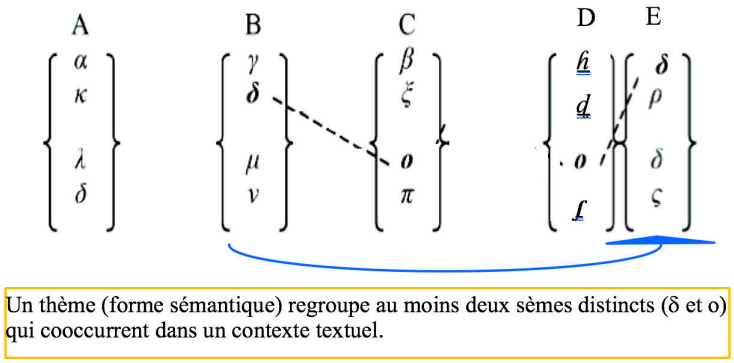


FIG. 2.6 – Repérer un thème dans un texte (graphe modifié à partir de son original proposé par VALETTE (2009))

« villes », « banlieue » et « ghettos », alors que le thème /ville botanique/ est composé de deux sèmes distinctes : /ville/ et /botanique/ (le dernier est représenté les vocabulaires cooccurrents de /ville/ : « arbre », « lilas des Indes », « la rose de Chine », « lys », « fleurs », etc.).

En disposant de l'hypothèse principale que les textes racistes et antiracistes partagent un même fond sémantique avec une certaine catégorie d'isotopies, mais se distinguent par les formes sémantiques manifestées par groupes stables de sèmes, Valette (2004) a mené une série d'analyses axées sur les trois paliers—palier macro-, méso- et micosémantique pour confirmer cette hypothèse.

Au palier macrosémantique, l'étude du genre se fait aux niveau intra- et infratextuel par contraste des textes racistes et antiracistes [14]. Les résultats d'analyses au niveau infratextuel montrent que les textes antiracistes citent davantage les textes racistes, alors que les textes racistes s'approprient directement le vocabulaire des antiracistes. L'analyse sur les données globales infratextuelles porte principalement sur l'organisation des textes et le style dédié à la page Web, observées par le biais des étiquettes multi-modales du code HTML. Autrement dit, la signature sémiotique du site. Il est observé que, par rapport aux textes classifiés comme racistes, les textes antiracistes sont plus structurés et mieux organisés. Par exemple, ces textes contiennent plus de balises comme <H1>, <H2>, <H3> (hiérarchisation des titres) ou <UL>, <OL>, <LI> (balises pour créer des listes), etc. Des balises comme <CITE>, <BLOCKQUOTE> sont spécialement utilisées dans les textes antiracistes, mais assez peu présentes dans les textes racistes. Dans leur aspect visuel (code couleur, typographie, utilisation des images), les textes à caractère raciste font majoritairement usage d'un style emphatique et dynamique : code couleur avec du rouge sang, image GIF en arrière plan, bannières, et foisonnement d'éléments d'hypertextualité (liens internes, adresse e-mail, etc.). Ce qui est à l'opposé du style plus classique et plus sobre des textes à caractère antiracistes.

La détection du genre se base aussi sur les éléments du niveau intratextuel. Les variables comme les points d'exclamation, les adverbes de négation ou l'évaluation emphatique (jamais, rien, peu, tout, trop, etc.) montrent que le genre idéologique

14. Les critères de sélection des textes « repose sur une critique des systèmes de filtrage, et notamment sur ceux qui recourent à de simples listes de mots-clés (CyberSitter, CyberPatrol). Ceux-ci témoignent en effet d'une approche naïve du texte raciste, suggérant qu'il y a des mots racistes et des mots qui ne le sont pas, sans considération pour leur mise en texte » (Valette, 2004).

privilégié par les auteurs racistes est le pamphlet (ou le libelle), avec une forte présence de la diatribe et la polémique.

C'est à partir du second niveau de palier — le palier mésosémantique — que sont explorés les isotopies et les thèmes, qui sont actualisés par les cooccurrences de morphèmes ou de mots. Afin d'obtenir les corrélats d'une lexie, il procède à une sélection des cooccurrents associés à un mot pôle, en l'occurrence, dans les textes racistes et antiracistes, le mot «*immigration*» et le lemme «*étranger*». À partir de ces cooccurrents, il déduit des thèmes comme « immigration invasion » et « immigration croissante » pour les textes racistes, et « flux migratoire » et « fermeture de frontière » pour ceux à caractère antiraciste. Ceci permet de constater la différence des thèmes abordés dans les deux discours confrontés. L'auteur met en évidence la variété des formes sémantiques associées à une même unité lexicale dans deux corpus de textes contrastés de type raciste *vs* antiraciste. Dans l'exemple montré ci-dessous, le discours sur l'« étranger » relève d'un fond sémantique qui varie en fonction des sous-corpus.

Exemple : Le contraste des cooccurrences du mot « étranger »

Tab. 2.1 – Tableau des cooccurrents du mot « étranger »

Corpus	Cooc de « étranger »
Textes antiracistes	irrégularité, régularisation
Textes racistes	illégalité, naturalisation

Dans l'article « Representations of the future in English language blogs on climate change», Fløttum et al. (2014) étudient les points de vue de plusieurs communautés discursives au sein de la blogosphère sur le sujet du changement climatique. Pour cela, l'auteur décrit d'abord la conceptualisation de la notion de « futur » dans les blogs portant sur les changements climatiques. Elle étudie les traits lexicaux (formes sémantiques) et les cooccurrents du mot-pivot « futur », et découvre que le /futur/ constitue une isotopie dans l'ensemble du corpus, qui forme ainsi le fond sémantique des textes. Les vocabulaires cooccurrents de « futur » peuvent être scindés en deux avis contradictoires : l'avis positif est porté par le mot « opportunité » et l'avis négatif par les mots « risque, danger, menace ». Quelque soit les avis, ces discours dénotent que le changement climatique est d'origine anthropique. La diversité des discours liés à des points de

vue différents illustre l'hétérogénéité qui caractérise la blogosphère climatique, et montre que le problème du changement climatique est discuté par plusieurs communautés discursives, chacune ayant sa propre perspective et conceptualisation du changement climatique.

Après l'explication des concepts théoriques et la présentation des travaux pratiques réalisés par d'autres chercheurs dans le domaine, nous allons, à partir des parties suivantes, introduire la méthode quantitative — la textométrie — et les outils techniques qui nous assistent pour notre analyse qualitative.

2.4 Textométrie et Outils

Comme dit Rastier (2001) : « le quantitatif et le qualitatif ne s'opposent aucunement : seule une analyse qualitative peut rendre significatifs des phénomènes quantitatifs remarquables ». La textométrie propose des méthodes qui permettent de mettre à profit des principes fondamentaux de la SI, et réciproquement, la SI permet de conceptualiser la méthode de la textométrie. Dans la section suivante, nous allons lister quelques aspects de la textométrie qui renforcent sa compatibilité avec la SI.

2.4.1 *Textométrie*

La textométrie, aussi appelée *logométrie* ou *statistique textuelle*, doit son origine à la *lexicométrie*. Elle est née dans les années 1970 et a été successivement développée par Pierre Guirard (1954, 1960), Charles Muller (1968, 1977) et Jean-Paul Benzécri (1973). La textométrie regroupe « plusieurs perspectives d'herméneutique textuelle issues des problématiques des sciences sociales au contact de la linguistique » (Maingueneau,1991). Elle est « un ensemble de mesures, de traitements statistiques effectués sur des textes » (Salem,1986). La textométrie travaille sur des corpus de textes intégraux et cherche surtout à décrire et à interpréter des phénomènes linguistiques du corpus. En articulant les procédures quantitatives et les moyens de parcours et d'interprétation qualitatives basés sur la SI, elle met à profit des procédures de tris et des calculs statistiques outillés par des logiciels TAL (Traitement Automatique des Langues). L'objectif est d'assister et d'outiller la lecture humaine dans le respect des données recueillies, de suggérer des points d'appui, des pistes d'investigation. Les résultats sortis aident à mettre en évidence des régularités et des spécificités qui pourraient être négligées par

l'humain. La textométrie fournit une première connaissance d'ensemble du corpus sur les catégories méta-textuelles (genre, style, domaine, etc.). Elle constitue « une démarche empirique qui regroupe un ensemble diversités de pratiques documentaires automatiques et de programmes statistiques, articulés autour d'un nombre limité de principes unificateurs » (Tournier, 1980).

La segmentation du corpus (textes) en unités sémantiques (habituellement de l'ordre des mots) permet de recourir à certaines fonctionnalités fondatrices de la textométrie : le calcul de la cooccurrence, le calcul de la spécificité (cf. Annexe 6, Calcul des spécificités), celui des segments répétés, et la recherche des motifs dans sa concordance voisine (contextualisation locale), ou par rapport au contexte linéaire ou non-linéaire du texte (contextualisation globale). De même, pour étudier les paramètres génériques liés au genre textuel, nous pouvons tirer partie des informations quantitatives sur ces variables qualitatives (lexicales, syntaxiques, modales, sémiotiques, morphologiques, etc.) qui sont propres aux caractéristiques saillantes de tel ou tel genre textuel. Comme dit Denise et François (2001) : « la textométrie permet d'observer l'incidence sémantique du genre textuel [...] les analyses factorielles sur les décomptes et les mesures effectués sur le corpus permettent de confirmer la détermination du global sur le local et les interrelations transverses aux paliers de description ».

Les logiciels textométriques (voir section 2.4.3 Outils textométriques) peuvent effectuer sur les textes des calculs statistiques et l'ingénierie du TAL. Dans ce qui suit, nous allons présenter et expliquer les calculs statistiques essentiels proposés dans les outils textométriques dont nous nous servons pour effectuer les analyses quantitatives.

- Le calcul des **spécificités** (SP) (Lafon,1980) est fondé sur le résultat du calcul de la probabilité d'une fréquence d'une forme du corpus. À chaque forme est attribuée un indice de spécificité, positif ou négatif (cf. Annexe 6 Calcul des spécificités). Cet indice permet d'établir les listes de formes sur-employées ou sous-employées dans chaque partie du corpus par rapport aux autres. Nous pouvons nous servir du calcul de spécificité pour faire ressortir les variables lexicales ou sémiotiques saillantes qui caractérisent une typologie du genre textuel du corpus.
- La **concordance** (CD), est définie par André Salem comme « l'ensemble de lignes de contexte se rapportant à une même forme-pôle ». Il s'agit d'ob-

server, grâce au retour au texte, la localisation topographique des mots ou des parties en question. La fonction « Section » proposée par l'outil Lexico5 [15] permet d'accéder au contexte d'apparition du mot-pivot (cf. Le graphe 4.7 Concordance du verbe nominalisé 保护环境 (protéger l'environnement) dans les textes Ins).

– La **cooccurrence** (COOC) se présente sous forme d'un groupement des formes en présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, etc.). La mesure de cooccurrence permet de repérer les thèmes dans le corpus. La fonction « Thème » du logiciel Hyperbase sert à repérer des corrélats sémantiques pour une description thématique des isotopies et de ses cooccurrents (Rastier, 2001). En outre, les comportements sémantiques représentés par les isotopies, c'est-à-dire des répétitions de l'information qui ne sont pas facilement saisie par la lecture « manuelle », peuvent être identifiés dans un corpus de taille volumineuse grâce à des calculs quantitatifs. La calcul des cooccurrents est crucial, car il illustre le passage du quantitatif (les cooccurrents) au qualitatif (les corrélats). Les cooccurrents ne sont élevés à la dignité de corrélat que s'il est possible d'établir une relation d'isotopie avec d'autres cooccurrents (Rastier, 1995). Le passage de l'analyse lexicale à l'analyse thématique conduit de signes non interprétés à des unités sémantiques qui résultent d'un parcours interprétatif. Par exemple : pour relever les thèmes principaux présentés par notre corpus, nous avons calculé les cooccurrents de l'isotopie /pollution de l'air/ afin de relever lexiques qui cooccurrent de manière récurrente avec elle (cf. section 5.2 Méthode de travail pour l'identification du thème).
– Les **segments répétés** (SG) sont constitués par « une suite de formes dont la fréquence est supérieure ou égale à 2 dans le corpus » (André Salem,1994). Ils mettent en évidence les collocations principales de chaque forme présente dans le corpus, qui peuvent ensuite donner des indices aux analyses du style typologique du genre textuel du corpus.
– **Analyse Factorielle des Correspondances** (AFC), fournit des graphiques des proximités lexicales des diverses parties des corpus selon les facteurs les plus représentatifs (André Salem, 1994, cf. Annexe 7 AFC (Ana-

15. http://lexi-co.com/ressources/manuel-3.41.pdf

lyse factorielle des correspondances)). Ces relations corrélatives présentées sous forme des graphes donnent des indices globaux au niveau macro- et microsémantique. Ces indices sont utiles lorsque nous étudions le genre textuel (à travers les variables lexicales et autres) et les thèmes textuels.

Voici le tableau proposé par Pincemin (2011) afin de décrire le corpus par la méthodologie textométrique :

Tab. 2.2 – Description graphique du corpus

Type de Graphie	Graphie Lexicale	Topographie	Graphie Diachronique
Fonctionnalité symbolique textométrique	Cooccurrence	Section/Segments Répétés	AFC
Présentation dans le corpus	Chaîne de caractères+séparateur	Parition du corpus[1]	Index temporels
Exemple	Mot+blanc	Partition en syntagme	horaire
Exemple	Mot+ponctuation	Partition en phrase	Journal
Exemple	Lemme+étiquette morphosyntaxique	Partition en paragraphe	Mensuel
Exemple	Lemme+étiquette d'Entité nommée	Partition en paragraphes apparentés	Trimestriel
Exemple		Partition par texte	Annuel
Exemple		Partition par sous-corpus	Périodique
Nivea de localisation	LOCAL	LOCAL / GLOBAL	LOCAL / GLOBAL

[1] Le découpage du corpus en parties est un moyen de rendre compte de catégories méta-textuelles et de caractéristiques philologiques (comme le genre, l'auteur, la période historique) (Pincemin,2012).

2.4.2 *Rapport compatible entre l'étude qualitative et le caulcul quantitatif*

Il faut noter que les résultats calculés par les outils techniques ne constituent que des indices quantitatifs : la fréquence, la spécificité, etc. Ces résultats requièrent toujours une démarche qualitative d'interprétation à travers des concepts théoriques. Le tableau illustre le rapport compatible entre l'étude qualitative et le calcul quantitatif dans le cadre du présent travail. Il énumère aussi les outils techniques utilisés pour réaliser ces calculs statistiques.

Études qualificatives	**Macrosémantique**	**Micro- et Mésosémantique**
	Analyse du genre textuel ↓ Variables linguistiques et sémiotiques ↑ Quatre composantes sémantiques	Analyse des thèmes ↓ sème identique récurrent → **Isotopie** + sèmes distincts cooccurrents→**Thème** ↑ Quatre composantes sémantiques
Calcul quantitatif	SP, AFC, Concordance	SP, COOC, SR, AFC
Outils techniques	Hyperbase, Lexico 5	Lexico5, Trameur, Hyperbase

Fig. 2.7 – Rapport entre les études qualitatives et les calculs quantitatifs

2.4.3 *Outils textométriques*

Aujourd'hui, de plus en plus de linguistes sont amenés à constituer des corpus numériques et à les étudier au moyen d'outils techniques. Parallèlement, il existe de plus en plus d'outils offrant une large variété de fonctionnalités textométriques plus ou moins complexes. Nous proposons ici une rapide description des outils d'exploration de corpus utilisés dans le cadre de notre recherche. Étant donné que chaque outil a des avantages et une présentation des résultats qui lui est propre (graphique ou tableau), nous choisissons au cas par cas celui qui convient le mieux pour la réalisation de nos tâches, de façon à faciliter nos analyses.

2.4.3.1 Lexico

Lexico [16], développé depuis 1984 par André Salem, est un outil textométrique en téléchargement libre. Étant l'un des outils d'analyses des données textuelles (ADT) les plus diffusés, il inclut des fonctionnalités textométriques classiques, telles que le calcul de spécificités des formes ou des parties de corpus, l'AFC (analyse factorielles des correspondances) sur une partition du corpus, la carte de sections, les segments répétés, des groupes de formes, les inventaires distributionnels, etc. Dans notre présent travail, nous avons utilisé quatre fonctionnalités du Lexico pour :

1. calculer la spécificité des mots et observer la ventilation de ces mots en fonction de la partition du corpus (par exemple, cf. figure 4.5 Ventilation de quatre types de noms dans les quatre genres textuels pour une partition par genre, année ou ville) ;
2. projeter la carte de sections de manière à observer la distribution des mots cibles dans les partitions de corpus (cf. figure 4.10 Carte de section des termes modaux et de négation dans GOV) ;
3. extraire des groupes de formes avec des mots-clés de sorte de rassembler les mots qui sont dans une même catégorie selon leurs tags (cf. annexe 11.1 Conjonctions) ;
4. générer le graphe AFC des mots cibles dans une partition du corpus (cf. figure 4.6 Verbes et adverbes modaux dans les quatre genres textuels selon l'indice de spécificité) ;

16. Site officiel de l'outil Lexico : `http://lexi-co.com`

5. exploiter la concordance à l'aide du retour au contexte original du mot ou segment d'étude (voir section 4.7 Concordance du verbe nominalisé 保护环境 (protéger l'environnement) dans les textes Ins).

2.4.3.2 Trameur et TXM

Le Trameur et TXM sont tous des outils de l'ADT. Comme d'autres logiciels dans ce domaine, ces deux outils textométriques intègrent de nombreuses fonctionnalités statistiques, documentaires et graphiques visant à l'exploration de la sémantique et des thèmes et à l'interprétation de ces derniers. Par rapport aux autres outils dans la même catégorie, la particularité du Trameur est sa performance de calcule des segments répétés. La présentation des résultats statistiques que l'outil propose, sous forme de tableau, correspond à nos besoins. Nous utilisons aussi la fonctionnalité de « Section » pour repérer des expressions caractéristiques des sous-corpus (voir section 5.5.1 Études sémantiques du thème 1 dans le sous-corpus Ins du Chapitre 5). Enfin, l'option *exporter les parties contenant le motif* permet d'extraire les parties des textes contenant des mots-clés ou des expressions prédéfinies. Quant à TXM, il dispose d'une option permettant de générer un graphe AFC en fonction de l'étiquette syntaxique ; ceci permet d'effectuer des études relatives au genre textuel dans le Chapitre 4 (cf. graphe 4.14 AFC des variables intratextuelles des quatre genres discursifs).

2.4.3.3 Hyperbase

Hyperbase (web édition [17]) est un logiciel en ligne d'exploration documentaire et de statistique des textes, développé par Laurent Vanni [18]. La fonction « Thème » de l'outil est utilisée dans nos études afin de générer des nuages de mots. Ce nuage de mots est composé d'un mot-pivot et de ses cooccurrents. Ce type de visualisation graphique permet, selon les concepts de Rastier, d'identifier un thème dans une partition du corpus (cf. figure 5.5 Cooccurrents de 雾霾 dans les textes du genre Profane), à partir des corrélats sémantiques induits par des sèmes distincts lexicalisés. Par ailleurs, le graphe de la distribution des maladies et symptômes dans les 33 régions en Chine provient aussi d'Hyperbase grâce à la fonctionnalité « Distribution » (cf. figure 1.7 Distribution des maladies par

17. http://hyperbase.unice.fr//?lang=fr
18. La version initiale de l'Hyperbase est développée par Étienne Brunet en 1989

région en Chine dans le sous-corpus WEIBO).

2.4.4 *Travaux combinant la SI et la textométrie*

Plusieurs travaux combinant des aspects d'analyse sémantique textuelle et la méthode textométrique montrent l'intérêt et les potentialités pour l'étude linguistique de corpus volumineux (l'étude des caractéristiques du genre textuel, du fond sémantique et la détection des thématiques, etc.). Ces travaux nous offrent des pistes de réflexion pour nos propres analyses.

L'objectif de l'article « Les mots de la controverse sur le changement climatique» (Chetouani, 2007) est double — a) manifester la polémique des discours scientifiques, politiques, indépendants et technocratiques à propos du changement climatique, en particulier l'effet de serre ; b) interroger le mode de fonctionnement lexical de ces discours pour relever : 1) Qui en parle ? 2) Comment en parle-t-on, avec quels mots et quels arguments ? L'auteur cherche à identifier et comparer les points de vue antagoniste (optimistes/pessimistes) dans les discours. L'étude se fait sur quatre registres de la polémique, c'est-à-dire sur quatre thèmes : « les responsabilités », « les conséquences et l'ampleur du phénomène », « les solutions » et « les projections spatio-temporelles de l'effet de serre ». Pour cela, l'auteur a utilisé la méthode textométrique, ici en l'occurrence l'AFC et le calcul de spécificité. Il a sélectionné les mots « polémique », « controverse », « débat enflammé » et « diatribe » comme mots-pivots. Ces mots-clés forment l'isotopie /controverse/ du corpus. Le calcul des cooccurrents des mots isotopants permet de comparer l'opinion des optimistes (« modéré ») à celui des pessimistes (« alarme », « catastrophe », « urgence », « menace ») face au problème de l'effet de serre. La comparaison se fait aussi sur les styles de rhétorique employés des deux camps. Il constate que les pessimistes mettent l'accent sur «*la désignation radicale des adversaires*, *l'ethos de soi*, le «*je*» *comme marqueur de subjectivité*, *la modalité dépréciative et l'ironie*» pour réfuter l'idée des optimistes, alors que les optimistes utilisent «*la démarche déductive logique et les arguments de l'autorité*» pour démontrer leur point de vue. À l'aide du graphe généré par l'AFC, qui indique les distributions des lexiques par rapport aux quatre types de sous-corpus, l'auteur obtient des résultats montrant que les partis politiques optimistes forment un bloc, qui fait clivage avec les pessimistes scientifiques.

Dans l'article « Approche lexicométrique des controverses climatiques» (Scotto,

Giancarlo et Grégory, 2014), l'auteur a associé les modes opératoires de l'approche linguistique à la méthode statistique quantitative. Son objectif est d'étudier un corpus portant sur des controverses climatiques abordées par quatre catégories d'acteurs : *discours écologique*, *discours politico-anthropologique*, *discours climato-sceptique* et *discours non-climato-sceptique*. D'abord, en se basant sur le dictionnaire généré à partir du corpus, l'auteur a identifié deux thèmes à partir des mots les plus fréquents : « environnement » et « politico-économique » ; ensuite, l'étude des segments répétés confirme la prédominance des vocabulaires spécifiques (selon le calcul de la spécificité). Elle permet également de récupérer des sous-thèmes : « changement climatique », « gaz à effet de serre », « réchauffement climatique », « controverse », etc. L'AFC est utilisée pour identifier les lexiques qui sont associés à chaque acteur et analyser les points de vue antagonistes. À travers le résultat montré sous forme de graphe AFC, on voit bien que le clivage entre les **sceptiques**, qui expriment leur incertitude en projetant des scénarios ou des conséquences sur la société de demain (rhéthorique du changement et vision du futur/prospective), et **le reste des acteurs**, qui constatent des risques liés au changement climatique. Toutefois, il faut noter que la notion de « risque » est utilisée de façon très hétérogène (risque/alerte/danger) en fonction de la nature du discours dans lequel elle s'inscrit.

En s'appuyant sur un jeu de données produit par les internautes de WEIBO de manière spontanée, Renaud (2016) analyse les structures lexicales (sémantiques), sociales (conversationnelles) et spatio-temporelles des échanges autour de deux *mèmes Internet* [19]. Il se sert du JIEBA [20] pour effectuer des pré-traitements sur le corpus (segmentation, élimination des *stop-words* spécifiques à WEIBO), et pour calculer les coocurrents de mots-pivots qui forment un réseau sémantique des *mèmes*. Par exemple, pour analyser le *mème* « 杜甫很忙 » (Dufu est très occupé), il prend 杜甫 [21] comme mot-pivot, et génère un réseau sémantique autour de 杜甫. L'auteur a ensuite relevé plusieurs sous-thèmes : « blague » (haha,

19. La notion de *mème* est proposée par Dawkins (1976) pour définir une unité minimale de propagation des cultures. « L'utilisation du terme *mème Internet* pour décrire la diffusion de messages ne recouvre pas nécessairement la dimension culturaliste du concept initial, mais garde l'idée générale d'une circulation virale d'idées parmi les groupes d'individus » (Renaud, 2016).

20. (JIEBA, « bégayer » en français), un outil de segmentation du chinois développé en Python, qui permet à la fois de segmenter le texte en mots et d'attribuer une étiquette syntaxique à ces segments. https://github.com/fxsjy/jieba

21. 杜甫 (Dufu) est un célèbre poète chinois de la dynastie des Tang.

humour, rire, etc.), « poète » (李白[22]), « réseau social » (internautes, tag, créativités, Kuso[23], etc.). En plus de l'identification des réseaux sémantiques des *mèmes*, l'auteur s'intéresse à la localisation géographique de ces *mèmes* ainsi qu'à leur évolution et leur dispersion spatio-temporelle à travers les échanges des internautes en Chine.

Signorini, Segre et Polgreen (2011) ont construit un corpus numérique constitué d'un large échantillon de *tweets* publics collectés à partir d'une série de mots-clés en lien avec les maladies courantes de l'épidémiologie. Ces mots sont par exemple : *flu* (rhume), *swine* (grippe porcine), *illness* (maladie), *influenza* (grippe), *symptom* (symptôme), etc. Avant d'aborder les analyses, l'auteur procède à une série de pré-traitements du corpus afin de limiter le bruit dans les données. Les *tweets* qui contiennent moins de 5 mots, et ceux qui ne sont pas encodés en ASCII, sont supprimés ; les éléments typiques du microblogging ne portant pas beaucoup d'informations (#hashtag, @, lien, etc) sont également éliminés. Grâce aux données de géolocalisation des *tweets*, les auteurs ont exploré l'évolution géographique de ces maladies mentionnées en mots-clés dans des contextes sociaux particuliers : *travel/trip* (voyage), *fly* (avion), *cruise* (croisière), *ship* (bateau) ; en plus, ils ont analysé l'inquiétude des consommateurs sur le porc : *porc* (porc), *bacon* (bacon) ainsi que les contre-mesures portées par ces consommateurs : *hygiene* (hygiène) et *mask* (masque). À l'aide des labels horodatés, ils peuvent observer l'évolution temporelle des ces trois thèmes discutés dans les *tweets* correspondants.

Wang, Paul et Dredze (2014) cherchent à identifier et à caractériser les sujets majeurs autour des problèmes de santé discutés sur WEIBO avec des *topic* modèles probabilistes[24]. Pour cela, l'auteur récupère sur cinq ans un million de messages *weibo* contenant les mots-clés 流感 (grippe), 生病 (être malade), 医生 (docteur). Ces messages sont ensuite annotés grâce à un dictionnaire médical visant à identifier les maladies, symptômes et autres terminologies médicales.

22. 李白 (Libai) est un autre célèbre poète chinois de la dynastie des Tang, avec 杜甫 (Dufu), ils sont considérés tous deux les poètes les plus connus de la dynastie des Tang.

23. Kuso (, ㄑ) est un mot japonais, qui signifie « merde » ou « conneries ». Ce mot est utilisé en Asie de l'Est pour désigner la culture Internet qui comprend généralement tous les types de camps et de parodies. kuso est souvent prononcé comme une interjection. Il est également utilisé pour décrire des sujets scandaleux et des objets de mauvaise qualité. Traduit selon la définition du KUSO proposé par le site https://baike.baidu.com/item/KUSO.

24. Le *topic model*, autrement dit la modélisation d'un corpus de documents et des différents thèmes ou topics qu'ils contiennent (Francesiaz, Graille et Metahri, 2015).

Les résultats montrent que le « rhume » ou la « grippe » constitue la maladie la plus discutée dans les *weibo* obtenus. Ces mots apparaissent fréquemment avec un vocabulaire désignant les symptômes relatifs à ces maladies : 头痛 (maux de tête), 咳嗽 (toux), 发烧 (fièvre). Ces derniers sont également associés à des verbes ou des noms liés à des remèdes : 多吃 (manger plus de), 生姜 (gingembre), 蜂蜜 (miel) ; ou encore à des noms désignant des personnes victimes : 妈妈 (maman), 宝宝 (bébé).

À l'aide de la méthode textométrique, Hồ-Đình et Valette, (2014, 2017) réalisent une étude interprétative sémantique dans une sphère multidimensionnelle (linguistique, technique, social, culturel, etc.) pour analyser et contraster deux types de discours institutionnels et informels (forum de discussion). Leur corpus est composé des textes français et vietnamiens sur la même thématique : la prévention médico-sanitaire du VIH. En repérant et caractérisant les comportements linguistiques au niveau des acteurs (dialogique) et de la temporalité (dialectique) dans les deux types de discours, les auteurs constatent que « les discours normés que produisent les institutions sont concurrencés par les discours informels », et « les discours du web social peuvent compléter les discours institutionnels ».

Dans l'article « What health-related information flows through you every day ?» (Yang, Yang et Zhou, 2015), les auteurs utilisent l'approche théorique du modèle de croyance en Santé(HBM pour « Health belief model ») pour explorer comment les informations concernant le brouillard de pollution sont discutées par les différents types d'organisations dans les réseaux sociaux en Chine. Le HBM propose six catégories de classification : susceptibilité perçue, gravité perçue, avantages perçus, obstacles perçus, auto-efficacité et indices d'action. En se basant sur ces catégories, Yang, Yang et Zhou analysent un corpus constitué de 756 messages *weibo* sur les relations entre trois types d'organisations (gouvernementales, non-gouvernementales, entreprise privée) à propos du brouillard de pollution et de sa menace pour la santé. Les résultats de cette étude indiquent que les *weibo* publiés par les entreprises privées, qui sont pour la plupart des publications de vente, sont classés dans la catégorie « avantages perçus » (La catégorie « avantages perçus » porte sur la croyances des individus que leurs comportements leur apporterais des avantages potentiels («*perceived benefits, is about the individual's beliefs about the potential benefits and consequences of the behavior*» (Yang, Yang et Zhou, 2015))). Tandis que la petite quantité de *weibo*

publiées par les organisations gouvernementales sont classés « gravité perçue » (La « gravité perçue » se réfère aux opinions des individus sur les graves conséquences si l'on ne réagit pas pour prévenir les problèmes de santé («*Perceived severity refers to individual's beliefs on the serious consequences of not acting on preventing the health problem*» (Yang, Yang et Zhou, 2015))).

Les travaux de Eensoo et Valette (2015) sur des corpus issus de discussions médicales et sanitaires sont axés sur deux agonistes : disphorique et euphorique, favorable et défavorable. Les auteurs mobilisent la sémantique textuelle et les analyses textométriques pour classer ces textes subjectifs. Les critères de classement sont choisis selon deux composantes sémantiques. La composante **dialectique** manifeste son rôle d'argumentation et présente les évaluations modales avec les éléments descriptifs ou argumentatifs : véridictoire : le vrai/faux ; thymique : positif/négatif. Alors que la composante **dialogique**, représente les acteurs et le positionnement énonciatif à l'aide des pronoms personnels et possessifs, permettant de fonder la typologie des énonciateurs représentés. Le tableau ci-dessous, issu de l'article mentionné, met en évidence les relations entre les composantes sémantiques, le genre textuel et le vocabulaire caractéristique.

Tab. 2.3 – Caractéristiques des genres textuels des discours dysphoriques et euphoriques issues des travaux de Eensoo et Valette (2015)

	Dialogique	**Dialectique**
Dsyphorique	**Egocentrer**	**Extraction de l'action**
	- surreprésentation de la 1er personne du singulier	- On me dit que
Euphorique	**Acteur-énonciateur**	**Texte descriptif ou argumentatif**
	- surreprésentation de la 2ème personne du singulier - partage des expériences personnelles - temoignage personne intertextualisé avec hyperlien	- après avoir - par contre

2.5 Conclusion

Dans ce chapitre, nous avons présenté le cadre théorique dans lequel s'inscrivent nos analyses qualitatives et quantitatives. Dans ce cadre, nous faisons un usage opportuniste de la SI en exposant certaines propositions clés qui sont adéquates à notre recherche : le **principe général** : « le global détermine le local, le sens est contextuel », et **les concepts fondamentaux** de la théorie SI : le sème, l'isotopie, le thème, le genre textuel, les trois paliers sémantiques et les quatre composantes sémantiques, ainsi que les relations internes entre ces aspects. Toutefois, « le quantitatif et le qualitatif ne s'opposent aucunement », ils sont en effet complémentaires de manière réciproque. La textométrie propose des procédures de tris et des calculs statistiques outillés par des logiciels TAL qui permettent de mettre en œuvre des principes fondamentaux de la SI ; réciproquement, la SI permet de conceptualiser la méthode de la textométrie. Ainsi, nous avons introduit la méthodologie quantitative de la textométrie ainsi que des outils techniques qui sont pertinents pour nos analyses du corpus. En plus de la présentation théorique et stratégique des concepts de la SI, nous avons aussi listé une série de travaux qui ont intégré directement ou indirectement les concepts théoriques de la SI, en combinant (ou non) l'approche textométrique. Ces travaux précédents sont des traits de lumière qui nous inspirent dans nos propres analyses.

Chapitre 3

Constitution du corpus et Outils

3.1 Introduction

L'enjeu de la sémantique de corpus consiste à étudier les textes en eux-mêmes et pour eux-mêmes. Un corpus adéquat doit correspondre à des critères de représentativité ou d'homogénéité qui reposent sur : le choix des textes, le mode de nettoyage, l'encodage, l'étiquetage, la structuration du corpus, la correspondance aux attentes scientifiques. Ainsi, en suivant les critères proposés par Rastier, dans ce chapitre qui est consacré à la présentation générale du corpus, nous expliquerons d'abord le choix de nos sources (section 3.2 Choix des sources). Ensuite, nous présenterons en détail les sites Internet desquels sont tirés les textes constituant le corpus, et nous relèveront les caractéristiques de chaque site (sections 3.2.2, 3.2.4, 3.2.3 et 3.2.5). Nous expliquerons comment collecter des données textuelles depuis ces sites et avec quels outils correspondants (section 3.3 Collecte des données et choix des outils), et comment nettoyer le corpus grâce à une série de prétraitements. Ces traitements sont suivis de la présentation détaillée du processus de segmentation et d'annotation (section 3.7 Organisation du corpus et outils) du corpus. Nous terminerons ce chapitre en exposant les méthodes en vue d'organiser le corpus à partir des métadonnées prédéfinies (section 3.7.1 Extraction des métadonnées).

3.2 Choix des sources

3.2.1 *Critères du choix des sources*

D'après Tournier (1980), la pertinence d'un corpus repose sur le choix adéquat des sources de données textuelles. Ces dernières doivent non seulement partager un ou plusieurs **caractères communs** et comporter une certaine **ré-**

gularité, mais aussi représenter la **diversité** et l'**équilibre** des informations textuelles. Pour éviter des écarts d'analyses, la règle de composition du corpus doit garder l'**homogénéité** générale des sources d'information. De plus, il faut prendre en compte la **variation diachronique**, la **variation des énonciateurs**, des types de discours ou des divisions internes aux textes, etc.

Conformément aux consignes données par TOURNIER, nous avons sélectionné quatre sites Web (homogénéité des sources d'information et caractère communs), dont un réseau social (diversité), pour collecter des textes/*weibo* relatifs au « brouillard de pollution en Chine » (équilibre des sources d'information). Ces textes ou *weibo* datés de 2006 jusqu'à 2018 (variation diachronique) appartiennent à quatre genres textuels (variation d'énonciateurs) : GOV [1] pour sous-corpus Ins [2], PEOPLE pour sous-corpus InsM, SOHU pour sous-corpus InfM et WEIBO pour sous-corpus Profane. Dans la partie suivante, nous allons présenter brièvement les quatre sites sources. La présentation commence par une description générale de la nature, du rôle et des caractéristiques principales des quatre sites sources. Puis, nous nous focaliserons, au niveau infratextuel, sur l'architecture du site, la composition de la page d'accueil pour chaque site, ainsi que l'emplacement et la structure du texte publié sur la toile. Un récapitulatif des caractéristiques des quatre sites sources sera donné à la fin pour recenser leurs points communs et leurs divergences.

3.2.2 *Présentation de GOV*

GOV est le site officiel du gouvernement chinois, administré par le Conseil d'état de la République populaire de Chine. Il joue le rôle de porte-parole des dirigeants chinois et transmet des informations à la population sur le gouvernement de Chine, tant au niveau central qu'au niveau provincial. Le site possède au total huit rubriques : « Conseil des affaires d'État », « Premier Ministre », « Actualités », « Politiques », « Interaction », « Services publics », « Archives », « Situation de Chine ». GOV est proposé en deux langues : le chinois et l'anglais [3].

La plupart des articles publiés sur GOV sont produits par les éditeurs internes au site, qui travaillent pour le gouvernement. À travers ces articles publiés sur la toile, les autorités transmettent directement leurs décisions et stratégies

1. Site officiel du GOV : www.gov.cn.
2. Nous allons employer l'acronyme GOV, PEOPLE, SOHU, WEIBO ou Ins, InsM, InfM, Profane pour faire référence aux quatre sites.
3. Voir le site : http ://english.gov.cn/Page/Uuid/e15c646a-446e-11e4-8156-03a6019c7a4e

politiques, économiques et sociales. Le tableau 3.1 Tableau récapitulatif des métadonnées d'un article du GOV et les figures 3.1 Page d'accueil du GOV [4] et 3.2 Exemple d'un article publié sur GOV [5] montrent respectivement l'architecture HTML du site (principalement les informations relatives aux métadonnées incluses dans les balises), un aspect de la page d'accueil, et la mise en page type d'un article publié sur GOV. À travers ces supports, nous pouvons constater que la structure du site et l'organisation du article du GOV sont du style classique. Le site délimite clairement les éléments composants : rubrique d'origine, titre, date de publication et le contenu principal du texte. Cette structure régulière des données facilite à la fois la tâche de récupération des articles, et celle de catégorisation données récupérées (voir section 3.7.2 Format et balisage du corpus).

TAB. 3.1 – Tableau récapitulatif des métadonnées d'un article du GOV

Encodage	<meta http-equiv="Content-Type" content="text/html; charset=utf-8">;<meta http-equiv="x-ua-compatible" content="IE=edge" >
Adresse URL	<link href="http ://www.gov.cn/govweb/xhtml/favicon.ico" rel="shortcut icon" type="image/x-icon">
Titre de l'article	<title> 河南省研判近期环境空气质量状况 </title>
Résumé de l'article	<meta name="description" content=" 为积极应对今年首轮重污染天气，河南省 12 日召开了重污染天气应急应对视频会议，研判近期环境空气质量状况，并于当日零时启动了重污染天气橙色预警，部署重污染天气应对措施。2018-01-16-21 :01 :00" />
Rubrique et catégorie	<meta name="catalog" content="c1443"><meta name="lanmu" content=" 滚动新闻">
Auteur de l'article	<meta name='author' content=" 刘淼">
Date de publication et de modification	<meta name='firstpublishedtime' content="2018-01-16-21 :01 :00">; <meta name='lastmodifiedtime' content="2018-01-16-21 :01 :00">

FIG. 3.1 – Page d'accueil du GOV [6]

6. Page consultée en février 2018.

FIG. 3.2 – Exemple d'un article publié sur GOV [7]

7. Page consultée en février 2018 : http://www.gov.cn/xinwen/2017-12/25/content_5250284.htm.

3.2.3 *Présentation de SOHU*

Fondé en 1998 [8], SOHU est un site web qui fournit une vaste plateforme d'information, de divertissement et de communication à des millions d'utilisateurs chinois. Doté d'environ une cinquantaine de rubriques diverses sur la société, la finance, la culture, la santé, le sport, etc., il publie non seulement des nouvelles de Chine et des pays étrangers, mais aussi les actualités étroitement liées à la vie quotidienne du peuple chinois. Les articles publiés sur SOHU sont tous et uniquement rédigés en chinois. Il faut noter que le mode de production des articles sur SOHU combine celle du GOV et du WEIBO (cf. section 3.2.5 Présentation de WEIBO) : ceux qui sont affichés sur le site sont produits par des éditeurs spécialisés, et ceux qui sont stockés dans les forums ou blogs hébergés par SOHU sont rédigés par les utilisateurs. Toutefois, il existe encore une troisième modalité, certains articles sur SOHU sont tirés d'autres sources web de nature informelle ; la source originale de ces articles s'affiche alors juste après le titre principal de l'article. Nous pouvons donc faire un tri des articles purement SOHU à partir de ces indices.

Nous avons établi nos premières observations en étudiant la page d'accueil (figure 3.3 Page d'accueil du SOHU [9]) et la structure type d'un article (figure 3.4 Exemple d'un article de SOHU [10]) sur SOHU. D'abord par rapport au GOV, SOHU a beaucoup plus de rubriques (50 sur SOHU contre 8 sur GOV). Cela signifie plus de variétés et de diversités en matière de thème et de contenu textuel sur SOHU. Par ailleurs, il faut noter que les articles de la rubrique « Blog » du SOHU (voir figure 3.4 Exemple d'un article de SOHU [11]) sont en général produits par des particuliers pour exprimer leurs centres intérêts personnels et opinions personnels. Ceci constitue une différence fondamentale entre SOHU et GOV, puisque dans ce dernier la publication des articles est contrainte et doit suivre la demande des autorités hiérarchiques.

8. Source d'information : https://baike.baidu.com/item/%E6%90%9C%E7%8B%90?fromtitle=sohu&fromid=216383#2_16.

FIG. 3.3 – Page d'accueil du SOHU [12]

12. Page consultée en février 2018.

FIG. 3.4 – Exemple d'un article de SOHU [13]

13. Page consultée en février 2018 : http://www.sohu.com/a/122255118_255621)

3.2.4 *Présentation de PEOPLE*

PEOPLE[14] a été mis en ligne en 2000. Il doit son origine au Quotidien du Peuple 人民日报 *(Renmin Ribao)*. C'est un journal officiel du Parti communiste chinois et le journal le plus connu en Chine. En tant que « doyen » de la presse écrite nationale, le Quotidien du Peuple est l'un des dix plus grands journaux du monde avec un tirage de trois millions d'exemplaires[15]. Le site web PEOPLE, version dérivée en ligne du journal Quotidien du Peuple, hérite de la fonction principale du journal : informer le peuple sur la vie politique en Chine et l'actualité mondiale, et commenter les grands sujets nationaux et internationaux. La nouvelle forme de présentation de la page facilite la distribution de l'information à son audience. PEOPLE couvre presque tous les aspects de la vie sociale en Chine grâce à ses 31 rubriques d'informations, telles que « Politique intérieure », « Actualité internationale », « Point de vue et opinion », « Économie », « Science et éducation », « Société », « IT », « Protection de l'environnement », « Armée et défense », « Culture et loisir », « Vie au quotidien », etc. Le site est également disponible en 16 langues étrangères. Cette caractéristique multilingue favorise la diffusion des discours du gouvernement chinois dans le monde.

À propos des caractéristiques du PEOPLE, nous avons fait les constats suivants : d'un côté, PEOPLE possède beaucoup de rubriques similaires à GOV, notamment liées aux activités des dirigeants (par exemple « Nouvelles politiques », « Dirigeants chinois » et « Affaires du personnel gouvernemental »). Cette similitude s'explique par la nature institutionnelle du PEOPLE. De l'autre, PEOPLE partage l'aspect médiatique et co-édition de SOHU, puisque les articles publiés sur ce site proviennent aussi partiellement d'autres sites web. En plus de ce point commun, nous avons aussi observé des similitudes entre PEOPLE et SOHU au niveau de l'architecture de la page d'accueil (voir figure 3.3 Page d'accueil du SOHU[16] et figure 3.5 Page d'accueil du PEOPLE[17]), ainsi que la concordance des thèmes abordés dans les rubriques de même nom (idem). En ce sens, PEOPLE partage les caractéristiques institutionnelles du GOV et celles médiatiques du SOHU. Nous allons investiguer dans le Chapitre 4 et le Chapitre 5 si PEOPLE garde toujours son double statut au niveau des caractéristiques du genre et de la sémantique discursives.

14. Site officiel du PEOPLE : www.people.com.cn.
15. Source d'information : http://french.peopledaily.com.cn/209354/311716/index.html. Page consultée en janvier 2017.

FIG. 3.5 – Page d'accueil du PEOPLE [18]

18. Page consultée en février 2018.

FIG. 3.6 – Exemple d'un article publié sur PEOPLE [19]

19. Cf. site http://paper.people.com.cn/rmrb/html/2017-12/20/nw.D110000renmrb_20171220_3-01.htm. Page consultée en février 2018.

3.2.5 *Présentation de WEIBO*

Lancé en août 2009, SINA WEIBO[20] est le premier site de microblogging (*weibo*, 微博 en chinois, abrégé de messages WEIBO, équivalent de *tweet*) qui a été mis en ligne en Chine. Il a été immédiatement adopté par les Chinois, qui sont séduits par la rapidité de l'actualisation des informations et par le dynamisme de l'interface. Sur WEIBO, les utilisateurs peuvent recevoir, partager, ou commenter les nouvelles ou des actualités des célébrités, des entreprises commerciales, des sociétés de médias, des organisations à but non lucratif et également des organismes gouvernementaux. Ils peuvent aussi produire eux-mêmes des *weibo* sous forme de courts messages limités à 140 caractères chinois. Selon les statistiques publiées par l'entreprise Sina, WEIBO compte plus de 500 millions d'utilisateurs[21], 20 millions de nouveaux inscrits par mois ; plusieurs millions de messages sont publiés chaque jour sur la plateforme. Par rapport au mode de transmission de l'information proposé par les médias traditionnels, WEIBO offre une plateforme de communication qui rend plus facile et plus rapide la diffusion et l'actualisation de l'information. D'ailleurs, il faut noter que même si la forme des *weibo* est similaire à son équivalent *tweet* (limite de mots, possibilité de combiner les textes avec d'autres supports numériques tels que l'audio ou la vidéo), ils diffèrent l'un de l'autre en matière de contenu, la plateforme Twitter met l'accent sur les sujets d'actualité (Yu, Asur et Huberman, 2015), tandis que les contenus les plus échangés et discutés sur WEIBO sont plutôt multidimensionnels (Sullivan, 2013). Selon les résultats de recherche de Li et al. (2015), les *hot weibo* sont classés en huit catégories : le divertissement, les actualités sociales, la mode, la vie et la santé, le renseignement et la recherche d'aide, la promotion des ventes, le Fengshui et la fortune. Les catégories « santé » et « actualités sociales » représentent à elles seules 65% des messages *weibo* qu'ils ont classés. Ce qui montre que « les problèmes sociaux (qui font partie de la catégorie « actualités sociales ») ainsi que « la santé publique » sont des sujets auxquels les internautes portent beaucoup d'attention.

Comme ce que montrent les images suivantes (cf. figures 3.7 Page d'accueil de WEIBO[22] et 3.8 Exemple d'un *weibo* publié sur le site WEIBO[23]), par rapport

20. Site officiel du WEIBO : www.weibo.com.
21. Source d'information : https://thenextweb.com/asia/2013/02/21/chinas-sina-weibo-grew-73-in-2012-passing-500-million-registered-accounts/. Page consultée en mai 2019.

aux trois sites précédents, l'architecture de la page de WEIBO est beaucoup plus complexe. Contrairement aux trois sites web statiques précédents où les textes sont seulement constitués de caractères chinois (ou de mots anglais) et de ponctuation classique, WEIBO, en tant que réseau social, est construit de façon plus dynamique. Ce dispositif permet de diversifier les modes de publication, de commenter et de transférer des informations, et nous pouvons choisir entre textes écrits, images, audio, vidéo, voire combiner deux ou plusieurs modes. De plus, WEIBO propose trois fonctions (voir *infra*) pour varier et diffuser davantage les informations. Pour ce faire, la plateforme WEIBO crée un réseau de relations hétérogènes, qui englobe des utilisateurs issus de différentes couches sociales.

- Créer des topics en utilisant deux croisillons ##
  Exemple :
  Message original : # 雾霾天 # 雾霾天我出不了门了 : (
  Traduction : #jourdesmog# Je ne peux pas sortir à cause du brouillard de pollution : (

- Mentionner une ou plusieurs personnes avec @+alias de la personne en commentaire ou en transfert
  Exemple :
  Message original : 今天有雾霾要记得戴口罩哦 @miaomiao123
  Traduction : N'oublie pas de porter ton masque aujourd'hui@miaomiao123

- Ajouter des émoticônes pour enrichir des émotions
  Exemple :
  Message original : 今天没雾霾 [高兴][转圈]
  Traduction : Il n'y pas de brouillard de pollution aujourd'hui[joie][tourner]

Une autre spécificité qui distingue WEIBO des trois autres sites Web, c'est la section consacrée exclusivement au profil des utilisateurs. Sur WEIBO, les informations basiques sur l'utilisateur sont enrichies par le sexe, l'âge, la région d'origine ou la région de localisation, et éventuellement par le secteur de travail. Ces informations, notamment celles qui concernent la localisation de l'utilisateur ou sa région d'origine, nous permettent d'explorer notre corpus à travers des paramètres métatextuels, comme par exemple la distribution géographique de

certains thèmes lexicaux. Les métadonnées associées à chaque publication *weibo*, telles que la date et l'heure de publication ou le nombre de partages, nous permettent également d'observer les tendances évolutionnelles d'un sujet donné.

FIG. 3.7 – Page d'accueil de WEIBO [24]

FIG. 3.8 – Exemple d'un *weibo* publié sur le site WEIBO [25]

24. Page consultée en février 2018.
25. Page consultée en février 2018.

3.2.6 *Conclusion*

Le premier site GOV de nature institutionnelle est le site officiel du gouvernement chinois. Les articles publiés sur ce site, considéré comme porte-parole des dirigeants chinois, servent à transmettre les idées politiques, économiques ou sociales des hiérarchies au peuple chinois. Dans le but de comparer les informations contenues dans différents types de médias, nous avons choisi comme seconde source le site SOHU, dont le genre diffère diamétralement du site institutionnel GOV. Si les deux sites précédents s'opposent l'un et l'autre par leur nature, le troisième site est de type mixte, c'est à dire qu'il partage la nature institutionnelle avec GOV, mais dispose d'un mode de transimission similaire à SOHU. Pour notre quatrième source, nous avons choisi le premier réseau social WEIBO, caractérisé par sa rapidité de mise à jour des informations, son interactivité entre utilisateurs et son caractère divertissant. Ce quatrième site source s'opposent aux trois autres sites quel que soit au niveau de la forme de présentation ou de publication.

3.3 Collecte des données et choix des outils

3.3.1 *Introduction*

Nous avons utilisé trois méthodes différentes pour récupérer les données textuelles : par outil de crawling, par script R, et par requête de mots-clés dans la base de données. Dans cette partie, nous allons décrire notre travail de constitution du corpus, de l'aspiration des données textuelles jusqu'à leur organisation, en passant par le nettoyage, la segmentation sémantique, l'annotation. Chaque étape, réalisée avec un outil technique considéré par nous comme le plus performant, sera expliquée de manière détaillée. Nous récapitulerons à la fin de ce chapitre l'ensemble des informations quantitatives relatives au corpus.

3.3.2 *Collecte des données à l'aide d'un crawler*

Étant un outil de *crawling*, Gromoteur [26] permet non seulement de télécharger les pages web d'un site, mais aussi d'effectuer des prétraitements et analyses statistiques et linguistiques. Pour collecter en mass les articles publiés sur les trois sites statiques : GOV, PEOPLE et SOHU, nous nous somme servis de la

26. Site du Gromoteur : http://gromoteur.ilpga.fr/.

fonctionnalité principale de Gromoteur : *Crawler*. Trois éléments de requête sont demandés par l'outil pour déclencher le *crawler* :

- L'adresse URL du site web (ex : www.gov.cn) ;
- Les mots-clés (ex : 雾霾) ;
- La quantité [27] de pages web à crawler (ex : 100).

Par exemple, pour *crawler* (ou aspirer) les articles du site GOV au sujet du « brouillard de pollution en Chine », il faut donner au robot d'indexation le mot-clé suivi d'une adresse URL (tel que : « 雾霾 site : www.gov.cn ») comme requête d'entrée. La quantité de pages web à crawler est demandée par l'outil avant de déclencher la recherche [28]. Il est possible de prédéfinir une série de conditions dans l'outil afin d'affiner la récupération des données et d'éliminer les pages qui ne sont pas pertinentes par rapport à nos attentes. Par exemple, nous pouvons forcer l'outil à n'aspirer que les pages web avec un **encodage** donné, comme l'UTF-8. De même, il est possible de contrôler la **langue d'édition** et de ne garder que les pages éditées en chinois. On peut aussi spécifier le **format des données** textuelles inclut dans la page : en effet, les articles peuvent parfois être au format PDF ou EXCEL, et l'outil nous laisse la possibilité de conserver ou d'ignorer ces types de page, selon nos besoins. L'outil nous offre la possibilité de fixer la taille à crawler en mégabit ou en nombre de phrases de la page, ainsi que les rubriques spécifiques à prendre ou à éviter [29].

3.3.3 *Prétraitement du corpus avec Gromoteur*

Lorsque le crawler a fini son travail de récupération, il affiche toutes les pages téléchargées dans la fenêtre « résultat » de l'outil. Toutefois, ces pages obtenues contiennent non seulement l'article même, mais aussi des informations surérogatoires aspirées du site, puisque l'outil a conservé toutes les données brutes renfermées dans chaque élément HTML depuis le <head> jusqu'au </html>. L'ensemble des pages nécessite ainsi des prétraitements en vue d'éliminer le bruit et conserver la partie porteuse de données utiles, à savoir le contenu de l'article. Ce nettoyage peut être réalisé à l'aide de la fonctionnalité « Select », qui permet

27. En général, chaque page web contient un seul article, ainsi la quantité des pages web en d'autres termes, signifie la quantité des articles.

28. D'après notre expérience, il ne faut pas surcharger l'outil et rester en deçà de 1 000 pages à chaque requête

29. Ceci se réalise grâce à la possibilité offerte par l'outil permettant de spécifier l'URL : URL matches ou URL doesn't match

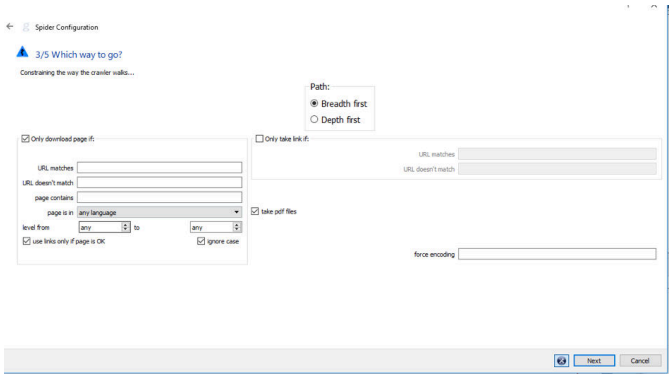


FIG. 3.9 – Fenêtre de la configuration de Spider du Gromoteur

d'analyser automatiquement les balises HTML et propose des choix de conservation ou d'élimination de la partie comprise dans telle ou telle balise. En examinant manuellement l'organisation de la page des trois sites traditionnels, nous avons constaté que le contenu de l'article est en général stocké dans la balise <p> [30](cf. figure 3.10 Exemple des codes sources de la page wab du GOV [31]), alors que les informations descriptives d'un article sont encapsulées dans la balise <head>, où figurent généralement la **date de publication**, l'**origine de l'article** (s'il s'agit d'un article transféré d'un autre site), la **ville** (optionnel), l'auteur (optionnel), etc. Ces informations constituent les métadonnées d'un texte à partir desquelles nous allons organiser notre corpus (voir section 3.7.1 Extraction des métadonnées).

La fonction « Select » nous aide à enlever 90% des informations surérogatoires, telles que les publicités, les liens externes, des images, des QR code, etc. Le reste du bruit présent dans les articles est dû soit à la présence de doublons, soit à des informations que les articles et les messages *weibo* ont en commun tels que les @ ou les mots-dièse (hashtag #), soit à des signes de typographie comme les ponctuations. Nous considérons ce type d'informations à part et détaillons leur traitement dans les sections 3.3.6 Pré-traitement sur les doublons et 3.4.2 Gestion des éléments technodiscursifs.

30. La vérification manuelle est bien nécessaire : même si dans la plupart des cas, le texte principal est stocké dans la balise <p>, le contenu de l'article peut également apparaître dans la balise .

31. Un exemple des codes sources d'un article publié sur GOV : `http://www.gov.cn/xinwen/2018-01/16/content_5257282.htm`

FIG. 3.10 – Exemple des codes sources de la page wab du GOV

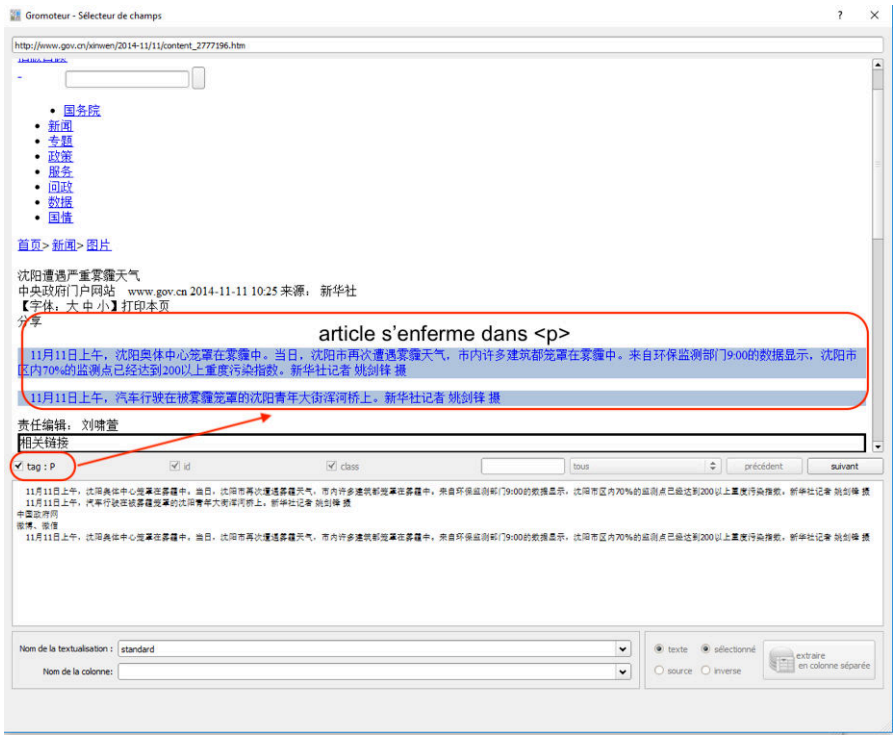


FIG. 3.11 – Prétraitement d'un article du GOV avec « Select »

À l'aide du Gromoteur, et après la vérification manuelle[32], nous avons obtenu 1095 articles sur GOV, 835 articles sur PEOPLE et 626 articles sur SOHU. Tous les articles chinois sont au format texte UTF-8 et produits entre 2006 et 2016.

3.3.4 *Collecte des données à partir d'une base de données*

Si les articles ont été collectés grâce à des procédures semi-automatiques, la récupération des *weibo* a quant à elle été réalisée avec deux méthodes différentes, que nous allons présenter dans cette partie.

Un premier échantillon de *weibo* a été obtenu depuis la base de données BCC[33] (BLCU Corpus Center, BCC) développée par l'Institut de Big Data et l'Université des Langues et Cultures de Beijing (BLCU). La BCC est un service en ligne qui dispose d'une collection d'échantillons de textes d'environ 15 milliards

32. Il arrive souvent que la page récupérée ne contienne que des entêtes et des publications, où l'article est absent. Ce type d'erreurs qui sont non-aperçues ou ignorées par l'outil doivent être examinées à la main.

33. http://bcc.blcu.edu.cn

de mots écrits en plusieurs langues (principalement en chinois). Ces données textuelles proviennent des sources diverses, comme les micro-blogs (*weibo*), les articles scientifiques et technologiques, les textes littéraires et la presse journalistique. Grâce au moteur de recherche proposé par BCC, l'utilisateur peut formuler des requêtes constituées d'unité de caractères, de mots, de mots syntaxiquement étiquetés, ou d'un binôme {expression régulière+caractère/mots/mots annotés } (voir Annexe 3 Exemples de requête dans BCC), et ainsi extraire les données textuelles correspondantes.

Nous avons créé notre propre liste de requêtes à partir d'un ensemble de mots désignant le brouillard de pollution, tels que 雾霾 (brouillard de pollution), 霾 (brouillard), 空气污染 (pollution de l'air), 大气污染 (pollution atmosphérique), PM2,5 (Particular Matter Ø 2,5 um), PM10 (Particular Matter Ø 10 um). La figure 3.12 Capture d'écran du résultat de recherche avec le mot-clé 雾霾 sur BCC présente le résultat de recherche avec 雾霾 dans BCC. Pour obtenir dans la mesure du possible les *weibo* contenant le mot-clé dans un contexte exhaustif, il vaut sans doute mieux noter l'expression régulière telle qu'on la saisit dans l'outil avec dix fois de «.», signifiant 10 caractère quelconque[34], à gauche et à droite du mot-clé 雾霾. Cette version allongée de requête nous permet d'élargir le contexte de production de *weibo*, autrement dit, elle nous permet d'obtenir plus d'information sur un seul message quel que soit son état original[35]. Contrairement aux données du type « article » crawlées depuis les trois sites web, qui demandent une série de prétraitements de manière à conserver uniquement le contenu textuel des articles, les messages *weibo* que nous avons collectés depuis la BCC sont d'ores et déjà nettoyés. Ce type de sous-corpus ne requiert donc pas de pré-traitement. Finalement, nous avons recueilli un total de 18,898 de messages *weibo* depuis la base de données BCC.

3.3.5 *Collecte des données par script R*

La deuxième partie des *weibo* a été collectée automatiquement à l'aide d'un script R (Weibo API R Scraper) réalisé par Turenne. Le script a recours à la li-

34. En fonction de la longueur du mot-clé défini, et d'après notre test, 10 «.» est l'élément extremum que nous pouvons rajouter de chaque côté du mot-pivot.

35. Nous avons deux états originaux pour les *weibo* : soit il se présente tout seul avec son contenu d'origine ; soit le message original est entouré par les commentaires « retwittés » des autres internautes.

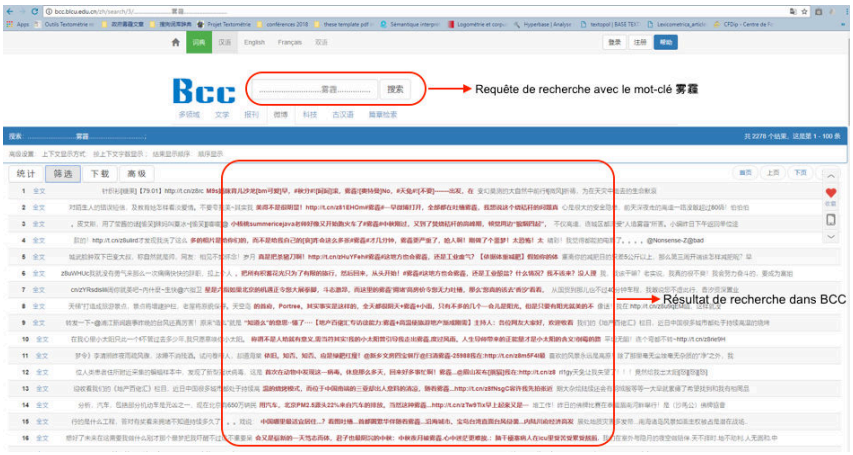


FIG. 3.12 – Capture d'écran du résultat de recherche avec le mot-clé 雾霾 sur BCC

brairie Phantom qui simule l'usage d'un navigateur web et télécharge directement des pages en générant des URI. Cet outil nous a permis d'obtenir 1 million de *weibo* de novembre 2017 à décembre 2018 [36]. Le corpus a ensuite été dédoublonné en fonction des premiers caractères communs à chaque message.

3.3.6 *Pré-traitement sur les doublons*

Le travail de dédoublonnage s'est effectué non seulement sur les articles mais aussi sur les *weibo*. Pour les articles, nous avons simplement supprimé manuellement des articles doublons. La production des doublons des *weibo* est principalement issue de la composition combinatoire de la requête de recherche, notamment les expressions régulières composées d'une chaîne des «.» [37] en tant que paramètre d'incertitude. Une requête composée de «.» risque de produire deux types de résultats : 1) des messages constitués des mêmes mots (caractères) mais pas de la même longueur ; 2) des messages de même longueur mais les mots (caractères) qui les composent sont différents. Dans le cadre de notre étude, les doublons résultent de la première situation : ils ont le même contenu et le même sens. Pour vérifier l'existence de ce type de doublons dans notre sous-corpus Weibo, nous avons procédé à l'opération suivante.

D'abord, un tri selon la longueur des messages *weibo* (des plus courts au plus

36. Nous n'avons gardé que dix mille *weibo* pour nos analyses.
37. Dans Jieba «.» signifie n'importe quel caractère.

longs) a été effectué dans l'ensemble du sous-corpus Weibo, cette opération nous permet de comparer la longueur des deux messages qui sont visuellement composés des mêmes caractères chinois, puis d'extraire les doublons. Il faut noter que deux messages au contenu identique sont tantôt de longueur identique, tantôt de longueur différente avec 1 à 3 formes d'écart (voir exemples suivants) :

1. Forme d'écart invisible sous forme d'un espace (en rouge)
   - Exemple :
     - 自打有了雾霾天, 就爱上了大风天
     - 自打有了雾霾天, 就爱上了大风天_ [38] (espace)

   - Traduction : (Je/On) commence à aimer les jours venteux quand il y a des jours de smog.
2. Forme d'écart visible sous forme d'un caractère chinois (en rouge) [39]
   - Exemple :
     - 雾霾中对健康有害主要是气溶胶粒子, 易引起鼻炎, 支气管炎等病症, 极重度雾霾还是心脏杀手, 颗粒污染物可能引发心肌梗
     - 雾霾中对健康有害主要是气溶胶粒子, 易引起鼻炎, 支气管炎等病症, 极重度雾霾还是心脏杀手, 颗粒污染物可能引发心肌梗死

   - Traduction : Les substances dangereuses du smog concernent principalement les particules d'aérosols, qui peuvent causer la rhinite, la bronchite et d'autres maladies, le smog grave pouvant même causer une cardiopathie, les particules polluantes pouvant entraîner un infarctus du myocarde.

Cette méthode de tri par longueur de phrase nous aide à enlever la plupart des doublons *weibo*. Pour le reste des informations bruitées, de même que pour celles présentées à la fin du pré-traitement des articles, nous expliquerons la démarche de traitement dans la partie suivante : 3.4 Dépouillement du corpus.

38. _ ici signifie un espace
39. Souvent phrase incomplète qui n'est pas finie.

3.4 Dépouillement du corpus

3.4.1 *Introduction*

En tenant compte des spécificités de la langue chinoise en matière d'écriture, de l'encodage, des signes de ponctuations, des lettres et des chiffres, etc., nous allons présenter dans cette partie le travail de traitement sur les données textuelles bruitées. En général, ces traitements s'axent sur trois grandes lignes : la gestion des éléments technodiscursifs (mots-dièse et arobases, liens html, etc.), la gestion des émoticônes et l'uniformisation des signes typographiques (les ponctuations, les lettres alphabétiques et les chiffres) du corpus.

3.4.2 *Gestion des éléments technodiscursifs*

Le travail de Paveau (2012, 2012b, 2013a) désigne les éléments technodiscursifs dans l'environnement du site web ou du réseau social. Il liste une série d'éléments technodiscusifs. Ces éléments incluent 1) les informations du compte de l'utilisateur telles que l'avatar de l'abonné, les noms et pseudos de l'abonné ; 2) la date de publication du message ; 3) le texte principal du message y compris la mention d'autres utilisateurs sous la forme de @+nom de l'abonné, les mots-dièse, et les liens HTML ; 4) la liste des opérations possibles signalées par des mots-consignes sous le texte, et assorties de leurs icônes : «*Ouvrir ou Afficher le média, Afficher la conversation, Voir le résumé ou Voir la photo, Répondre/Transférer/Favori/Plus*» (cf. figures 3.13 Exemple des mots-consignes de la page web du SOHU et figure 3.14 Exemple des mots-consignes de la page web du WEIBO). Notre tâche de gestion des éléments technodiscursifs consiste à identifier, puis éliminer ces éléments qui sont présents dans l'ensemble de notre corpus. L'élimination de ces éléments a pour objectif d'obtenir un corpus adapté au format d'importation dans l'outil textométrique, mais nous gardons toujours les statistiques et ces données pour les analyses sémiotiques du genre textuel du corpus dans les autres parties.

Nous avons repéré dans notre corpus trois types d'éléments technodiscursifs : 1) les liens HTML ; 2) les mots-dièse ou *hashtag*, utilisés dans les *weibo* pour créer des *topics* de discussions ; 3) les mots-consigne et les arobases @. Ceux-ci apparaissent dans les articles comme élément nécessaire de l'adresse mail, ou à l'intérieur du *weibo* pour mentionner une personne, de manière à répondre à sa question, à

l'inclure dans une conversation ou une activité, à l'inclure dans une conversation ou une activité, ou bien à transférer son message original. Nous avons éliminé ces éléments à l'aide d'expressions régulières.

3.4.2.1 Gestion des arobases @

Dans notre corpus les seules données à caractère potentiellement personnel sont les adresses mail et les pseudonymes des utilisateurs de WEIBO. Ces informations sont composées d'une suite de caractères alphanumériques ; les pseudonymes commencent par un @ (voir 3.8 Exemple d'un *weibo* publié sur le site WEIBO [40]). Pour protéger la confidentialité et anonymiser les *weibo*, nous avons enlevé les surnoms à l'aide de l'arobase initial et d'expressions régulières.

3.4.2.2 Gestion des mots-consigne

Nous repérons une série de mots-consigne dans la page de WEIBO. Ces mots-consignes se présentent souvent comme des expressions figées récurrentes (voir figures 3.13 Exemple des mots-consignes de la page web du SOHU et 3.14 Exemple des mots-consignes de la page web du WEIBO) :

- 转发 (transférer/transfert) ;
- 回复 (répondre) ;
- 评论 (commenter/commentaire) ;
- 分享到 (partager sur) ;
- 返回首页 (revenir à la page d'accueil) ;
- 回到顶部 (revenir en haut de la page)

Comme ces mots-consignes ne sont pas liés à notre sujet d'étude, nous les avons enlevés. Si les autres éléments sont plus faciles à éliminer automatiquement avec des simples expressions régulières, 10% des mots-consigne doivent faire intervenir la supervision manuelle, car chaque site a ses propres mots-consigne, il est donc possible que certains mots aient échappé à notre nettoyage.

Fig. 3.13 – Exemple des mots-consignes de la page web du SOHU

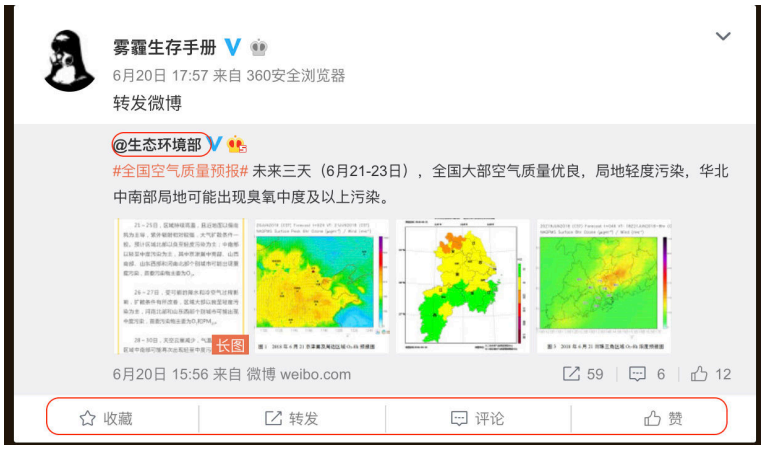


FIG. 3.14 – Exemple des mots-consignes de la page web du WEIBO

3.4.3 *Gestion des signes d'émoticônes*

« Une émoticône est une courte figuration symbolique d'une émotion, d'un état d'esprit, d'un ressenti, d'une ambiance ou d'une intensité, utilisée dans un discours écrit [41] ». L'utilisation des émoticônes s'avère une spécificité des *weibo*, où les signes d'émoticônes sont autorisés et souvent utilisés pour dynamiser le message ou l'échange entre les internautes. Nous avons remarqué trois types d'émoticônes dans notre corpus lors de notre recherche :

1. l'émoticône de structure : [AAA] constituée d'un mot expressif ou d'une interjection et de crochets. La légende de ces émoticônes apparaissent dans les forums ou les réseaux sociaux lorsque les images ne se chargent pas ;
   Phrase originale : 今天又有雾霾 [哭].
   Traduction : Encore la pollution de l'air aujourd'hui[snif].

2. l'émoticône Kaomojis : elle est constituée de signes de ponctuation, tels que **:)**, avec parfois des lettres latines en capitale, comme par exemple **:-D**.
   Phrase originale : #smogday# 嗓子疼 **:-(**.
   Traduction : #smogday# Mal à la gorge **:-(**.

41. Définition proposée par https://fr.wikipedia.org/wiki/%C3%89motic%C3%B4ne#cite_note\protect\discretionary{\char\hyphenchar\font}{}{}1.

3. l'émoticône graphique sous forme d'images GIF [42], statique ou animée : Bravo ! :  , .

L'élimination des émoticônes a été réalisée avec deux méthodes : suppression avec des expressions régulières pour les émoticônes du type 1 et 2 et suppression manuelle pour le type 3 à cause de leur format image. Cependant, nous avons gardé les statistiques de ces signes pour l'étude du genre textuel dans le chapitre suivant (cf. section 4.6.3 Variables sémiotiques du Chapitre 4).

3.4.4 *Uniformisation des signes typographiques*

La représentation informatique du système d'écriture chinois diffère de celle qui est utilisée par le système d'écriture occidental, autant dans l'aspect graphique que par le système d'encodage utilisé. Dans le système d'écriture occidental, chaque lettre et signe de ponctuation occupe un seul octet. À l'inverse, les caractères chinois occupent soit 1 (caractère demi-chasse (半角 en chinois simplifié)) soit 2 octets (caractères pleine chasse (全角 en chinois simplifié) [43]. Dans les fontes à chasse fixe, les caractères demi-chasse occupent la moitié de la chasse des caractères pleine chasse (voir figure 3.15 Lettres alphabétiques et chiffres pleine chasse (fullwidth) et demi-chasse (halfwidth)). Les caractères pleine chasse occupent deux colonnes chacun (appelés « cellule de rendu ») tandis que les caractères des écritures occidentales qui occupent 1 octet sont classés dans une colonne (appelé « demie cellule de rendu ») [44] (cf. tableau 2 Tableau des Formes à demi et pleine chasse). Dans notre corpus chinois, le mixe de codage sur 1 ou 2 octets sur les signes alphabétiques, numériques ainsi que certaines ponctuations rend les derniers non-identifiables par nos outils de traitement, car ces outils sont programmés pour supporter la représentation sur un seul octet. Cette hétérogénéité entraînera une interprétation erronée des textes. Par conséquent, il est nécessaire d'effectuer une opération d'uniformisation des codages afin d'homogénéiser ces éléments typographiques. Ces traitements permettent d'une part de rendre accessible notre corpus dans les outils techniques, de l'autre, de rendre le décompte statistique plus fiable. Pour ce faire, nous avons utilisé un

42. Le Graphics Interchange Format (littéralement « format d'échange d'images »), plus connu sous l'acronyme GIF, est un format d'image numérique couramment utilisé sur le web.

43. Définition proposée par https://zh.wikipedia.org/wiki/%E5%85%A8%E5%BD%A2%E5%92%8C%E5%8D%8A%E5%BD%A2.

44. Source d'information : https://fr.wikipedia.org/wiki/Formes_%C3%A0_demi_et_pleine_chasse. Page consultée en avril 2019.

script perl pour transcoder correctement ces signes typographiques.

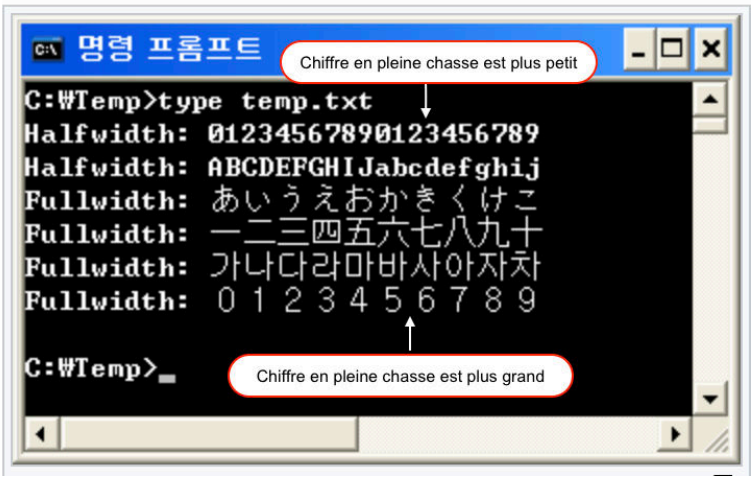


FIG. 3.15 – Lettres alphabétiques et chiffres pleine chasse (fullwidth) et demi-chasse (halfwidth)

3.4.4.1 Homogénéiser la dénomination de PM2.5/pm2.5/pm 2.5 ->PM2o5

En plus du mélange des chasses des signes typographiques, l'écriture de la virgule fractionnaire chinoise en un point se confond avec la ponctuation «.» désignant le point final. Afin d'éviter ce type de confusion, nous avons homogénéisé toutes les virgules fractionnaires chinoise en lettre « o » latine en minuscule, et changé toute forme de pm en PM majuscule sans espace.

3.5 Segmentation du corpus

3.5.1 *Segmentation du corpus + outil*

D'après ANDRÉ SALEM (1994), la segmentation du texte est l'opération qui permet de découper le texte en unités minimales ; nous parlerons de mots pour le chinois. Par rapport aux langues latines où l'on trouve un espace simple inséré entre les mots comme frontière d'unité, la langue chinoise est une langue *scriptio continua* qui ne possède que les ponctuations standards comme délimiteurs entre phrases, les mots sont par contre tous collés les uns aux autres. Ainsi, un travail de segmentation de notre corpus en chinois est nécessaire pour que les outils

d'analyse textométrique identifient les mots. Dans le cadre de notre travail, nous avons testé plusieurs segmenteurs pour la langue chinoise, et avons finalement sélectionné JIEBA [45] (« bégayer » en français) car c'est l'outil qui a démontré les meilleures performances. JIEBA est un outil segmenteur pour le chinois développé en Python et qui permet non seulement de segmenter le texte en unité de mot mais aussi d'attribuer une étiquette syntaxique (*tag*) aux unités segmentées. Afin de scanner dans la mesure du possible toutes les combinaisons de mots, puis d'en trouver la combinaison la plus probable en fonction de la fréquence des mots ainsi que des dictionnaires intégrés, l'algorithme implémenté dans JIEBA est basé sur une structure de dictionnaire de préfixes et sur un graphe acyclique dirigé (en anglais DAG, *directed acyclic graph*). En ce qui concerne les mots inconnus, l'outil a recours à un modèle basé sur un modèle de Markov caché (MMC) et sur l'algorithme de Viterbi [46].

Voici un exemple de segmentation effectuée par JIEBA :

- Phrase originale : 雾霾对健康的危害.
- Traduction : Les nocivités du brouillard de pollution sur la santé.
- Phrase segmentée par JIEBA : 雾霾 对 健康 的 危害 [47].

3.6 Annotation du corpus

3.6.1 *Rajout des dictionnaires personnalisés*

En plus du dictionnaire inclus par défaut dans JIEBA pour annoter le corpus selon la nature syntaxique (par exemple, **n** pour nom, **v** pour verbe, **adj** pour adjectif, etc.) ou pour détecter les entités nommées (**nt** pour nom d'établissement, **nr** pour nom de personne, etc.) [48], JIEBA peut segmenter le corpus en fonction d'un dictionnaire personnalisé ajouté par l'utilisateur. Nos dictionnaires personnalisés sont tous développés et créés à partir des documents disponibles dans la base de données SOGOU (搜狗细胞词库) à l'adresse https://pinyin.sogou.com/dict/. SOGOU est une large base de données en ligne disposant de dictionnaires spécialisés relatifs à de multiples domaines, tels que la géographie, le secteur médical, le secteur judiciaire, les néologismes, etc. Ces do-

45. Site officiel du JIEBA : https://github.com/fxsjy/jieba
46. Présentation de JIEBA sur : https://github.com/fxsjy/jieba
47. Ici, une version de segmentation sans *tag*.
48. Voir plus de catégories de *tags* dans le glossaire 2 Tags proposés par JIEBA.

cuments open-source sont fournis par des utilisateurs experts qui travaillent dans le domaine concerné. Les dictionnaires disponibles sur SOGOU sont au format «.scel » [49]. Puisque l'unique format de dictionnaire supporté par JIEBA est le format texte brut (.txt) en UTF-8, nous devons convertir les dictionnaires téléchargés via SOGOU pour rendre nos données compatibles avec cet outil. Pour ce faire, nous avons opté pour l'outil IMEWLCONVERTER [50], un logiciel qui permet de convertir les documents d'extensions « scel », « bdict », « dat », ou « uwl » au format « txt ». Les fichiers texte ainsi convertis par l'outil IMEWLCONVERTER contiennent un seul mot par ligne. Cependant, il arrive que la version convertie du dictionnaire déroge à cette règle à cause de la présence d'un élément résiduel, en l'occurrence la transcription en *pinyin* [51] du vocabulaire toponymique d'un dictionnaire (voir figure 3.16 Capture d'écran du résultat de conversion de dictionnaires « Subdivision de l'organisation territoriale de Chine.scel » en format du texte). Nous devons enlever l'élément résiduel et ne garder que le mot en lui-même, car un dictionnaire en entrée du segmenteur JIEBA doit nécessairement contenir deux éléments par ligne : un mot et son étiquette (*tag*) thématique, séparés par un simple espace (par exemple : 陕西　province). L'ajout des *tags* pour chaque mot sont effectué automatiquement grâce à une expression régulière lorsque cela était possible, et manuellement pour les mots restants. Par exemple pour annoter automatiquement des noms de maladies, nous avons utilisé le mot-clé : 病 [52] pour repérer les mots contenant ce caractère du corpus, puis les extraire en rajoutant un simple espace et le *tag* « disease » (pour maladie) (cf. Tableau 3.2 Tableau des catégories des *tags*).

Voici un exemple de segmentation effectuée par JIEBA :

- Phrase originale : 近日，北京雾霾严重，肺炎和呼吸道疾病发病率增高。
- Traduction : Ces derniers jours, la pollution de l'air de Beijing est grave, il en résulte que de plus en plus de gens ont atteint la pneumonie ou les maladies respiratoires.
- Phrase segmentée par JIEBA : 近日 **t** [53], 北京 **city** [54] 雾霾 **denowu-**

49. «.scel » est un format spécial adopté par les dictionnaires SOGOU.
50. Voir : `https://github.com/studyzy/imewlconverter.`
51. *Pinyin* est un système de transcription phonétique de la langue chinoise en Chine
52. En chinois, les noms de maladies contiennent en général le caractère « 病 » correspondant au suffixe «-pathie » pour la désignation des types de maladies en français.
53. Voir le glossaire 2 Tags proposés par JIEBA pour l'explication des abréviations des *tags*. Nous allons mettre en gras tous les *tags* pour les souligner.
54. JIEBA annote le corpus en fonction du dictionnaire toponymique rajouté. Une fois. Lors-

mai[55] 严重 **a**, 肺炎 **disease** 和 **c** 呼吸道疾病 **disease** 发病率 **Dns** 增高 **v**.

3.6.2 *Présentation des dictionnaires personnalisés*

Nous avons créé et ajouté trois dictionnaires personnalisés : « Dictionnaire de dénomination de brouillard de pollution », « Dictionnaire toponymique des villes/régions de Chine » et « Dictionnaire des maladies et symptômes ». L'identification des différentes dénominations du brouillard de pollution est réalisée à partir des annotations effectuées à l'aide du « Dictionnaire de dénomination de brouillard de pollution » (60 entrées, cf. section 3.6.2.3 Présentation du « Dictionnaire de dénomination de brouillard du pollution »). Ces dénominations constituant ensemble l'isotope /brouillard de pollution/ nous serviront à procéder à l'étude du repérage de thème (voir section 5.2 Méthode de travail pour l'identification du thème) au Chapitre 5. À l'aide du « Dictionnaire toponymique des villes/régions de Chine » (116,402 entrées, cf. section 3.6.2.1 Présentation du « Dictionnaire toponymique des villes/régions de Chine ») et du « Dictionnaire des maladies et symptômes » (31,295 entrées, cf. section 3.6.2.2 Présentation du « Dictionnaire des maladies et symptômes »), nous observons la distribution géographique et la distribution régionale des maladies/symptômes causés par le brouillard de pollution, et l'évolution temporelle de certains problèmes de santé. Dans la partie suivante, nous allons présenter de manière détaillée les trois dictionnaires, notamment sur la composition et le choix des étiquettes associées aux mots concernés. Un exemple concret qui combine les trois types de *tags* extraits de notre corpus sera donné à la fin.

3.6.2.1 Présentation du « Dictionnaire toponymique des villes/régions de Chine »

Le dictionnaire des toponymes que nous avons récupéré depuis la plateforme est exhaustif : avec 116,402 d'entrées, il englobe toutes les subdivisions de l'organisation territoriale de la Chine, de la province jusqu'au village naturel[56]. Comme

qu'une ville est citée et détectée, JIEBA associe le *tag* « city » au mot.

55. JIEBA annote le corpus en fonction du dictionnaire des maladies et symptômes rajouté. Lorsqu'une maladie est citée et détectée, JIEBA associe le *tag* « disease » au mot.

56. Les subdivisions de l'organisation territoriale de la Chine sont classées de manière hiérarchique du plus grand — les provinces (ou les régions autonomes) —, au plus petit — les villages naturels.

montré dans le tableau (voir Annexe 18 Tableau de subdivision de la structure territoriale de la Chine[57]), il y a cinq niveaux au total dans l'organisation territoriale de la Chine. Nous n'avons sélectionné que le premier niveau, c'est-à-dire le niveau provincial, qui représente 31 items au total[58], pour étudier la distribution géographique du brouillard de pollution ainsi que les problèmes de santé causés par ce dernier. Afin de simplifier l'appellation de différentes subdivisions de l'organisation territoriale, nous les appelons toutes « région » dans les parties suivantes. Il faut noter que notre groupe « région » appartenant au niveau provincial englobe non seulement des provinces, mais aussi des municipalités et des régions autonomes. Ainsi, en ce qui concerne les noms des *tags* des toponymes, nous avons gardé leur appellation initiale en anglais : « province » pour 22 provinces, « city » pour les 4 municipalités et « autoreg » pour les régions autonomes, le reste des subdivisions est étiqueté comme « NC » (non concerned).

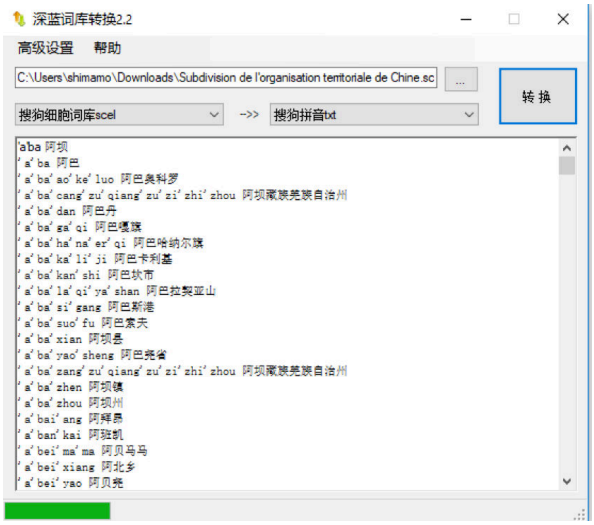


FIG. 3.16 – Capture d'écran du résultat de conversion de dictionnaires « Subdivision de l'organisation territoriale de Chine.scel » en format du texte

58. Les 31 régions sont composées de 22 provinces, 4 municipalités, et 5 régions autonomes. Par manque de données accessibles, la province de Taiwan et les 2 villes administrativement spéciales ne sont pas prises en compte, et nous allons concentrer nos recherches sur la Chine continentale.

3.6.2.2 Présentation du « Dictionnaire des maladies et symptômes »

En étudiant notre corpus, nous avons remarqué qu'un certain type de maladies est souvent associé à des symptômes correspondants, par exemple la maladie respiratoire (呼吸系统疾病) apparaît souvent avec des symptômes comme 咳嗽 (tousser/toux), 喘息 (haleter/halètement), 上呼吸道感染 (infection de l'appareil respiratoire supérieur), etc. Nous avons donc décidé de mettre les maladies et les symptômes dans un même dictionnaire avec des étiquettes différentes. En suivant le même processus de traitement que pour le premier dictionnaire des villes, nous avons téléchargé et converti le dictionnaire ICD-10 疾病编码 (Classification Internationale des Maladies ICD-10) et 各类基本医学词汇 (Glossaire des termes standards médicaux) sur SOGOU [59], et en avons fait notre deuxième dictionnaire : « Dictionnaire de maladies et symptômes » avec 31,295 entrées. Lors de la révision manuelle du résultat d'étiquetage de ces deux dictionnaires professionnels, nous avons constaté que certains vocabulaires quotidiens employés dans le corpus WEIBO sont absents de notre dictionnaire. Nous l'avons donc complété avec ces 105 termes. Quatre étiquettes ont été définies : « disease » pour les maladies, par exemple 呼吸系统疾病 **disease** (maladie respiratoire), « symptom » pour les symptômes, tel que 咳嗽 **symptom** (toux/tousser), « Dns » pour les terminologies médicales, par exemple 呼吸道 **Dns** (voie respiratoire), « Dv » pour les locutions verbales relatives, par exemple 看病 **Dv** (voir le médecin) (cf. Tableau 3.2 Tableau des catégories des *tags*).

3.6.2.3 Présentation du « Dictionnaire de dénomination de brouillard du pollution »

Contrairement aux procédures supervisées (automatique et manuelle) de la production des deux dictionnaires précédents, la création du « Dictionnaire de dénomination du brouillard de pollution » est entièrement manuelle. En plus des 6 mots-clés : 雾霾 (brouillard de pollution), 霾 (smog), 大气污染 (pollution atmosphérique), 空气污染 (pollution de l'air), PM2o5 et PM10 dont nous avons parlé plus haut, nous avons retrouvé 54 mots désignant le brouillard de pollution avec indépendamment 雾 (brume) ou 霾 (brouillard). Nous avons ainsi obtenu 60 mots dans le dictionnaire de dénomination du brouillard de pollution. Les

59. Ces dictionnaires sont disponibles sur le site https://pinyin.sogou.com/dict/detail/index/654.

intitulés des *tags* sont donnés soit en fonction du *pinyin* du mot, par exemple 大雾 (brume) est *taggé* avec «**denowu**», soit selon la traduction en anglais, 空气污染 **denopollu** (pollution de l'air).

Le tableau ci-dessous résume les trois catégories de dictionnaires, les *tags* définis pour chacun ainsi que des exemples concrets extraits du corpus.

TAB. 3.2 – Tableau des catégories des *tags*

Catégorie de mot-clé	Tags	Example
Terme de Maladie	**disease** : mot+**disease** terminologie médicale : mot+**Dns**[60]	肺炎 **disease** (pneumonie) 呼吸道 **Dns** (voie respiratoire) 看病 **Dv** (voir le médecin)
ville/province	city : mot+**city** province : mot+**province**	南京 **city** (nanjing) 河北 **province** (hebei)
dénominations de smog	brouillard de pollution : mot+**denowumai** brume : mot+**denowu** brouillard de pollution : mot+**denomai** pollution de l'air : mot+**denopollu** particule fine mot+**denopm**	雾霾 **denowumai** (brouillard de pollution) 大雾 **denowu** (brume) 霾 **denomai** (brouillard de pollution) 空气污染 **denopollu** (pollution de l'air) PM2o5**denopm** (Particule fine Ø 2,5 um)

3.7 Organisation du corpus et outils

Afin d'analyser et contraster les quatre sous-corpus de manière plus exhaustive, nous avons testé et finalement choisi quatre types d'outils qui sont performants dans leurs domaines. Ces outils s'acquittent chacun de tâches spécifiques : les deux outils textométriques (Trameur et Lexico5) se chargent des analyses textométriques telles que les segments répétés, le dictionnaire des mots (fréquence et spécificité), la distribution de certains mots selon des critères à définirà l'aide de fréquence ou de spécificité. Hyperbase, grâce à sa performance cartographique, se charge de générer des graphes et des histogrammes ; nous avons par exemple utilisé le « WORDCLOUD » pour générer un nuage de cooccurrents de mot-pivot, ainsi que la fonction « DISTRIBUTION » pour observer la distribution de

certains thèmes lexicaux envisagés selon des paramètres donnés. Quant à TXM, il nous est utile pour générer les graphes d'AFC en fonction des *tags* de chaque partition de corpus[61], par exemple la partition par année ou par nature de sous-corpus. Chaque outil prend en entrée un format de corpus différent en fonction de la tâche qu'il réalise, nous devrons adapter notre corpus à ces divers formats. Dans les parties suivantes, nous allons présenter comment extraire les métadonnées en fonction du type de corpus (article ou *weibo*, cf. 3.3 Nature et type du corpus), puis expliquer comment adapter la structure des corpus au format requis par les outils utilisés.

Tab. 3.3 – Nature et type du corpus

Corpus						
	Nature				Type	
	Ins	InsM	InfM	Profane	article	*weibo*
GOV	√				◯	
PEOPLE		√			◯	
SOHU			√		◯	
WEIBO				√		◯

3.7.1 *Extraction des métadonnées*

D'après Fiala (1994), « les comparaisons et les calculs statistiques ne s'établissent que dans un corpus limité, daté, lié à une thématique ». Dans cette partie, nous présentons le processus de l'extraction des métadonnées du corpus ainsi que les types de métadonnées associées aux articles et aux *weibo*.

Dans la phase d'aspiration des articles médiatiques à partir des trois sites traditionnels, nous avons conservé les adresses URL, car les éléments composants les URL de ces sites comprennent les métadonnées élémentaires d'un article.

Prenons un exemple de l'URL de la rubrique 服务 (service) du GOV :

L'adresse ci-dessus est découpable en trois segments délimités par «/», qui indiquent trois types d'informations :

– **site d'origine** : la racine `http://www.gov.cn` marque le site d'origine, c'est-à-direla provenance de l'article ;

61. La partition du corpus peut être effectuée selon différentes métadonnées que nous avons stockées dans un fichier « metadata.csv ».

FIG. 3.17 – Composants de l'URL du GOV

- **rubrique d'appartenance** : /fwxx/ est l'acronyme de fuwuxinxi (服务信息 l'information du service) en *pinyin* entouré par deux « barres obliques » (//). Ceci montre la rubrique d'appartenance de l'article ;
- **date de publication** : /2011-12/14/, suivi du premier bloc de « barres obliques », indique la date de la publication de l'article — le 14 décembre 2011.

Pour les *weibo* récents qui datent de fin 2017 à la fin du mois septembre 2018, nous avons extrait les métadonnées captées par le Weibo API R *Scraper*, telles que les variables « loc » (la région de publication) et « time » (la date de publication). En revanche, les *weibo* issus de la base de données BCC (notre deuxième source de données) ne sont pas accompagnés des métadonnées et ne nous délivrent donc aucune information de temps ou de localisation. Nous constatons néanmoins que tous les *weibo* provenant de la BCC sont produits avant 2017 ; nous avons donc défini une balise <an=av2017> comme étiquette de temps pour ces messages dans l'intention de contraster les *weibo* avant 2017 et après 2017. Nous définisson ainsi nos types de métadonnées comme ci-dessous :

TAB. 3.4 – Tableau des types de métadonnées du corpus

Nom de métadonnée	Fonctionnalité	Lexico5	Trameur	Hyperbase	TXM
id	identifiant	√	√	√	√
site	site d'origine	√	√	√	√
nature[1]	nature d'origine	√	√	√	√
type	type de données textuelles	√	√	◯[2]	√
rubrique	rubrique d'appartenance	√	√	◯[2]	√
an	année de publication	√	√	◯[2]	√
mois	mois de publication	√	√	◯[2]	√
date	date précise de publication	√	√	◯[2]	√
categorie	extraction de la catégorie générale à laquelle un certains groupe d'articles font partie[3]	√	√	◯[2]	√
loc	lieu de publication	√	√	√	√

[1] Nous avons quatre types de nature caractérisant chaque corpus : GOV : Ins ; PEOPLE : Mins ; SOHU : Minf ; WEIBO : Profane.

[2] L'outil Hyperbase a été utilisé principalement pour générer des nuages de cooccurrents de certains mots-pivots dans quatre sous-corpus, les critères temporels et catégoriels ne sont pas nécessaires dans l'outil.

[3] Dans la partie du corpus constitués par des articles journalistiques, il y existe des rubriques similaires qui peuvent être classées dans une même catégorie, par exemple : <rubrique=sohu_economy>, <rubrique=sohu_finance> et <rubrique=sohu_business>, une catégorie générale a été ainsi rajoutée, puis balisée comme <category=sohu_ECONOMY> pour attribuer une classification thématique à des articles correspondants.

3.7.2 *Format et balisage du corpus*

Dans le cadre de notre recherche, l'organisation du corpus constitue une étape cruciale, il s'agit de la mise en œuvre du balisage du corpus en fonction de l'outil adopté. Du fait que le format exigé par les outils textométriques se distinguent les uns des autres, nous avons préparé trois formats de corpus. La préparation du corpus se réalise à partir des métadonnées extraites montrées dans le tableau 3.4 Tableau des types de métadonnées du corpus.

- **Format TXM** : L'outil a besoin de deux types de fichiers pour faire des analyses : un fichier « csv » pour stocker les métadonnées de l'ensemble du corpus, et un fichier « txt » pour le corps textuel. Chaque ligne du fichier « csv » contient des items séparés par un délimiteur ;
- **Format Hyperbase** : Chaque fichier ne contient qu'un texte, et le titre de chaque fichier est composé de ses propres métadonnées pour chacune des 31 régions [62]. Ces métadonnées sont : le site d'origine, la région d'origine, l'identifiant numéral du texte et son type (article ou message *weibo*). Un délimiteur est utilisé pour séparer chaque composant, comme montré dans la figure 3.18 Exemple du nom du fichier en format Hyperbase. L'outil peut identifier les métadonnées du corpus à travers les composants des noms de fichiers ;
- **Format Lexico5/Trameur** : Ces deux outils ne demandent qu'un seul fichier XML ; toutes les métadonnées figurent alors dans le corpus même, à l'intérieur de balises délimitant chaque article ou message.

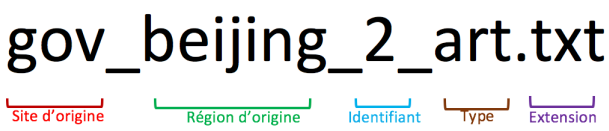


Fig. 3.18 – Exemple du nom du fichier en format Hyperbase

62. Du fait que l'Hyperbase est plus performant pour générer des graphes de distribution géographique, nous n'avons gardé que quatre types de variables (voir 3.4 Tableau des types de métadonnées du corpus) : « site », « id », « province/city/autoreg » et « art/message » pour lui.

3.7.3 *Partition du corpus*

3.7.3.1 Partition du corpus par an

L'organisation des trois formats de corpus, contrastive et partitionnée, nous permet d'explorer et contraster notre corpus en fonction de la nature (ex : Ins vs Profane), du type (article vs *weibo*), du temps (ex : annuel/mensuel), de l'aspect spatial (ex : Beijing vs Shanghai), ou encore d'une combinaison spatio-temporelle (ex : le graphe 1.1 Distribution régionale du brouillard de pollution en Chine de 2013 à 2018 [63] montre l'évolution temporelle et spatiale du Brouillard de pollution de Chine de 2013 à 2018). Dans la partie 3.6.2.1 Présentation du « Dictionnaire toponymique des villes/régions de Chine », nous avons mentionné notre objectif de partitionner le corpus par région ou par période pour effectuer des études spatio-temporelle, par exemple, pour observer la distribution spatiales ou l'évolution temporelle des problèmes de santé causés par le brouillard de pollution, et étudier s'il existe une corrélation entre les problèmes de santé et une certaine région concernée.

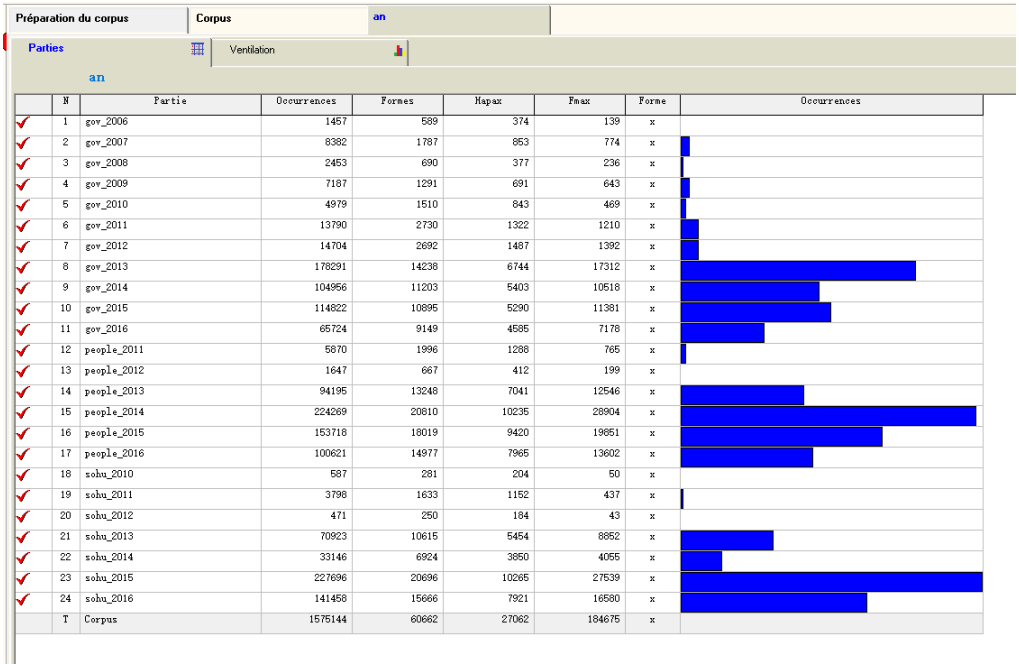
La partition temporelle du corpus par **an** est réalisée à l'aide de balises <an> prédéfinies dans le corpus. Les outils dont nous avons parlés en haut peuvent reconnaître ces balises et les interpréter comme un indice de temps. Prenons l'outil Lexico5 comme exemple, il nous suffit juste avant d'entrer dans l'analyse, de choisir dans l'onglet « Partition du corpus » le mot **an** comme clé, l'outil va découper le corpus selon notre besoin, comme le montre l'image suivante :

3.7.3.2 Partition du corpus par région

L'organisation du corpus par région a été réalisée à l'aide de l'outil Trameur. Prenons le corpus WEIBO comme exemple :

Tout d'abord, afin de limiter les choix des endroits cibles, nous n'avons choisi que les 31 régions dans le but de simplifier nos analyses et explications dans les parties *infra*. Ensuite, en prenant une ville/province/autoreg cible [64] comme requête de recherche dans l'onglet « Section » de Trameur, nous pouvons extraire tous les *weibo* qui contiennent cette métadonnées, puis regrouper les *weibo* qui proviennent de cette province [65].

64. Ici un exemple de « 河北 province »
65. Tous les articles et *weibo* ne disposant pas de référence géographique, nous n'avons utilisé



	N	Partie	Occurrences	Formes	Hapax	Fmax	Forme	Occurrences
✓	1	gov_2006	1457	589	374	139	x	
✓	2	gov_2007	8382	1787	853	774	x	
✓	3	gov_2008	2453	690	377	236	x	
✓	4	gov_2009	7187	1291	691	643	x	
✓	5	gov_2010	4979	1510	843	469	x	
✓	6	gov_2011	13790	2730	1322	1210	x	
✓	7	gov_2012	14704	2692	1487	1392	x	
✓	8	gov_2013	178291	14238	6744	17312	x	
✓	9	gov_2014	104956	11203	5403	10518	x	
✓	10	gov_2015	114822	10895	5290	11381	x	
✓	11	gov_2016	65724	9149	4585	7178	x	
✓	12	people_2011	5870	1996	1288	765	x	
✓	13	people_2012	1647	667	412	199	x	
✓	14	people_2013	94195	13248	7041	12546	x	
✓	15	people_2014	224269	20810	10235	28904	x	
✓	16	people_2015	153718	18019	9420	19851	x	
✓	17	people_2016	100621	14977	7965	13602	x	
✓	18	sohu_2010	587	281	204	50	x	
✓	19	sohu_2011	3798	1633	1152	437	x	
✓	20	sohu_2012	471	250	184	43	x	
✓	21	sohu_2013	70923	10615	5454	8852	x	
✓	22	sohu_2014	33146	6924	3850	4055	x	
✓	23	sohu_2015	227696	20696	10265	27539	x	
✓	24	sohu_2016	141458	15666	7921	16580	x	
	T	Corpus	1575144	60662	27062	184675	x	

FIG. 3.19 – Découpage des trois corpus par année avec Lexico5

3.8 Informations quantitatives du corpus

Nous disposons d'un corpus composé de quatre sous-corpus, 2556 articles numériques journalistiques et 118,898 *weibo* publics postés par les utilisateurs sur WEIBO, soit 10,122,515 items (formes+délimiteurs), 4,622,566 occurrences de forme, 117,000 formes (les mots différents) datant de 2006 à 2018. Ils se répartissent en quatre types de sous-corpus : Ins, InsM, InfM et Profane. Le sous-corpus Ins se divise en «*articles de presse, documents officiels, rapports*», etc. Le sous-corpus InsM est constitué pour la plupart des «*articles journalistiques des presses institutionnelles*». Si les textes Ins et InsM rapportent les affaires et les informations à propos de la politique du gouvernement chinois, les textes InfM et Profane englobent tout type d'informations, de la politique jusqu'au domaine privé.

que ceux qui disposent de cette information pour faire les analyses correspondantes

TAB. 3.5 – Informations quantitatives du corpus

Corpus	Nombre d'articles/de *weibo*	Nb d'occurrences	Nb de formes	Types de métadonnées [1]	Empan temporelle [1]
GOV	1095	516756	28234	9	2006-2016
PEOPLE	835	580437	39181	9	2011-2016
SOHU	626	478082	32629	9	2010-2016
WEIBO	118,898	3057252	88678	7	2013-2018
Totalité	2556 articles/118,898 weibo	4622566	117000	9	2006-2018

[1] cf. 3.4 Tableau des types de métadonnées du corpus.

3.9 Conclusion

Dans cette partie, nous avons présenté les méthodes de travail ainsi que les outils utilisés dans chaque phase, depuis le choix des sources jusqu'à la constitution du corpus. Nous avons obtenu un corpus composé de quatre sous-corpus prêts à être analysés dans les prochains chapitres.

Chapitre 4

Étude du genre textuel du corpus

4.1 Étude de la sémantique interprétative des quatre genres textuels

4.2 Introduction

Notre corpus est constitué de textes provenant de quatre types de sous-corpus contrastés (Ins, InsM, InfM et Profane) qui traitent d'un même sujet écologique. Nous posons comme hypothèse que les quatre sous-corpus partagent des thèmes communs mais ils se distinguent par leur interprétation sémantique. Ces différences sémantiques, qui sont actualisées sous forme de unités lexicales variées, ont pour source fondamentale la divergence entre différents sous-corpus en termes du genre textuel. Afin de vérifier notre hypothèse, nous nous pencherons sur les trois paliers mentionnés au Chapitre 2 section 2.3 Trois Paliers sémantiques et étudierons et interprèterons la relation sémantique entre le genre textuel (travail global au palier macrosémantique) et les thèmes manifestés par le corpus (travail local aux paliers méso- et microsémantique). Nous consacrerons deux chapitres pour réaliser ce travail de vérification (chapitres 4 et 5) : le premier, c'est à dire celui dont il est actuellement question, vise à étudier le genre textuel de chaque sous-corpus à travers une série de variables réparties sur deux niveaux — infratextuel et intertextuel ; le deuxième (soit le chapitre 5) est axé sur l'identification des thèmes et l'analyse sémantique des thèmes de chaque sous-corpus.

4.3 Choix des variables

Nous procédons à l'analyse du genre textuel du corpus au travers d'un jeu de variables. Dans le cadre de nos analyses, le choix des variables s'effectue sur les deux

niveaux : infratextuel et intratextuel. Plus précisément, au niveau infratextuel, l'analyse portera sur 1) l'interface du site pour observer le *time-line* (chronologique/antéchronologique) et la configuration du texte (fragmentaire/linéaire), 2) l'organisation hiérarchique de la page web au travers des variables hyperstructurelles (balises HTML) de trois sites web (Ins, InsM, InfM) [1] ; au niveau intratextuel, les études porteront sur les traits lexicaux, sémiotiques, rhétoriques, modaux et syntaxiques. Toutes ces variables choisies comme critères d'analyse sont soit issues des *tags* effectués par JIEBA (voir partie 3.6 Annotation du corpus et Annexe 2 Tags proposés par JIEBA), soit identifiées automatiquement par les outils Lexico5 ou Le Trameur. Il peut s'agir par exemple des variables rhétoriques de parallélisme, repérées avec le « segment répété » (voir partie 4.6.5 Variables rhétoriques), ou de traits sémiotiques déjà inclus dans le corpus, comme la ponctuation. Certains indices de variables sont calculés par nous-même, comme par exemple la longueur moyenne de mots et de phrases (cf. section 4.6.6 Variables syntaxiques). Voici le détail des variables que nous avons sélectionnées :

H

TAB. 4.1 – Variables caractéristiques du genre

Début du tableau							
Niveau	**Variable**	**Signe**	**Description**	**Ins**	**InsM**	**InfM**	**Profane**
Infra-textuel	**Variable-(Para-textuel)**	**TITL**	<H> (<h1>,<h2>,<h3>, etc.) : étiquette titre	-	+	+	◯ [2]
		TAL	/ / : étiquette tableau	+	+	+	◯
		LIST	<dl>/<dt>/<dd> : étiquette liste de définition			+	-
		COL	<color> : couleur de fond	+	+	-	◯
		ACC	/ : effet renforcé ou accentué			+	◯
	Variable-(Hyper-textuel)	**IMG**	<img> : logo, portrait [3]	+	+	-	◯
		HYP	<href> : lien interne	+	+		
			hyperlien/lien externe			+	+

1. Le sous-corpus Profane provient d'un site basé sur un langage informatique dynamique, il ne possède pas une structure figée, donc ne permet pas d'extraire les balises hiérarchiques.
2. En raison du langage dynamique utilisé par le sous-corpus Profane, nous n'examinons pas les variables hyperstructurelles de ce sous-groupe.
3. Ici, nous regardons seulement les images en logo et en portrait.

Continuation de tableau							
Niveau	**Variable**	**Signe**	**Description**	**Ins**	**InsM**	**InfM**	**Profane**
Intra-textuel	**Variable lexique**	**ABR**	Acronyme et Abréviation(j) [4]	+			
		CON	Connecteurs(c) : addition, cause, conséquence, conclusion, disjonction, opposition, etc.	+			
		ENG	Mots anglais(eng)				+
			Mots d'emprunt(nz)				+
		EXP	Terminologie politique(i/l)	+	+		
			Terminologie technique/scientifique(nz/q)			+	+
			Expressions proverbiales(i)		+		
		Nom	Nom d'état(ns+ 国 pays)	+		+	
			Nom d'établissement(nt)	+	+		
			Nom de personne(nr)			+	
			Toponyme(ns)	+			+
			Nom de localité(f)	+	+		
			Nom propre(nz)	+	+		
		Nomi	Transformation de verbe en nom(vn)	+	+		
		NEO	Néologie				+
		PRON	Pronom personnel : je/tu				+
			Nous : toi et moi		+	+	
			Il/Ils		+	+	
			Elle/Elles			+	+
			Pronom interrogatif(r : liste des mots interrogatifs)		+	+	
		TEM	Temps(t) : le futur	+			+
			le future proche	+			+
			le passé	+			+
			l'imparfait	+			+
			le présent	+			+
			le temps descriptif	+			+
	Variable modale	**MOD**	Modaux : énonciatif			+	
			Impératif (falloir, devoir, vouloir)	+	+		
			Argumentatif (enfin, donc, également, or, mais, pourtant, jamais, rien, etc.)	+	+		
			Instructif		+	+	
			Exclamatif (onomatopée)				+
			Interrogatif(?)				+

4. Les tags syntaxiques sont définis et étiquetés par le tokenizer JIEBA. voir http://sighan.cs.uchicago.edu/bakeoff2005/data/pku_spec.pdf.

Continuation de tableau							
Niveau	**Variable**	**Signe**	**Description**	**Ins**	**InsM**	**InfM**	**Profane**
	Variable sémiotique	**PON**	Deux points	+	+		
			Crochets, guillemets, parenthèse, slashs	+	+	+	
			Point d'exclamation/d'interrogation/de suspension				+
			Répétition de point d'exclamation (!!!) /d'interrogation (???) /de suspension (……)				+
			Combinaison de points d'exclamation(!?) /d'interrogation(?!) /de suspension (… !), etc.				+
		EMO	Emoticône				+
		TEC	@, #				+
	Variable rhétorique	**RHT**	Rhétorique : parallélisme	+	+		
		MET	Métaphore	+	+		
		HOM	Homonyme			+	+
	Variable syntaxique	**LON**	Longueur de phrase : courte				+
			Moyenne			+	
			Longue	+	+		
			Subordination	+	+		
La fin du tableau							

4.4 Démarche et outils d'analyse

L'analyse des variables se fait à deux procédés — qualitatif et quantitatif. En ce qui concerne le procédé qualitatif, nous analyserons les variables à l'aide des composantes sémantiques suivantes : 1) tactique : elle permet de définir des rythmes sémantiques. Ces rythmes, (non) séquentiels ou (non) linéaires, marquent l'ordre de la production des unités sémantiques ; 2) dialogique : elle concerne les acteurs et le positionnement énonciatif représentés par les pronoms personnels ; 3) dialectique : elle est composée de trois types d'éléments — a) éléments descriptifs et argumentatifs, b) éléments délimitant des intervalles temporels et le déroulement aspectuel, et c) éléments marquant les évaluations modales de la vérité (vrai/faux) et celles thymiques (positif/négatif). Quant au procédé quantitatif,

nous allons effectuer trois types de calculs pour chaque descripteur [5] :

- la fréquence de chaque variable tous les 100 000 mots ;
- la fréquence relative d'un paramètre dans un sous-corpus par rapport à l'ensemble des quatre sous-corpus ;
- l'indice de la spécificité de chaque type de trait

Dans la partie qui suit, nous allons d'abord expliquer, avec des exemples concrets tirés de chaque sous-corpus, la composante sémantique à laquelle appartient chaque variable, puis examiner les comportements linguistiques de ces variables dans chaque genre de sous-corpus. À la fin, nous présenterons un bilan sur les convergences mais aussi les divergences des paramètres caractérisant chaque genre textuel.

4.5 Étude des variables au niveau infratextuel

Dans cette section, nous nous intéressons aux variables infratextuelles constituées par des éléments paratextuels et hypertextuels [6]. La notion du paratexte est défini par Genette en 1987 comme « un seuil entre le texte et le hors-texte, zone indécise entre le dedans et le dehors » qui a pour fonction principale de « rendre présent [le texte], pour assurer sa présence au monde, sa réception et sa consommation ». Les éléments paratextuels englobent titres, sous-titres, noms d'auteur, indications génériques, illustrations, quatrièmes de couverture, dédicaces, notes de bas de page, correspondances d'écrivains, etc. D'après Malrieu (2004), il est important d'inclure dans la description du genre la combinaison des variables sémiotiques, dont les variables hypertextuelles. Par exemple, les tableaux, le gras et l'italique font partie des consignes d'interprétation sémantique de portée variable qui sont normées par le genre.

Les variables infratextuelles sont constituées principalement par des étiquettes HTML. Ces variables donnent des indications fondamentales sur la structure des textes, et constituent aussi de précieux indices stylistiques. [7]. Le détail des

5. Sauf les calculs pour les variables syntaxiques qui sont effectués par nous-même, tous les autres calculs statistiques sont effectués et outillés par les logiciels textométriques mentionnés dans la partie 2.4.1 Textométrie. Les résultats de calculs seront présentés sous forme graphique ou sous forme de tableau.

6. Le détail des deux catégories d'éléments est présenté dans le tableau 4.1 Choix des variables.

7. Pour rappel, nous analysons le code HTML de tous les corpus, sauf pour WEIBO, dont le code source est dynamique

variables est donné dans le tableau 4.1 Choix des variables). Afin de mieux comprendre la structure hiérarchique et le style qui sont propres à chaque site, nous examinerons le code source d'une page par type de sous-corpus, et observerons les phénomènes spécifiques au genre textuel de chaque sous-corpus.

En suivant l'ordre de l'emplacement [8] des éléments infratextuels (du haut en bas et de gauche à droite), nous allons analyser et contraster les composants représentatifs de chaque genre textuel.

4.5.1 *Caractéristiques infratextuelles de l'Ins et de l'InsM*

Du fait que le sous-corpus InsM rejoint son équivalent Ins sur plusieurs aspects, nous examinons donc ces deux genres textuels ensemble pour discuter de leurs caractéristiques globales au niveau infratextuel.

- Sur l'organisation hiérarchique, l'Ins et l'InsM suivent une structure hiérarchique classique balisée en <h>, <ol>, <ul>, <li>. À cette hiérarchie élémentaire s'ajoutent dans ces sites plusieurs autres étiquettes qui viennent renforcer leurs aspects structurés et organisés. Par exemple, on dénombre plusieurs niveaux de <h> numérotés ; de cette façon, <h1>, <h2>, <h3> marquent plusieurs niveaux et plusieurs tailles de titres et sous-titres.
- Le logotype balisé en <type= "image/x-icon">, placé en haut à gauche de la page web, correspond au logo identitaire du site. Pour les deux sites Ins et InsM, ils utilisent tous l'icône représentative du pays : l'emblème national et la place de *Tian An Men*.
- Les images des dirigeants chinois sont délimitées par la balise <img> dans les deux sites. Les portraits statiques des dirigeants chinois figurent en haut au centre de la page, avec un fond de couleur vive, souvent rouge et jaune, qui rappelle la composition des couleurs du drapeau national de la Chine. La taille des images est assez importante qui peut s'étendre sur toute la largeur de la page (avec l'attribut "pannel-image").
- Les rubriques thématiques sont délimitées par <ul> . Pour Ins, les thèmes des rubriques sont assez homogènes dont la plupart sont relatifs à la politique. Par contraste, les sujets abordés des rubriques du InfM sont

8. L'emplacement des éléments infratextuels occupe une place très importante, car les genres diffèrent les uns des autres en fonction du privilège de l'emplacement des éléments compositionnels.

beaucoup plus variés. En plus des actualités politiques, il rapporte des nouvelles économiques, écologiques, éducatifs, culturelles, etc.

- Pour Ins et InsM, la barre de navigation est consacrée exclusivement à la recherche de contenu intérieur au site.
- Les moyens de contact avec les utilisateurs sont stockés par <ul> . En tant que sites politiques, Ins et InsM mettent en place un dispositif qui permet aux visiteurs d'écrire un message à l'attention du Premier Ministre de la Chine ou du dirigeant local (régional, par exemple).
- Le corps des textes suit un ordre antéchronologique, c'est-à-dire du plus récent au moins récent. Ce qui permet de mettre en avant les actualités politiques et celles des dirigeants chinois.
- Les sites institutionnels privilégient les liens internes et les redirections vers d'autres sites institutionnels. Ainsi, sur la page d'InsM, nous trouvons des liens vers d'autres sites institutionnels en mode ruban roulant, juste après la zone de contact :

En résumé, nous abordons d'abord les points communs des deux sites. Au niveau de l'organisation et de la configuration globale des deux sous-corpus institutionnels, ceux-ci montrent une disposition mixte associant la structure **séquentielle** et **fragmentée**. Cette mise en page renforce leur identité politique. La structure séquentielle et le time-line antéchronologique permettent de mettre en avant les actualités politiques les plus récentes et importantes des dirigeants chinois. Les blocs fragmentés facilitent la centralisation des images visibles des dirigeants chinois. Le logotype de l'emblème national, le fond de couleur rouge, des liens internes et externes institutionnels sont les marques les plus saillantes du sous-corpus institutionnel. En ce qui concerne les différences des deux sous-corpus institutionnels, on notera qu'en répondant à l'exigence de production de normes standardisées, Ins suit un ordre strictement hiérarchique, qui accentue l'uniformité identitaire de la politique institutionnelle. De plus, le fait que « la barre de navigation » et « les moyens de contact » donnent exclusivement accès à l'intérieur du site accentue son caractère autocentré. Par contraste, le site InsM aborde des thèmes plus variés que son équivalent en raison de son rôle de média. La disposition du blog et du forum permet aux lecteurs une interaction plus dynamique entre eux.

4.5.2 *Caractéristiques infratextuelles de l'InfM*

Nous analysons les caractéristiques infratextuelles du site informel-médiatique en suivant le même ordre de l'emplacement des variables.

- En plus des balises classiques, telles que <h>, <ol>, <ul>, <li> dans l'organisation hiérarchique, nous avons repéré des étiquettes comme <dl>, <dt>, et <dd>, marquant les listes de définition, utilisées pour renforcer l'aspect structuré du site ;
- Le logotype balisé en <type="image/x-icon"> du site InfM est placé en haut à gauche du site sous forme du logo identitaire :
- A la différence du site Ins dans lequel la taille des images des dirigeants présentées dans la page d'accueil est importante, le site informel-médiatique contient des images statiques plus petites. Quant à la position de ces images, celles-ci se situent soit dans les deux côtés soit dans les coins de la page. En plus des illustrations du texte, le site InfM se sert de la capture d'écran d'une vidéo balisée en <video-focus-pic> ou des images des stars en ruban roulant pour gagner l'attention des utilisateurs :
- Les rubriques thématiques sont beaucoup plus variées dans le site informel-médiatique, qui peuvent aller de la politique au sport, de l'économie au divertissement ;
- Par rapport aux sites institutionnels, où la barre de navigation est consacrée exclusivement à la recherche de contenu intérieur au site, le SOHU permet d'effectuer des recherches en dehors du site, et qui fonctionne comme le navigateur Google ;
- Les moyens de contact sont disponibles aux lecteurs sous forme du cadre réservé aux commentaires. Ce dispositif, dynamisé par l'ajout des émoticônes, favorise l'interaction ponctuelle entre le producteur des textes et les lecteurs ou entre les lecteurs eux-mêmes ;
- Les site InfM contiennent un grand nombre de liens vers des plateformes ou sites externes, qui enrichissent leur contenu.

L'organisation structurelle de la page d'accueil de SOHU est fragmentaire. Cette mise en forme a pour but de maximiser et de diversifier les types d'informations dans sa page d'accueil, en espérant attirer plus de lecteurs et faire augmenter le taux de fréquentation. Au niveau du contenu et du style du SOHU, nous pouvons résumer ses caractéristiques avec trois mots-clés : **multidimen-**

sionnalité, **hypertextualité** et **variété**. La multidimensionnalité est due à sa vocation à diffuser les informations en multidimensions (politique, économique, technique, sport, éducation, etc.) façon d' inciter les lecteurs ou les utilisateurs des secteurs multiples. Ensuite, son hypertextualité se caractérise par la possibilité de se déplacer entre les pages et les sites internes et externes. Sa variété s'avère par la diversité des balises stylistiques et par la multiplicité des images (images statiques ou images dynamiques en *gif*), par exemple on se sert des étiquettes ou pour mettre l'accent sur certaines informations importantes. Les caractères de **multimodalité** et d'**interactivité** commencent à émerger dans le site informel-médiatique grâce au dispositif du blog et du cadre de commentaires réservés aux lecteurs. On peut intégrer un enregistrement audio ou une vidéo dans son blog, ou bien ajouter des émoticônes en gif dans son commentaire.

4.5.3 *Caractéristiques infratextuelles du Profane*

Du fait que le site WEIBO emprunte des codes sources dynamiques ayant des origines différentes (ce qui n'est pas le cas des trois sites précédents), nous restons sur la configuration et la structure d'apparence du site pour étudier ses caractéristiques au niveau infratextuel. L'analyse suivra le même d'ordre des emplacements comme ci-dessus.

- L'organisation de la page du site WEIBO est mixte de style fragmentaire et linéaire. Globalement les différents blocs d'informations sont formés de manière fragmentaire. Par contre, la publication des *weibo*, qui fournit le contenu principal du site, est placée au centre du site sous forme linéaire. Les *weibo* publiés suit en général un ordre antéchronologique. En plus de cette règle, on classe les *weibo* également selon la popularité du sujet dont ils traitent ;
- Le logotype se situe en haut à gauche du site, correspond au logo identitaire du site :
- On utilise davantage les images animées, multicolores et de petite taille, de nombreuses vidéos, et certaines audios dans le site. Ces éléments permettent de dynamiser les pages pour contrebalancer l'état statique des textes et ainsi d'attirer l'attention des utilisateurs :
- Les sujets abordés, qui portent sur des domaines hétérogènes, sont encore

plus riches et variés que ceux proposés par SOHU. Ces dernières années, les sujets qui intéressent le plus les utilisateurs ont tendance à relever du divertissement et des faits-divers des vedettes ;

– La barre de navigation est consacrée exclusivement à la recherche de contenu intérieur au site pour fouiller des *weibo* relatifs ;
– WEIBO privilégie le rôle de la communication entre les utilisateurs. Le cadre réservé aux commentaires, qui se trouve au milieu juste au dessous de chaque *weibo*, consacre une large couverture dans le site WEIBO. Le site met à disposition une boîte aux lettres en ligne permettant de privatiser une communication entre le producteur du *weibo* et les lecteurs ou bien entre les « amis » ;
– À l'instar du SOHU, le site WEIBO donne la possibilité aux utilisateurs d'ajouter des liens externes dans leur texte pour enrichir leur contenu.

Par les réaménagements et l'hybridation des genres numériques médiatiques qu'il opère, le genre Profane offre une grande diversité architecturale, thématique et procédurale. Par rapport aux trois genres précédents, le genre Profane adopte une tactique spéciale : la structure, globalement fragmentaire, est caractérisée par la centralisation linéaire de la section principale des *weibo* et des commentaires publiés par les utilisateurs. Le sous-corpus Profane partage avec le sous-corpus InfM toutes les caractéristiques montrées dans le tableau 4.2 Caractéristiques génériques des quatre genres textuels, de manière incrémentale. Les singularités du genre Profane se manifestent par quatre aspects : la **rapidité**, la **multidimensionnalité**, la **multimodalité** et l'**interactivité**. Pour attirer davantage d'utilisateurs, la rapidité de la mise à jour et la multidimensionnalité s'avèrent indispensables. Grâce à son dispositif « user-generated-content » et en suivant la logique de valorisation des informations en temps réel, les informations publiées sur le site WEIBO se renouvellent très rapidement. La multimodalité se manifeste par la possibilité d'intégrer du son, des clips vidéos et des images animés aussi bien dans les publications que dans les commentaires. Enfin, l'interactivité—la caractéristique saillante du genre Profane, résulte de son dispositif divers qui est en faveur de la communication dynamique et en temps réel : par commentaire ou par message privé.

4.5.4 *Récapitulatif des caractéristiques de chaque genre textuel au niveau infratextuel*

Nous résumons les caractéristiques singulières de chaque genre textuel dans le tableau suivant :

TAB. 4.2 – Caractéristiques génériques des quatre genres textuels

Caractéristique générique	Ins	InsM	InfM	Profane
Fragmentaire	√	√	√	√
Antéchronologique	√	√	√	√
Identitaire	+	+	-	+
Multidimensionnel	-	+	+	+
Hypertextuel	-	-	+	+
Interactif	-	-	+	+
Multimodal	-	-	-	+
Rapide	-	-	-	+

4.6 Étude des variables caractéristiques au niveau intratextuel

4.6.1 *Introduction*

En plus des traits infratextuels, les variables intratextuelles sont aussi prises en compte pour décrire les caractéristiques récurrentes des quatre genres textuels. Les descripteurs intratextuels sont répartis en cinq catégories : lexicale, sémiotique, modale, rhétorique et syntaxique. Ils sont actualisés dans les quatre genres textuels au moyen de diverses variables que nous présenterons et détaillerons dans les sous-sections suivantes. Un récapitulatif des variables analysées au niveau intratextuel est donné dans le tableau 4.1 Choix des variables.

4.6.2 *Étude lexicales*

4.6.2.1 Analyse de l'acronyme et de l'abréviation

En chinois, les mots abrégés et les acronymes sont formés des signes arbitraires ou des lettres (caractères) initiales des mots de l'expression. Afin d'explorer l'utilisation de l'acronyme ou l'abréviation dans les quatre genres des sous-corpus,

nous avons recensé la fréquence des mots étiquetés comme tels[9]. Le résultat montre que les textes institutionnels, pour introduire des établissements publics, des pays ou des villes/régions, emploient davantage les abréviations (674 sur 10 000 mots, contre moins de 400 dans les autres corpus) sur trois types principaux : des verbes nominalisés et des terminologies spécifiques (voir *infra*). Du fait que ces noms propres d'établissements, toponymes et terminologies spécifiques sont fréquents dans les textes institutionnels, ils deviennent en quelque sorte des expressions propres à ces institutions. On se sert d'abréviations afin de gagner du temps et de l'espace et de rendre ces mots plus faciles à mémoriser.

- L'abréviation des substantifs :
  - sur les établissements publics : 环保部 pour 环境保护部 (Ministère de la Protection de l'Environnement) ou 发改委 pour 发展改革委员会 (Comité national de Développement et de Réforme) ;
  - sur les acronymes des toponymes, notamment les toponymes provinciaux : 冀 pour 河北 (Province du Hebei), 沪 pour 上海 (Shanghai) ;
  - sur les noms d'état : 中美 pour 中国 (Chine) et 美国 (États Unis), 中俄 pour 中国 (Chine) et 俄罗斯 (Russie) ;
  - sur les noms des dirigeants ou gouvernants : 外长 pour 外交部长 (ministre du Ministère des Affaires étrangères) ;
- L'abréviation des verbes nominalisés : 煤改气 (changement de charbon en gaz), 环保 pour 环境保护 (protection environnementale) ;
- L'abréviation des mots anglais : GDP pour 国内生产总值 (Produit Intérieur Brut) ; pm2,5 pour *Particulate Matter* .

4.6.2.2 Analyse des conjonctions

Deux types de conjonctions sont pris en compte[10] :

- Les conjonctions de coordination peuvent marquer l'union (et), l'opposition (mais, pourtant), l'alternative (ou), la négation (ni, jamais), la conséquence (donc) ou la conclusion (enfin, ainsi, finalement).
- Les conjonctions de subordination servent à introduire une proposition subordonnée conjonctive. Par exemple : si, lorsque, comme, quand, pour que, afin que, etc.

9. Les deux types de données partagent le même tag «**j**».
10. voir plus de détails sur le site https://www.espacefrancais.com/les-conjonctions-de-coordination-et-de-subordination/.

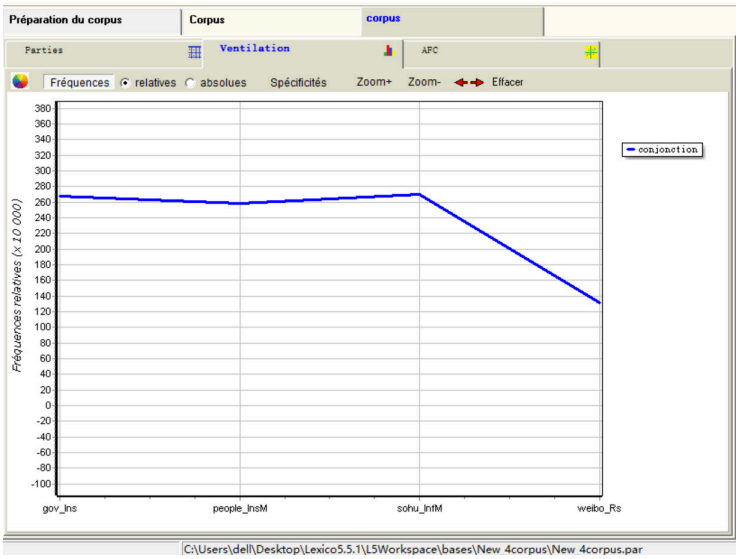


FIG. 4.1 – Conjonctions dans les quatre sous-corpus [11]

Le graphique ci-dessus montre les fréquences relatives des conjonctions détectées dans l'ensemble du corpus selon leur étiquette morphosyntaxique [12]. Nous pouvons observer que le genre InfM, à l'instar de l'Ins et de l'InsM, en tant que discours médiatique, utilisent beaucoup plus de conjonctions que Weibo. En effet, les sites traditionnels médiatiques sont tenus de respecter des normes grammaticales plus strictes que les réseaux sociaux : les conjonctions y sont utilisées pour mieux structurer les informations et garantir un certain niveau de langue. On observe ainsi que les phrases sont plus longues 4.6.6 Variables syntaxiques [13]. Inversement, contraint par une limitation du nombre de mots, un *weibo* présente souvent un contenu concis et dans un style oral, donc un contenu moins formel qu'Ins. Pour ces deux raisons, les conjonctions sont marginales dans le sous-corpus Profane.

4.6.2.3 Analyse des mots d'emprunt

Les variables des mots d'emprunt sont principalement composées de mots anglais. Quelques-uns proviennent d'autres langues étrangères telles que le japonais,

11. Le calcul de la distribution des conjonctions est sur l'indice de la fréquence relative
12. Les conjonctions sont annotées par JIEBA avec la lettre « c ».
13. Le résultat du calcul de comparaison de la longueur moyenne des phrases correspond à l'analyse de l'utilisation et de la distribution des conjonctions.

le coréen ou le russe. Les emprunts à l'anglais sont généralement utilisés pour introduire des expressions ou des terminologies techniques. Par exemple, dans les textes médiatiques (Ins, InsM et InfM), nous avons repéré des terminologies techniques telles que NO2, O3, CO, ou AQI ; ces termes y sont utilisés pour transmettre des connaissances scientifiques ou techniques à propos du brouillard de pollution. Cette même terminologie technique est utilisée dans les textes profanes pour représenter la qualité de l'air d'une région ou d'une ville (cf. Exemple 5.5.4 Études sémantiques du thème 1 dans le sous-corpus Profane). Toutefois, nous trouvons dans le sous-corpus Profane un autre type de recours aux emprunts : il arrive que les utilisateurs chinois empruntent un ou plusieurs mots qui sont familiers au public local, par exemple le signe japonais « の » ayant un sens similaire que « de » (voir Exemple 1 suivant), dans l'objectif d'apporter à leur texte une connotation « chic », ou de vendre un produit soi-disant importé du pays concerné (ici en l'occurrence un produit japonais).

- **Exemple 1** : 鲜肌の谜人胎盘鲜活洗面奶 快速進入肌膚底層，深度清潔肌膚内 PM2o5 污染物 ;
- **Traduction** : *Le gel nettoyant de marque* 鲜肌の谜 *(le secret de peau tendre) est un nettoyant bien frais, qui peut pénétrer rapidement la peau en profondeur, et élimine parfaitement les polluants PM2,5.*

4.6.2.4 Analyse du temps

Après avoir collecté les marqueurs temporels annotés « t », nous avons obtenu 1205 variables réparties en six types de termes temporels. Ces termes nominaux ou adverbiaux localisent un événement ou un état dans le temps, y compris le présent, le futur, le passé. En outre, il existe aussi des termes liés au temps descriptif[14]. En regroupant les termes temporels dans la graphe de ventilation par l'indice de spécificité, nous avons identifié les caractéristiques des quatre genres textuels sur le plan temporel : Ins utilisent davantage le vocabulaire descriptif pour marquer le futur ou le passé. On observe ainsi l'apparition fréquente des marqueurs de « saison ou mois hivernal » dans la liste des mots temporels descriptifs, ce qui s'explique par ce que le pic de pollution annuel apparaît durant la période hivernale. Nous nous avons repéré dans le sous-corpus institutionnel un

14. Le temps descriptif est constitué de mots qui portent une indication de temps vague, par exemple 秋冬季 (les saisons automnale et hivernale), 冬天里 (dans les jours hivernaux).

usage significatif de mots liés au futur (la fréquence des marqueurs du futur met en évidence une spécificité positive), comme par exemple « 未来 (futur)» doté d'une fréquence de 322 sur 633 au total, et « 将 (marqueur de futur)» d'une fréquence de 2163 sur 4627 au total. Cette observation est à mettre en lien avec les propositions de mesures et réformes politiques que nous trouvons fréquemment dans ce genre de sous-corpus. Cette caractéristique de l'Ins correspond à la zone distale (transcendante) de l'absence définie par Rastier (2001) : « La spécificité des langues et des cultures humaines réside dans leur capacité à parler de ce qui est absent, ce qui n'est pas là ».

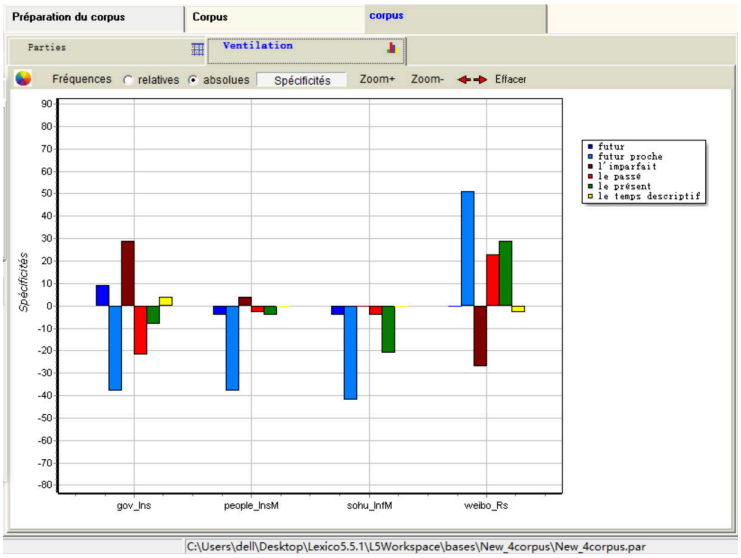


Fig. 4.2 – Caractéristiques de la temporalité des quatre genres textuels

Par contraste, les textes profanes adoptent essentiellement un discours d'ordre de l'*exposé*. Il se caractérise par des mots temporels marquant le présent ou le passé ; ce dernier figure en spécificité positive dans nos résultats, ce qui suggère que les internautes tiennent souvent un discours projectif, et expriment peut-être leurs souhaits ardents de résoudre le problème de *wumai*. Le genre Profane se situe donc dans la « zone identitaire de coïncidence », et se caractérise par un usage privilégié du temps du présent.

De nature médiatique, les textes InsM et InfM sont publiés pour rapporter et analyser des faits ou des nouvelles de la société, ainsi, on y trouve les termes temporels de toutes sortes.

4.6.2.5 Analyse des pronoms personnels

À l'aide des statistiques de fréquence de différents pronoms personnels(4.3 Pronoms personnels dans les quatre genres textuels [15]), et du graphe AFC(4.4 Exemple de l'utilisation des expressions proverbiales et des terminologies politiques dans un texte institutionnel), qui montre le rapprochement ou l'éloignement des mots personnels par rapport au discours d'appartenance, une nette distinction s'est établie entre le genre institutionnel et le genre Profane.

D'abord, la distinction se manifeste au niveau de la présence du pronom personnel ou du pronom impersonnel.

La nature interactive du genre Profane explique l'utilisation plus conséquente des pronoms personnels de la 1ère et 2ème personne au singulier dans ce genre de sous-corpus. Par ailleurs, nous avons également relevé le recours à la 3ème personne du féminin singulier dans les textes profanes. Ce qui montre que le sous-corpus Profane accorde une attention particulière au groupe féminin, en le distinguant des autres groupes avec les pronoms personnels 她/她们 (elle/elles). Cet aspect est parfaitement concordant avec nos résultats d'analyses sur les publics concernés (cf. Partie 5.6.3 Variation des publics sensibles) par le brouillard de pollution, où les textes profanes prêtent plus grande attention aux 妈妈 (maman), 母亲 (maman), 女性 (femme), 孕妇 (femme enceinte). Par contraste, les textes institutionnels, privilégient les pronoms impersonnels comme le 人们 ("on" en chinois)» ou « il » en mode énonciatif injonctif pour ordonner ou faire agir le peuple. Sur toutes les mesures préventives proposées, il y a un type de slogan qui est introduit par « il faut » ou par « on doit », par exemple « 治理雾霾需从我做起 » (Il faut commencer par soi-même pour lutter contre le brouillard de pollution)», « 我们必须高度重视并采取行动 » (On doit y attacher une grande importance et réagir immédiatement). La récurrence de ces slogans fait monter la fréquence du pronom personnel « il » et celui du « on ».

Tab. 4.3 – Pronoms personnels dans les quatre genres textuels [16]

Pronom personnel	Ins	InsM	InfM	Profane
Je	49	151	227	1527
Nous	136	355	319	164
Je+Tu	1	14	2	7
Tu	27	100	134	740
Vous	2	12	9	96
Il	80	118	137	129
Elle	13	18	44	85
Ils	4	50	58	26
Elles	1	2	1	20

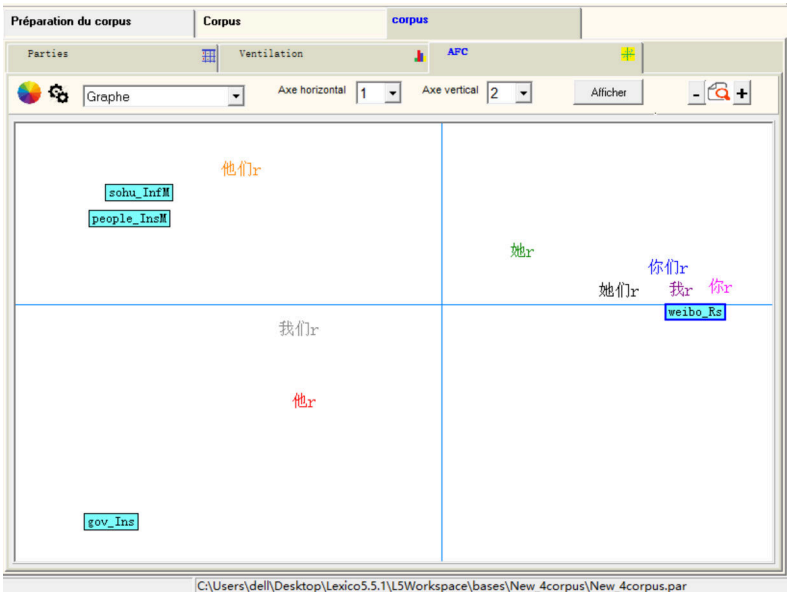


Fig. 4.3 – Pronoms personnels dans les quatre genres textuels

4.6.2.6 Analyse des collocations et des terminologies

Les normes discursives et le formalisme d'expression induisent l'utilisation de formes consensuelles telle que le 成语/chengyu/, qui désigne des expressions pro-

16. Le résultat statistique est basé sur la fréquence de chaque pronom personnel dans chaque genre textuel.

verbiales ou locutions[17]. Celles-ci sont formées en général de quatre caractères lexicalement indivisibles pour exprimer un sens unique. Dans le cadre de notre recherche, de nombreuses expressions proverbiales apparaissent dans les textes institutionnels (Ins et InsM) (cf. figure 4.14 AFC des variables intratextuelles des quatre genres discursifs). En plus des expressions proverbiales, les termes et expressions de l'ordre constituent une autre particularité des textes institutionnels (Ins et InsM). Ces expressions de l'ordre sont formées de quatre syllabes comme les expressions proverbiales. Ces expressions figées apparaissent fréquemment dans les textes institutionnels est pour exprimer les déterminations des autorités face à la situation de pollution atmosphérique, tels que 大力发展 d'une fréquence de 54 sur 99 (se développer vigoureusement), 综合治理 (la gestion globale) d'une fréquence de 37 sur 78, 关停并转[18] (Mesures à la chinoise visant à optimiser la structure industrielle et à redresser les entreprises) d'une fréquence de 12 sur 16, 全面推广 (généraliser) d'une fréquence de 13 sur 17. Voici un exemple concret extrait du sous-corpus Ins à l'aide de la fonctionnalité « carte de section »[19], qui permet de mettre en relief les expressions proverbiales (en bleu) et les mots d'ordre (en rouge) dans le texte.

Par rapport aux genres institutionnels, le genre Profane est caractérisé par la récurrence des terminologies techniques (colorées en rouge). Certains comptes du WEIBO publient chaque jour l'indice de la qualité de l'air des villes chinoises. Ci-dessous, un exemple concret extrait d'un *weibo* dans lequel se trouvent six terminologies techniques : AQI (indice de la qualité de l'air), 颗粒物 (particule fine), 微克 (microgramme), 立方米 (mètre cube), 良 (二级) (bon, deuxième degré).

- *weibo* original : 【江宁区环境空气质量】2017 年 12 月 21 日 09 时，江宁区环境空气质量指数(AQIeng) 为 83，空气质量为良 (二级) ，首要污染物为细颗粒物 (PM2o5) ，PM2o5实时浓度为 61微克/立方米。
- Traduction : 【La qualité de l'air dans le district de Jiangning】21/12/2017 9h00, l'indice de la qualité de l'air (IQA) de Jiangning est 83, la qualité de l'air est bonne (second degré) , les particules fines (PM2o5) constituent

17. En chinois, les expressions proverbiales sont issues de la tradition historique littéraire ; elles peuvent servir de sujet, de complément du nom ou de complément verbal. Le dictionnaire en ligne de Chine Nouvelle (http://www.chine-nouvelle.com/chinois/chengyu/dictionnaire) en contient 30000 entrées.

18. Sigle des quatre mots 关 闭 (fermer), 停 办 (cesser), 合 并 (combiner), 转 产 (transformer). Source d'information : https ://baike.baidu.com/item/%E5%85%B3%E5%81%9C%E5%B9%B6%E8%BD%AC/2322858

19. Chaque carré représente un seul texte.

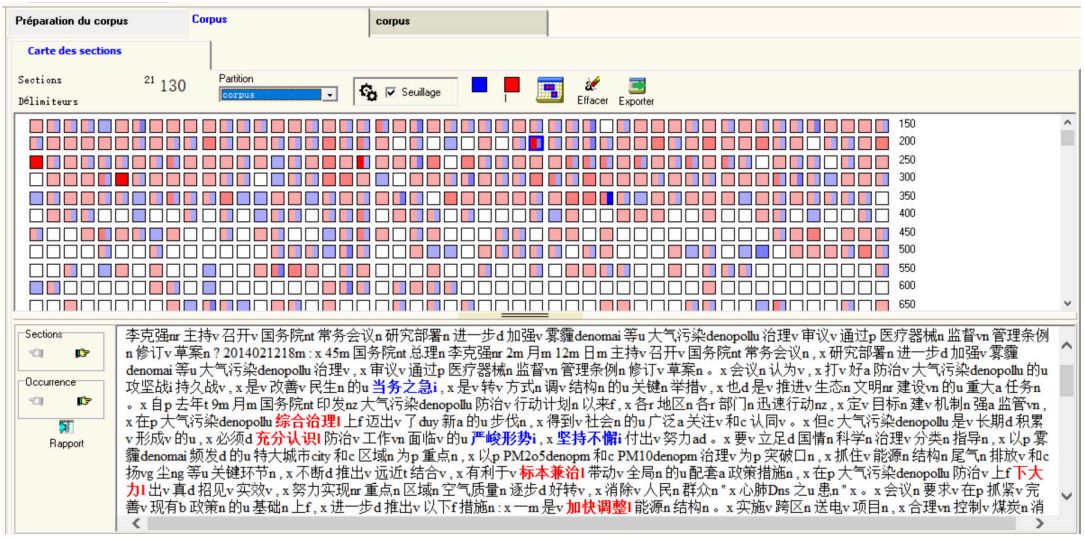


FIG. 4.4 – Exemple de l'utilisation des expressions proverbiales et des terminologies politiques dans un texte institutionnel

les polluants principaux, chaque mètre cube contient 61 microgrammes de particules fines.

L'apparition récurrente des expressions à quatre mots dans le sous-corpus institutionnel, quelque soit les expressions proverbiales ou les expressions de l'ordre, renforcent phonologiquement le style emphatique du genre institutionnel, ce qui permet d'accentuer la position hiérarchique de ce dernier. Tandis que le sous-corpus Profane s'adresse davantage au peuple dans sa vie quotidienne ; il prend en charge la diffusion des informations qui intéressent les citoyens, comme par exemple la qualité de l'air au quotidien. L'ajout des terminologies techniques permet de rendre l'information plus fiable aux yeux des lecteurs.

4.6.2.7 Analyse des noms

Six types de variables relatives à des noms sont repérés grâce aux tags :

1. Nom d'État annoté ***ns*** : 中国 **ns** (Chine), 法国 **ns** (France) :
2. Nom d'établissement (public) avec ***nt*** : 环保部 **nt** (Ministère de la Protection de l'Environnement), 发改委 **nt** (Comité national de Développement et de Réforme) ;
3. Nom de personne en ***nr*** : 李克强 **nr** (premier Ministre de la Chine), 柴静 **nr** (journaliste chinoise qui a rapporté la situation de la pollution atmo-

sphérique en Chine) ;

4. Nom propre en ***nz*** dans tous les domaines : 埃菲尔铁塔 **nz** (Tour Eiffel), 北京日报 **nz** (Le quotidien de Beijing), 淘宝 **nz** (site e-commerce Taobao), 百科 **nz** (l'encyclopédie) ;
5. Toponyme annoté ***city/autoreg/province*** 北京 **city** (Pékin), 河北 **province** (Hebei), 新疆 **autoreg** (Xinjiang) ;
6. Terme de localité en ***f*** : 西 **f** (l'Ouest), 北 **f** (le Nord).

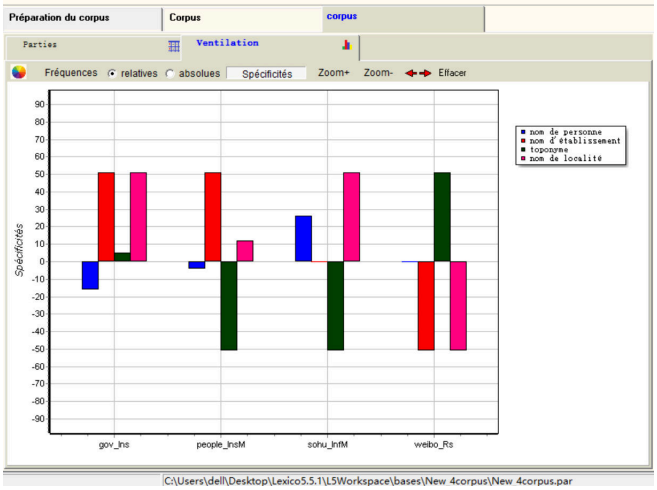


FIG. 4.5 – Ventilation de quatre types de noms dans les quatre genres textuels

Comme nous pouvons le vérifier sur le graphe 4.5 Ventilation de quatre types de noms dans les quatre genres textuels, Ins est caractérisé par les « noms des personnes (**nr**)», les « noms d'établissement (**nt**)», les « noms propres**n**», les « toponymes**city/province/autoreg**» ainsi que les « mots de localisation ou de position (**f**)». Dans un discours narratif, les quatre premières variables sont utilisées pour rendre compte de ce qui s'est passé ou de ce qui se passera (voir 4.6.2.4 Analyse du temps) « dans quel pays, entre quels établissements publics, et représenté par qui », alors que la présence de la cinquième variable (les mots de localité) marquent l'aspect « reportage météorologique » du sous-corpus institutionnel — si l'on considère le brouillard de pollution comme un phénomène naturel météorologique —, et informe de la tendance météorologique dans le bulletin quotidien (cf. Partie 5.5.1 Études sémantiques du thème 1 dans le sous-corpus Ins). Les « noms propres**n**» constituent un marqueur saillant du sous-corpus Profane, car

les publicités publiées sur WEIBO par les utilisateurs-vendeurs occupent une part importante. Ces publicités sont diffusées dans cette plateforme pour promouvoir des produits contre les effets nocifs du *wumai*, tels que le masque facial, la crème anti-poussière ou le gel nettoyant pénétrant, ou encore des purificateurs d'air (voir Exemple de la section 4.6.2.3 Analyse des mots d'emprunt).

4.6.2.8 Analyse des néologismes

L'utilisation massive de néologismes constitue une singularité des textes profanes. Les internautes se servent de néologismes lexicaux ou sémantiques pour créer un code interactif dans leur blogosphère, perçue comme une zone privée et individuelle.

Nous avons repéré deux types de néologismes dans les textes profanes [20]. Chaque type de néologisme dispose de sa propre règle de création. Voici quelques exemples concrets qui représentent ces deux types de néologismes dans les textes profanes.

1. **Le néologisme de forme** :
   - **par la dérivation** : 囧/jiong/ (embarras) est un mot dérivé du mot 冏/jiong/ (voir 4.6 pour observer l'évolution temporelle de l'écriture du mot 囧 depuis l'écriture ossécaille), qui signifie initialement « les lumières », mais qui ressemble étrangement à un visage embarrassé d'homme ; c'est pourquoi on lui attribue le sens d'« embarras ». Dans cet extrait d'un *weibo* « 雾霾险、中秋赏月险、人在囧途险 » (Le danger du brouillard de pollution, le danger de contempler la lune, le danger d'un voyage d'embarrassant.), le 囧 désigne ici l'état d'embarras de l'utilisateur qui n'arrivait pas à voir la lune à cause du brouillard de pollution.
   - **par emprunt d'un mot ou d'une expression d'une langue étrangère** : Dans le *weibo* « 冬天过去了，以为可以躲过北京的雾霾。图样图森破，生活中处处充满了惊喜.» (L'hiver est fini, j'espère me dégager du brouillard de pollution de Pékin. Trop simple trop naïf, la surprise est partout dans la vie.), l'expression néologique 图样图森破/tu yang tu sen po/ (Trop jeune, trop naïf) est calquée de l'expression originale anglaise « too young too simple », retranscrite avec des caractères phonétiquement correspondants en chinois :

20. https://www.espacefrancais.com/la-neologie/#Nologie-de-forme-et-nologie-de-sens

- **par troncation** : il s'agit d'un procédé de suppression d'une ou de plusieurs parties d'un mot. Le *weibo* suivant contient le néologisme « 艹/cao/ (putain)» issu de la troncation du mot « 草/cao/ (herbe)» qui tire parti de son homophonie avec le mot chinois signifiant « putain » ; les internautes se servent du sens dérivé vulgaire (sens équivalent à celui de « putain » en français) de ce dernier mot pour exprimer leur mécontentement en cas de la pollution de l'air. Le *weibo* complet est : « 这几天北京空气真好，雾霾稀少的就像 angelababy 的演技,艹.» (Qu'est-ce qu'il fait beau à Pékin, le niveau de la pollution de l'air atteint un niveau aussi bas que celui de la performance de l'actrice chinoise angelababy, putain.) :
- **par siglaison** : on fabrique un nouveau mot à partir des premiers éléments d'un mot ou d'une expression. Dans le *weibo* « # 雾霾 # 人艰不拆打扰了.» (#smog# Tellement la vie est difficile, il vaut mieux ne pas infliger un démenti.), 人 艰 不 拆 est un sigle acronyme de l'expression : 人生已经如此的艰难，有些事情就不要 拆穿 (Tellement la vie est difficile, il vaut mieux ne pas infliger un démenti.), on prend le premier caractère de chaque mot pour forger la nouvelle expression : 人艰不拆. Ces dernières années, cette méthode de création de néologismes est devenue de plus en plus populaire parmi les jeunes chinois ;

2. **Le néologisme de sens** : par la transformation et le détour du sens original, un nouveau sens inédit est attribué à un mot. Par exemple, dans le *weibo* « 蓝天. 雾霾. 绿龟 » (Le ciel bleu/La tortue bleue, le brouillard de pollution, et la tortue verte.), le sens original du mot 天/tiān/ est le ciel. Cependant, du fait que ce mot est composé de deux caractères 王 et 八 signifiant conjointement « la tortue » dans un sens péjoratif. Ainsi, pour garder la structure équilibre avec le mot 绿龟 (tortue verte), et pour éviter la censure des gros mots sur Internet, l'utilisateur a choisi délibérément le mot 天/tiān/ (ciel) — un mot peu commun pour remplacer le 天/tiān/ (ciel) — qui partage la même prononciation et la même signification avec lui. Le sens original du mot 天/tiān/ (ciel) est donc transformé en tortue 王八.

FIG. 4.6 – Évolution temporelle du néologisme 囧

La diversité des néologismes que nous avons analysée reflète le caractère dynamique, interactif, émotionnel et expressif du genre Profane.

4.6.2.9 Analyse de la nominalisation

En linguistique, la nominalisation est un moyen pour convertir (avec ou sans transformation morphologique) en substantif un mot qui initialement n'en est pas (par exemple, un verbe, un adjectif ou un adverbe).

À travers la distribution des étiquettes morphosyntaxiques (dans le graphe 4.14 AFC des variables intratextuelles des quatre genres discursifs les mots nominalisés sont annotés avec « vn ») effectuée selon l'indice de la spécificité de chaque étiquette, nous observons que le phénomène de la nominalisation constitue une saillance des textes institutionnels. Pour mieux appréhender ce phénomène et étudier en quoi il est intéressant d'utiliser ce procédé linguistique, nous avons exploré la concordance d'un verbe nominalisé — « 保护环境 (protéger l'environnement)» — comme mot-pivot dans le sous-corpus Ins. Nous avons constaté que, dans les textes institutionnels, le mot « 保护环境 (protéger l'environnement)», initialement une locution verbe-objet, est considéré comme un nom et est utilisé de manière différente :

- comme nom-sujet en tête de phrase : 保护环境事关人民群众健康和可持续发展。(La protection de l'environnement concerne non seulement la santé du peuple chinois, mais aussi le développement durable de l'économie.) :
- comme nom-attribut suivi de 的 (de) et du nom-cible : 加强和提高人们对保护环境的意识 (Renforcer et intensifier chez les citoyens la prise de conscience de la protection environnementale.) :
- comme nom-objet après le verbe : 呼吁保护环境 (sensibiliser (le grand public) à la protection de l'environnement).

Après la nominalisation, l'effet de la mise en œuvre des actions du verbe « 保护环境 (protéger l'environnement)» est affaibli. En revanche, l'état ou l'effet de la

mise en scène de la « protection de l'environnement » est accentué. En considérant le verbe nominalisé comme un nom et en l'utilisant de manière différente (voir les exemples montrés ci-dessus), les textes institutionnels ont diversifié et varié les informations concernant leur travail et les résultats obtenus après les missions du gouvernement. L'analyse de la concordance du mot-pivot 保护环境 dans le sous-corpus Ins témoigne de ce phénomène. Nous avons choisi deux exemples représentatifs (voir ci-*infra*).

1. Exemple 1
   - **Phrase originale** : 也表明了新一届中央政府治理雾霾，保护环境，还蓝天绿水于人们的坚定决心 ;
   - **Traduction** : Ce qui montre aussi la détermination ferme du nouveau gouvernement à propos de la résolution du brouillard de pollution, dela protection de l'environnement, de la restitution du ciel bleu et de l'eau limpide au peuple ;
   - **Effet et intention** : pour montrer la ferme détermination sur la protection de l'environnement, en l'occurrence, le problème atmosphérique.
2. Exemple 2 :
   - **Phrase originale** : 采取立法经济科技等多种手段进行保护环境 ;
   - **Traduction** : Diverses mesures législatives, économiques et techniques ont été prises en vue de la protection de l'environnement ;
   - **Effet et intention** : pour faire connaître toutes sortes d'initiatives mises en action par les institutions à propos de la protection de l'environnement.

Ainsi, la nominalisation constitue un phénomène saillant du genre institutionnel. Les mots nominalisés sont utilisés de manière à mettre en avant les résultats des mesures préventives adoptées par le gouvernement face à la pollution de l'air. Les expressions nominalisées permettent d'exposer des notions de façon brève et efficace, ce choix des mots s'avère ainsi important dans le sous-corpus institutionnel.

4.6.3 *Variables sémiotiques*

Les études sémiotiques développées ici visent à analyser l'utilisation de la ponctuation, des mots-consignes et des émoticônes. Nous avons deux types de ponc-

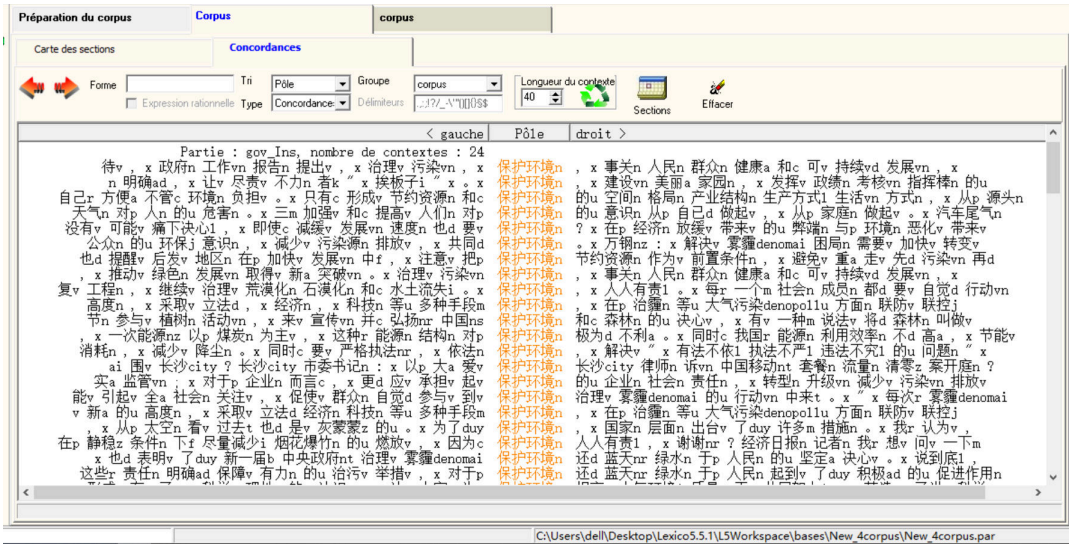


FIG. 4.7 – Concordance du verbe nominalisé 保护环境 (protéger l'environnement) dans les textes Ins

tuation :

- ponctuation expressive : pour exprimer des sentiments. Par exemple, les points d'interrogation (?) et d'exclamation (!), et ceux utilisés de manière conjuguée (?!/!?) ou répétitive (???/!!!), et les points de suspension [21] :
- ponctuation neutre : qui serve à introduire les discours des locuteurs, tels que les deux points et les doubles guillemets :

Deux mots-consignes qui sont utilisés exclusivement dans les réseaux sociaux sont pris en compte : # vise à proposer un «*topic*» ; @ [22] pour *tagger* une personne ou un compte.

À travers le diagramme (cf. figure 4.8 Comparaison de l'utilisation de la ponctuation) indexé selon la fréquence relative de chaque signe sémiotique (voir tableau 4.4 Comparaison de l'utilisation de la ponctuation), nous constatons que la particularité des textes profanes réside dans l'utilisation plus importante de ponctuation expressive liée à la subjectivité, à l'utilisation des mots-consignes

21. Les points de suspension détient plusieurs fonctions : pour exprimer des sous-entendus, étendre implicitement une énumération, s'exprimer par intermittence, dénoter une idée vague, confuse ou imprécise, marquer un moment de silence, prolonger la voix, ou encore pour indiquer qu'une phrase n'est pas achevée

22. La fonction de @ visant à composer l'adresse mail ne sera pas discutée dans notre analyse sémiotique.

et des émoticônes . En ce qui concerne la ponctuation expressive, il s'agit des points d'exclamation, les points d'interrogation, la répétition de ces signes (!+ , ?+ [23]), et la répétition des points de suspension (...+). L'emploi des mots-consignes (# et @) montre la forte interactivité du genre Profane. Une autre de ses singularités consiste en l'utilisation abondante des émoticônes (cf. figure 4.9 Ventilation des émoticônes dans quatre genres textuels selon la fréquence relative). Ces mini-signes se situent entre la modalité écrite et la modalité orale, car elles transmettent une information dans un format expressif visuel (expressivité du visage : sourire, clin d'œil, en colère ou encore triste, etc.) que l'écrit ne permet pas de transmettre aussi directement.

TAB. 4.4 – Comparaison de l'utilisation de la ponctuation

Ponctuation	Ins	InsM	InfM	Prof
!	177	832	525	1650
!+ [24]	2	33	5	151
?+	1	11	2	51
?!	0	6	4	19
!?	0	1	0	3
:	650	1150	768	1563
”	167	429	184	174
...+	0	18	63	192
@	15	14	172	749
#	1	32	8	960

23. Dans ces expressions régulières, le + permet de répéter le symbole précédent un nombre quelconque de fois

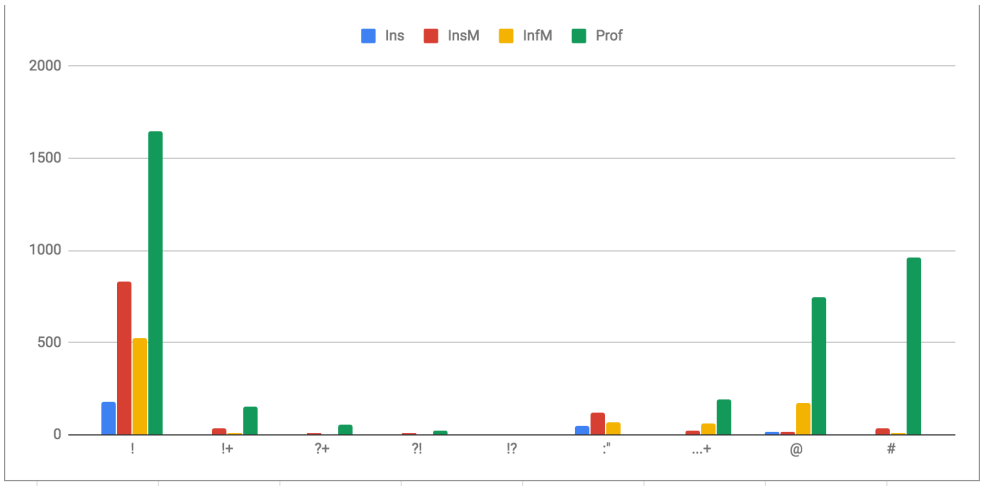


FIG. 4.8 – Comparaison de l'utilisation de la ponctuation

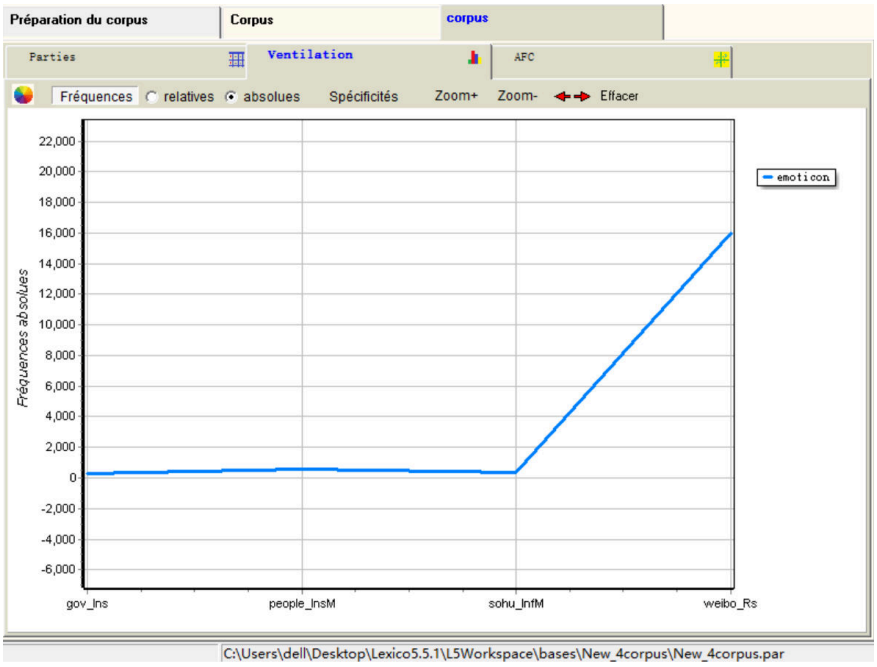


FIG. 4.9 – Ventilation des émoticônes dans quatre genres textuels selon la fréquence relative

Nous observons dans le sous-corpus Ins une présence accrue du groupe de la ponctuation neutre de « : “», utilisé pour rapporter directement un discours

(cf. Exemple 1). Par ailleurs, le point d'exclamation est beaucoup utilisé dans les textes institutionnels, qui suggère le recours fréquent à un style emphatique (cf. Exemple 2).

- **Exemple 1** : (extrait du GOV) 中国科学院植物研究所研究员陈佐忠则从城市的角度给出了治理空气污染的建议：“北京有几百万或者上千万株树，每株树坑 1 平方米，就是几百万或者上千万平方米的裸露土地。大风一起就是沙尘。不过，这些人工防控的措施，如果缺少人们节能减排的帮助，收效甚微“。
- **Traduction** : Le chercheur botaniste à l'Académie des Sciences de la Chine a apporté des propositions pour résoudre le problème de la pollution atmosphérique : “ Il y a des millions et des millions d'arbres à Pékin, chaque fosse d'arbre occupe un mètre carré, ce qui fait des millions de terres exposées. Beaucoup de poussières et de sables seront soulevés par le vent fort. Toutefois, ces mesures de prévention ont très peu d'effet si elles ne sont pas suivies par des gestes de la population pour économiser l'énergie et réduire les émissions."
- **Exemple 2** : (Extrait du PEOPLE) 连日的雾霾天，PM2o5 值接连爆表，新鲜空气难以寻觅！
- **Traduction** Le brouillard de pollution chasse le ciel bleu depuis plusieurs jours, et l'indice de PM2,5 atteint son pic. On ne peut trouver l'air frais nulle part !

En résumé, au niveau sémiotique, le sous-corpus Profane se caractérise par la présence importante de la ponctuation expressive, des mots-consignes et des émoticônes. Ces signes dénotent le caractère expressif, subjectif et interactif du genre Profane ; tandis que les textes institutionnels utilisent davantage de la ponctuation neutre «: et “ » pour introduire directement un discours. Son style emphatique est reproduit (à part des verbes modaux et des expressions de l'ordre) par l'utilisation du point d'exclamation.

4.6.4 *Variables modales*

Quatre types de modalités ont été détectés dans notre corpus ; ils caractérisent chaque genre textuel tantôt conjointement, tantôt séparément.

4.6.4.1 Modalité déclarative avec les mots de négation

La **modalité déclarative** est utilisée pour énoncer un fait, qui peut être affirmatif, négatif, dubitatif ou emphatique. Nous avons remarqué que l'utilisation des mots de négation constitue une caractéristique des textes institutionnels (Ins et InsM) (cf. Annexe 20 Termes de négation). Ces mots se présentent sous deux formes, soit des mots négatifs , tels que 杜绝 (éliminer), 禁止 (interdire), 拒绝 (refuser) pour donner un ordre formel (Exemple N1 et N2 ci-infra) ; soit avec des adverbes de négation — « ne pas » ou « non » — pour mettre une phrase à la forme négative de sorte à montrer une attitude déterminée, tels que 绝不 ((ne falloir) absolument pas), 绝非 (absolument non), 不需要 (ne pas nécessiter), 不必要 (ne pas falloir). Il faut noter que la deuxième utilisation est souvent accompagnée des verbes ou adverbes modaux (voir tableau 4.6 Verbes et adverbes modaux dans les quatre genres textuels selon l'indice de spécificité, adverbe de négation est coloré en rose, verbe ou adverbe modal est en vert). Par ailleurs, le sous-corpus institutionnel préfère emprunter des adverbes modaux pour renforcer sa tonalité affirmative. Par exemple l'adverbe 严格 (strictement) dans l'exemple N2, et 过度 (excessivement) dans l'exemple AN1, et 绝对 (absolument) dans l'exemple AN3. L'ajout de ces mots (verbe/adverbe modal) permet de renforcer l'effet emphatique du genre institutionnel.

1. Forme de négation avec des mots négatifs :
   - **N 1** : 禁止焚烧稻草秸杆的行为 ;
   - **Traduction** : Interdire les actes de brûler les pailles :
   - **N 2** : 严格杜绝超标排放 ;
   - **Traduction** : Éliminer strictement les émissions excessives.
2. Forme de négation avec des adverbes de négation :
   - **AN 1** : 面对空气污染, 不需要过度恐慌 ;
   - **Traduction** : En cas du brouillard de pollution, il ne faut pas être paniqué :
   - **AN 2** : 遏制雾霾, 绝非一日之功 ;
   - **Traduction** : La résolution du *wumai* ne s'avhève pas en un jour ;
   - **AN 3** : 电动自行车有问题, 可以规范, 但绝对不应该禁止.
   - **Traduction** : Au lieu d'interdire absolument les vélos électroniques, il faut les utiliser de manière raisonnable.

FIG. 4.10 – Carte de section des termes modaux et de négation dans GOV

4.6.4.2 Modalité interrogative avec les mots d'interrogation

La **modalité interrogative** est en général marquée par des pronoms interrogatifs. Avec l'indice des spécificités de chaque pronom interrogatif (cf. tableau 4.5 Pronoms interrogatifs dans les quatre genres textuels), nous remarquons que le recours aux pronoms interrogatifs constitue une spécificité caractéristique des textes médiatiques (cf. tableau 4.5 ci-dessous). Les textes médiatiques communiquent les actualités et expliquent les phénomènes qui s'observent en tous temps et lieux. Pour offrir ces informations concrètes et détaillées, on a recourt davantage aux pronoms interrogatifs pour poser différents types de questions à propos du sujet principal : il s'agit de qui (谁), de quoi (什么), l'actualité se passe où (哪里/哪儿) et comment (怎样/如何), et à la fin, on va s'interroger pourquoi (为什么/为何/缘何) l'affaire s'est produite, etc.

Tab. 4.5 – Pronoms interrogatifs dans les quatre genres textuels

Mot	Ins	InsM	InfM	Profane
为什么	-6.44	6.38	17.06	-13.11
为何	-2.75	11.26	12.68	-16.96
什么	-9.02	7.73	20.07	-13.75
哪儿	-2.34	2.44	4.83	-3.85
哪里	-5.22	1.32	1.73	1.82
如何	3.28	10.85	17.47	-23.32
怎么	-9	2.89	8.56	-2.76
怎样	-3.49	0.71	12.58	-6.92
缘何	1.72	2.08	1.83	-3.68
谁	-9.9	11.87	1.53	-3.93

4.6.4.3 Modalité exclamative avec les interjections

La **modalité exclamative** est utilisée pour exprimer des sentiments ; elle se concrétise à l'aide d'interjections. Ventilation des interjections dans quatre genres textuels selon la fréquence absolue nous montre que l'utilisation d'interjections constitue une caractéristique saillante du genre Profane (spécificité positive). Ces interjections sont utilisées dans le discours oral du sous-corpus Profane pour exprimer principalement des impressions et des sentiments vifs et négatifs : la crainte ou l'affliction (Exemple SN3), le dégoût ou le mécontentement (Exemple SN1 et SN5), l'aversion (Exemple SN4), la frustration ou le sentiment d'impuissance (Exemple SN2) ; ou d'autres sentiments : l'illusion (Exemple AS1 et AS4), l'ironie ou la dérision (Exemple AS2 et AS3), etc (cf. Annexe 31).

1. Sentiments vifs et négatifs :
   - **SN1** : 雾霾, 堵车, 上班, 头天上班心情不好啊 ;
   - **Traduction** : *wumai*, bouchon, travail. Fi ! je ne me sens pas très bien pour le premier jour de travail :
   - SN2 : 雾霾天气, 今天一天都毛焦火辣的, 心情极度不美丽人生真的是太无趣唉 ;
   - **Traduction** : Le temps de brouillard de pollution, je suis à bout de nerfs. Bon Dieu ! La vie est nulle.
   - SN3 : 雾霾天, 太让人闹心, 小盆友们都咳嗽感冒了, 哎 ;

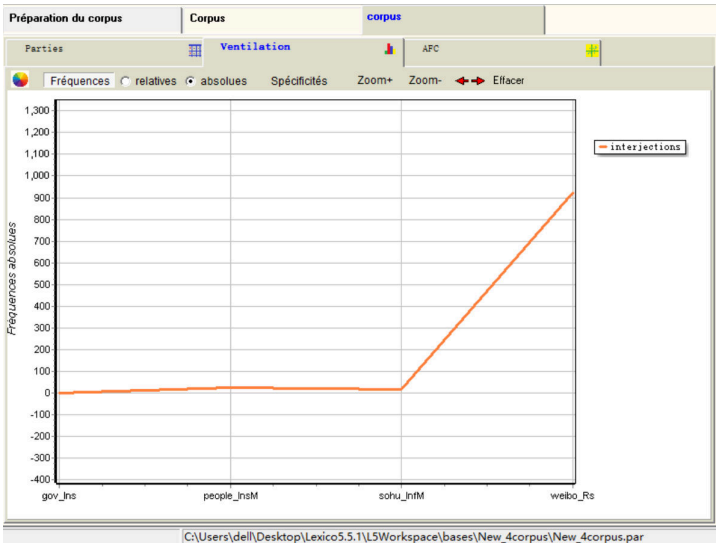


FIG. 4.11 – Ventilation des interjections dans quatre genres textuels selon la fréquence absolue

- **Traduction** : Le brouillard de pollution m'a fort agité. Ah ! les enfants toussent, ils sont tous enrhumés.
- SN4 : 哎这空气质量太差了 ;
- **Traduction** : Hélas, quelle mauvaise qualité de l'air.
- SN5 : 成都这两天雾霾天气太严重了吧 ;
- **Traduction** : Fi donc ! La pollution de l'air de Chengdu est un peu trop grave pendant ces deux derniers jours.
- SN6 : 北京雾霾依旧, 以后, 整个无车月吧 ;
- **Traduction** : Toujours le brouillard de pollution à Beijing. Ouf ! Pourrait-on avoir un mois sans voiture ?

2. D'autres sentiments :
   - AS1 : 好喜欢这霾, 连国贸都看不见了, 我在学校一片狼籍, 就多放一天假吧 ;
   - **Traduction** : Tellement j'aime le *wumai*, qui cache le bâtiment GUOMAO, je suis en chaos à l'école. Ah, donne-moi un jour de congé.
   - AS2 : 霾都全景 20130923 帝都, 重现世界末日哈哈
   - **Traduction** : Le 23 septembre 2013, l'apocalypse réapparaît dans la capitale de *wumai*, haha.
   - AS3 : 运动员们, 在雾霾中跑步不好受吧, 哈哈 , 向你们表示诚挚的慰

问 ;

– **Traduction** : C'est pénible de courir dans le *wumai*, non ? Haha, je vous réconforte, les joggeurs.
– AS4 : 原来天这么蓝达令很肯定的说这不是雾霾的北京, 这是在澳大利亚度假咧哈哈
– **Traduction** : Tellement le ciel est bleu, mon chéri a dit d'un ton ferme. Ho ! On n'est pas à Beijing : la capitale du brouillard de pollution, on est en vacance en Australie.

4.6.4.4 Modalité impérative avec les (ad)verbes modaux

La **modalité impérative** est utilisée pour donner un ordre, ou pour inciter, persuader, faire agir, répandre une propagande, etc. Outre les éléments usuels qui marquent un ton impératif, tels que l'usage en tête de phrase du verbe à la forme infinitive ; le subjonctif ou l'indicatif futur ainsi que les verbes modaux (falloir, devoir, exiger, etc.) contribuent à ce ton. D'après les calculs statistiques (voir le tableau 4.6 Verbes et adverbes modaux dans les quatre genres textuels selon l'indice de spécificité et la distribution des termes modaux (cf. graphe 4.12 AFC des verbes et adverbes modaux dans quatre sous-corpus), les verbes modaux recouvrant les sens de « falloir, devoir, exiger » caractérisent l'Ins. Nous observons un rapprochement avec l'Ins et le InsM du point de vue de leurs spécificités positive (sur-emploi).

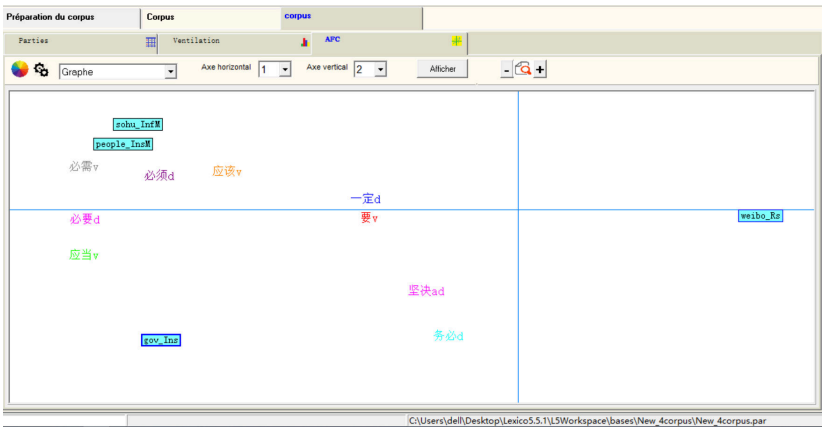


FIG. 4.12 – AFC des verbes et adverbes modaux dans quatre sous-corpus

Tab. 4.6 – Verbes et adverbes modaux dans les quatre genres textuels selon l'indice de spécificité

Verbe/Adverbe modal	Ins	InsM	infM	Profane
要	18.01	23.88	10.19	-37.56
应当	8.62	9.43	2.32	-15.79
必须	7.01	18.94	7.06	-25.84
必要	6.76	8.05	4.8	-15.16
一定	5.71	8.03	9.41	-16.79
坚决	5.22	3.86	-4.24	-4.31
务必	2.08	-1.04	-1.47	-0.8
强制	1.56	5.93	5.64	-10.15
必需	0.69	2.48	1.75	-3.01
应该	0.67	23.37	12.23	-27.58

4.6.5 *Variables rhétoriques*

La rhétorique se rapporte à l'action du discours sur les esprits. Les textes écrits en font un large usage afin de s'assurer une expression de meilleure qualité, à même de favoriser une meilleure transmission des idées qu'ils défendent, en espérant ainsi convaincre le lectorat. Dans les analyses suivantes, nous allons traiter de quatre types de figures de rhétorique présentes dans notre corpus : l'homonymie, l'ironie (effet emphatique), la métaphore et le parallélisme. Ces figures se retrouvent dans l'ensemble des textes que nous étudions ici.

4.6.5.1 Homonymie

Nous relevons le recours à l'homonymie dans les *weibo* et des textes de SOHU, dans lesquels les internautes ont inventé eux-mêmes divers surnoms pour le brouillard de pollution.

Par exemple 公雾源/gong wu yuan/, dont la prononciation est identique à celle de 公务员 (fonctionnaire), signifie littéralement l'origine publique du brouillard de pollution. On peut aussi citer l'exemple de : 尘疾思汗/chen ji si han/. Cette expression fait référence à 成吉思汗 (Genghis Khan), parce que les deux groupes de mots partagent une prononciation quasi identique. Les internautes ont remplacé

les deux premiers caractères du nom de Genghis par 尘疾, caractères signifient respectivement « poussières [du smog]» et la « maladie ». Un dernier exemple de ce jeu d'homonymie : 喂人民服雾/wei ren min fu wu/. La version initiale de cette expression est 为人民服务 (au service du peuple), le slogan historique du Parti communiste chinois. On a remplacé le premier caractère par 喂, qui signifie « faire prendre », et le dernier caractère par 雾, qui désigne le « brouillard ». Ainsi, le slogan du PCC est devenu désormais « faire respirer l'air pollué au peuple ».

4.6.5.2 Ironie

En plus du jeu d'homonymie en œuvre dans certains messages du WEIBO, nous avons aussi relevé des blagues publiées par certains internautes de SOHU en utilisant la figure de l'ironie. Elles se basent souvent sur l'emphase et l'exagération pour ménager un ton ironique. En voici donc deux exemples.

- **1er exemple** : 遛狗不见狗, 狗绳提在手, 见绳不见手, 狗叫我才走。
- **Traduction** : Quand je promène mon chien avec un laisse, je vois le laisse mais pas le chien, alors j'avance seulement au bruit du chien.
- **1er exemple** : 京城, 菜市, 一个犯人跪在地上, 即将被处决, 午时三刻已到, 行刑 !, 话音刚落, 蒙面的刽子手上前, 扯下了犯人的口罩。
- **Traduction** : Dans un marché de la capitale, un condamné, genoux à terre, était sur le point d'être exécuté. Une voix dit : « Il est déjà 3h, c'est l'heure de l'exécution » ! Le bourreau s'est donc rapproché du condamné et a enlevé son masque.

Derrière la plaisanterie permise par le jeu d'homonymie ou l'effet emphatique, se cache l'impuissance du grand public face à la grave situation de pollution de l'air, ainsi qu'aux faibles réactions du gouvernement contre le smog.

4.6.5.3 Métaphore

À la différence des sites informels, les textes institutionnels préfèrent un autre type de figure : la métaphore. Considérons ensemble le texte suivant :

- **Texte original** : 人民群众最不满意, 视腐败为最严重的政治雾霾。
- **Traduction** : C'est le peuple qui en est le moins satisfait, et on compare la corruption au « brouillard de pollution » le plus grave en matière politique.

Dans ce texte, le point commun entre la corruption et le brouillard de pollution tient à ce qu'ils sont tous deux qualifiés de « polluants ». Dans le cas de la

corruption, c'est l'environnement politique qui fait l'objet de la pollution, tandis que dans le second cas, il est question de l'environnement naturel.

4.6.5.4 Parallélisme

Outre la métaphore, le parallélisme constitue une autre figure rhétorique fréquemment utilisée dans les textes institutionnels. Le parallélisme consiste en la juxtaposition de deux phrases (courtes en général) symétriques en termes de structure. Le sens de ces phrases est soit similaire, soit opposé.[25]. La structure symétrique des deux phrases parallèles confère une valeur rythmique et poétique à la combinaison, ce qui facilite la transmission d'un message et sa mémorisation par le récepteur. Ces avantages font que le parallélisme est souvent adopté par les sites institutionnels. Les figures de parallélisme dans les textes institutionnels ont une fonction de type « slogan » pour unifier autour d'une idéologie politique. Nous allons présenter ci-dessous un exemple canonique de parallélisme utilisé dans les textes formels. Nous avons souligné en rouge les parties mises en contraste.

- **Exemple** : 大气环境保护事关人民群众根本利益, 事关经济持续健康发展,事关全面建成小康社会, 事关实现中华民族伟大复兴中国梦。
- **Traduction** : La protection de l'environnement atmosphérique concerne l'intérêt fondamental du peuple chinois, (concerne) le développement durable et vigoureux de l'économie, (concerne) la construction d'une société décente et productive, (concerne) la réalisation de l'objectif du grand renouveau national.

Cet exemple met en avant l'effet du parallélisme que veulent produire les textes institutionnels : un slogan rythmique, sonore, incitant, pour créer un effet incrémental au niveau de l'idéologie politique.

4.6.6 *Variables syntaxiques*

D'après Zhefei Fang (2006), en chinois, on se base sur le nombre de mots pleins d'une phrase pour juger de la longueur de celle-ci. Ainsi, une phrase comportant plus de sept mots pleins est considérée comme une phrase longue, alors qu'une phrase courte en comporte moins. En chinois, il existe 172 mots grammaticaux : il s'agit des conjonctions, des prépositions, des particules, des particules

25. cf.https://baike.baidu.com/item/%E5%AF%B9%E5%81%B6/3590947.

modales et des adverbes. Pour calculer la longueur moyenne d'une phrase et la longueur de phrase la plus fréquente, nous devons donc d'abord enlever ces mots vides et la ponctuation associée. Il faut noter qu'en ce qui concerne la ponctuation, nous avons gardé celles qui permettent de marquer la fin d'une phrase : le point final, le point d'interrogation, le point d'exclamation et les points de suspension. Cela étant fait, nous effectuons trois calculs quantitatifs : la longueur la plus fréquente de phrase, la longueur moyenne de phrase et la longueur moyenne de mot. Le résultat est recensé dans le tableau 4.7, accompagné d'un graphique permettant une meilleure visualisation.

TAB. 4.7 – Comparaison de la longueur de phrase

Longueur	Ins	InsM	InfM	Prof
Longueur moyenne de phrase	27	30	26	22
Longueur moyenne de mot	5	5	3	2
Longueur de phrase la plus fréquente	15	17	13	7

Comme le montre le graphique 4.13 Comparaison de la longueur moyenne de phrase et de mot, que ce soit au niveau de la longueur moyenne de phrase, de la longueur de phrase la plus fréquente, ou encore de la longueur moyenne de mot, l'InsM se situe au premier rang, qui est suivi par l'Ins, lui-même suivi par le Profane. Nous avons ainsi établi l'ordre suivant en termes de la longueur moyenne de phrase et de longueur de phrase la plus fréquente : InsM>Ins>InfM>Profane. Évidemment, par rapport au WEIBO dont le style est plutôt oral, les textes institutionnels et médiatiques, qui relèvent d'un registre soutenu (écrit), font un usage fréquent de la subordination complexe et privilégient les phrases ou expressions longues. Le résultat du calcul de la longueur moyenne des mots correspond aux résultats d'étude de la ponctuation (cf. Section 4.6.3 Variables sémiotiques) et de Analyse des collocations et des terminologies, où les textes institutionnels utilisent davantage la combinaison de «: “ » et les expressions proverbiales et les expressions de l'ordre.

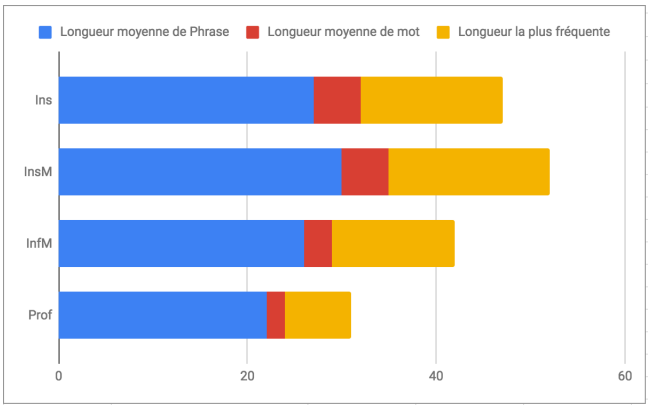


FIG. 4.13 – Comparaison de la longueur moyenne de phrase et de mot

4.7 Récapitulatif des caractéristiques des quatre genres textuels au niveau intratextuel

Comme montré le tableau 4.1 Choix des variables, nous avons cinq type de varibales comme traits intratextuels :

- **Variable lexicale** : l'acronyme et l'abréviation, la conjonction, les mots d'emprunt, marqueurs du temps verbal, le pronom personnel, la collocation et la terminologie, le nom, le néologisme et la nominalisation ;
- **Variable sémiotique** : la ponctuation, les mots-consigne et les émoticônes ;
- **Variable modale** : les mots de négation, les mots d'interrogation, l'interjection, les adverbes et les verbes modaux ;
- **Variable rhétorique** : l'homonymie, l'ironie, la métaphore et le parallélisme ;
- **Variable syntaxique** : la longueur moyenne de phrase/mot, la longueur de phrase la plus fréquente.

Ces multiples traits discriminants permettent de relever les points communs mais aussi distinctifs des différents sous-corpus et ainsi de caractériser chaque genre textuel. Nous avons effectué une analyse AFC (ci-*infra*) à l'aide de l'outil TXM. Le graphique généré par l'outil donne une vision globale de la distribution des variables intratextuelles dans les quatre genres de sous-corpus. À partir duquel, nous pouvons observer la relation corrélative entre les variables et le sous-corpus.

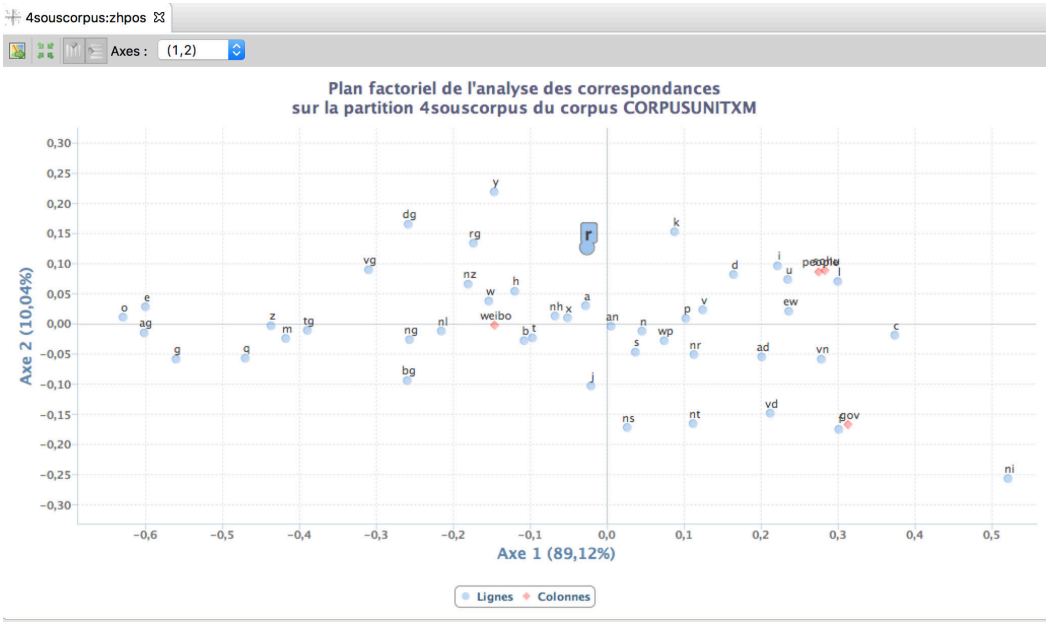


FIG. 4.14 – AFC des variables intratextuelles des quatre genres discursifs

En nous référant au graphique AFC des variables intratextuelles des quatre genres discursifs, nous arrivons à résumer les caractéristiques intratextuelles de chaque genre textuel à partir des deux composantes sémantiques : dialectique, dialogique.

4.7.1 *Caractéristiques intratextuelles du genre Ins*

Au niveau dialogique, les deux genres institutionnels (Ins et InsM) sont rattachés au discours énonciatif narratif prononcé par les énonciateurs officiels (nom d'établissement (**nt**), noms de personnes étiquetés en *nr*, l'abréviation étiquetée en **j**, toponymes, cf. section 4.6.2.7) en privilégiant un vocabulaire descriptif (cf. partie 4.6.2.4). La récurrence de slogans et de termes de l'ordre fait monter l'utilisation abondante du pronom personnel « il » et celui du « on » (pronom personnel, étiquetée en **r**, cf. section 4.6.2.5). La sur-emploi des mots qui localisent un événement ou un état dans le temps (**t**) (cf. section 4.6.2.4,) dénote son intervalle temporel. Cela montre d'ailleurs que le genre institutionnel est inscrit dans la disjonction spatio-temporelle de la situation au moment de la production du discours. Autrement dit, on peut dire « ce qui est absent » ou « n'est pas là » au moment de l'énonciation.

Au niveau dialectique, le genre institutionnel s'inscrit dans un cadre argumentatif 1) pour manifester sa modalité injonctive (verbe de modalité **v**[26], cf. section 4.6.4.4), 2) pour persuader, convaincre ou défendre sa position (expression et termes de l'ordre (**i**), cf. section 4.6.2.6) en développant un raisonnement structuré et logique (conjonction de coordination et de subordination (**c**), cf. 4.6.2.2), et 3) pour exprimer son opinion sur un ton affirmatif/négatif (mots de négation (**d**), cf. section 4.6.4.1) de manière emphatique (figure du parallélisme (**i**), cf. section 4.6.5.4 ; point d'exclamation **x**, cf. tableau 4.8 ; adverbe modal (**ad**), cf. section 4.6.4.1). De plus, l'Ins exige des normes écrites plus strictes (**i**), cf. section 4.6.2.6 ; longueur moyenne de phrase et de mot, cf. section 4.6.6). De par conséquent, le choix des mots est plus précis (nominalisation **vn**, voir section 4.6.2.9). En même temps, les procédés grammaticaux sont plus variés (métaphore, voir section 4.6.5.3) afin de mieux organiser les informations à un rythme soutenu (parallélisme) et de les transmettre de manière plus efficace. Tous ces éléments sont unifiés autour de son idéologie politique.

4.7.2 *Caractéristiques intratextuelles du genre InfM*

Étant donné qu'il emprunte la forme d'un récit épousant la réalité au plus près, le genre InfM est rattaché à la fois au discours médiatique en ce qu'il rapporte des faits, des événements, et au discours journalistique scientifique (terminologie technique) puisqu'il présente et transmet des savoirs et qu'il cherche à expliquer des phénomènes.

Au niveau dialogique, en tant que moyen médiatique, le pronom impersonnel « il » est utilisé davantage par le genre InfM pour introduire une phrase. Nous observons également un emploi massif des entités nommées telles que les noms de personne **nr**, les noms d'état **ns**, les noms d'établissements **ns**, les toponymes (cf. section 4.6.2.7) dès lors que les auteurs s'emploient à raconter un évènement.

Au niveau dialectique, l'InfM assume son rôle polémique de contestation et et de problématisation à travers la modalité interrogative traduite par l'utilisation abondante des pronoms interrogatifs (voir tableau 4.5). Son côté scientifique quant à lui se constate dans la multiplication de l'usage de signes techniques (**m**) , les terminologies techniques (cf. section 4.6.2.6) et les mots d'emprunt (**eng**) (cf. section 4.6.2.3), lesquels dénotent ses aspects proprement scientifiques.

26. La lettre en gras constitue l'étiquette morphosyntaxique.

4.7.3 *Caractéristiques intratextuelles du genre InsM*

Le genre institutionnel-médiatique, qui est d'origine institutionnelle, adopte la fonction et la forme du discours médiatique. Il se trouve dans une zone transitoire entre l'Ins et l'InfM.

Au niveau dialogique, d'une part, l'emploi massif du pronom « nous » et de la combinaison de « je+tu » permettant de former des slogans relève du institutionnel ; d'autre part, la présence abondante du pronom impersonnel « il » en tant que sujet phrastique, rejoint l'InfM pour introduire des phrases.

Au niveau dialectique, comme dans le genre institutionnel, le déroulement aspectuel du genre InsM est marqué par les conjonctions de coordination et de subordination. L'emphase et l'injonction sont marquées par l'utilisation davantage de verbes modaux (**v**), des expressions de l'ordre (**i**), du point d'exclamation (**x**), de la négation (**d**), de l'adverbe modal (**ad**), et de la figure du parallélisme. Alors que le style narratif se caractérise par l'emploi massif des deux points (**x**), des guillemets (**x**), de la phrase longue et de mots longs. Le genre média-institutionnel affiche son caractère médiatique à travers une série d'éléments représentatifs : mots d'emprunt scientifiques (**eng**), nom d'établissement (**nt**), nom de personne (**nr**), abréviation (**j**), nom propre (**nz**).

4.7.4 *Caractéristiques intratextuelles du genre Profane*

Au niveau dialogique, le genre Profane adopte essentiellement un discours de l'ordre d'*exposé*, il se caractérise par des mots fréquemment utilisés dans l'oral. Ce qui rend l'utilisation plus importante de la 1ère et 2ème personne du singulier et du pluriel (cf. section 4.6.2.5). La présence de ces pronoms personnels (**r**) indique le caractère interactif et communicatif du genre Profane, car l'utilisation de ces pronoms personnels est indispensable dans l'interaction et la communication des utilisateurs. La troisième personne au féminin singulier et pluriel montre qu'une attention particulière est accordée à la population féminine dans le genre Profane. En tant que discours oral, le Profane privilégie le temps de l'indicatif présent (**t**) (voir tableau 4.2), qui implique une relation de conjonction avec la situation d'énonciation produite dans le texte.

Le genre Profane provient d'une zone privée, individuelle et communicative. Ainsi, au niveau dialectique, cela dote l'échange et la communication d'un style oral (phrases courtes (cf. tableau 4.7), mots simples (cf. idem).), plus interactive

(mots-consignes (**x**) (cf. tableau 4.4)), plus dynamique (émoticônes (cf. figure 4.9)), plus expressive (interjection (**y** ou **e**) (voir section 4.6.4.3), onomatopées (**o**) (cf. figure 4.14), ponctuation expressive (**x**) (voir figure 4.8)). Les informations du sous-corpus Profane se présentent avec plus de créativité (néologismes (**x**), ponctuation libre (**x**) (voir section 4.6.2.8)), et plus de diversité, que ce soit au niveau de la forme (publicité : mots d'emprunt (**eng**), noms propres (**nz**) (cf. section 4.6.2.7) ; vulgarisation scientifique, terminologie scientifique), ou de style (homonymie, ironie). Pour conclure, tout cela reflète le caractère communicatif, interactif, dynamique, expressif, divers et créatif du genre Profane.

Avec les résultats d'études au niveau infratextuel et intratextuel à l'appui , nous résumons notre propos par le graphique suivant, qui catégorise les caractéristiques des quatre genres textuels en fonction de trois zones de discours proposées par Rastier (2001).

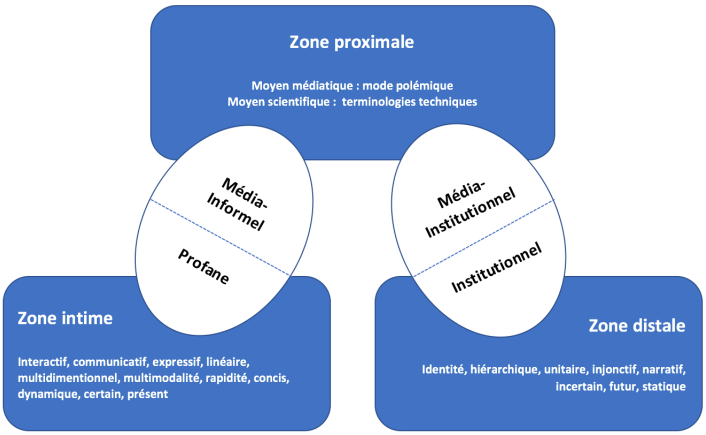


Fig. 4.15 – Caractéristiques des quatre genres textuels encadrées dans les trois zones de discours

4.8 Conclusion

Nous avons discuté et recueilli dans ce chapitre les caractéristiques macrosémantiques des quatre genres textuels à deux niveaux : infratextuel et intratextuel. Dans le chapitre suivant, nous allons relever les thèmes principaux dans chaque sous-corpus et interpréter leur sémantique tout en tenant compte des résultats obtenus dans ce chapitre.

Chapitre 5

Analyses sémantiques des thèmes principaux du corpus

5.1 Introduction

Dans le chapitre précédent, nous avons relevé et comparé les caractéristiques de différents genres textuels à partir de deux catégories de variables discriminantes au niveau macrosémantique. L'hypothèse présidant à cette étude est que les quatre genres de textes partagent des thèmes communs au niveau mésosémantique, mais qu'ils se distinguent les uns des autres au niveau microsémantique. Nous procédons dans ce chapitre à des études sémantiques du corpus aux niveaux méso- et microsémantique. Plus concrètement, nous identifierons les thèmes principaux qui s'établissent dans chaque sous-corpus ; puis, en nous basant sur les caractéristiques de chaque genre textuel, nous interpréterons chaque thème à partir des statistiques calculées par des outils textométriques. Pour chaque partie d'étude, nous présenterons successivement notre méthode de travail, les résultats obtenus, les analyses effectuées ainsi que la conclusion.

5.2 Méthode de travail pour l'identification du thème

En nous appuyant sur la méthode discutée dans la section 2.3.2.3 Démarche de recherche des thèmes du Chapitre 2, nous identifierons successivement les thèmes principaux véhiculés dans les quatre genres de sous-corpus.

Rappelons que notre corpus est collecté à partir des mots-clés correspondants à 6 mots différents de dénominations de «*wumai*» (cf. section 3.3.4 Collecte des données à partir d'une base de données). Par conséquent, la /pollution de l'air/ constitue l'isotopie fondamentale de notre corpus. Sachant que le thème est formé par un petit réseau sémantique, qui est établi de manière récurrente, par des

sèmes hétérogènes lexicalisés dans la suite de chaîne linguistique (phrase, texte, corpus) ; il suffit de trouver un autre sème qui apparaît de manière récurrente avec notre isotopie pour former un thème. En d'autres termes, l'identification des thèmes consiste en le calcul des cooccurrents de l'isotopie /pollution de l'air/. La récupération des thèmes principaux du corpus se fait à l'aide de la fonction « Thème » [1] proposée dans l'Hyperbase et de celle de Coocs [2] (calcul de cooccurrents) d'iTrameur [3]. À partir d'un mot-pôle donné, l'outil calcule les cooccurrents selon les statistiques de l'écart réduit [4]. Ces cooccurrents sont ensuite triés en fonction du score d'écart-réduit et de leur fréquence. Le résultat se présente sous forme d'un graphe de nuage de mots, chaque nuage contenant un thème (cf. figure 5.5 Cooccurrents de 雾霾 dans les textes du genre Profane).

5.3 Identification des thèmes

À l'aide des fonctions « Thème » et « Coocs », nous avons récupéré trois thèmes saillants partagés par les quatre genres de sous-corpus (cf. les statistiques des cooccurrents dans Annexe 13 Liste de COOC de 雾霾 du GOV, 14 Liste de COOC de 雾霾 du PEOPLE, 15 Liste de COOC de 雾霾 du SOHU, 16 Liste de COOC de 雾霾 du WEIBO) :

1. thème 1 : /cause/+/pollution de l'air/ : les causes de la pollution de l'air ;
2. thème 2 : /impacts/+/pollution de l'air/ : les impacts de la pollution de l'air sur la santé [5] ;
3. thème 3 : /mesures/+/pollution de l'air/ : les mesures préventives contre la pollution de l'air.

1. http://hyperbase.unice.fr/hyperbase/controller/action/recherche/theme.php
2. http://www.tal.univ-paris3.fr/trameur/iTrameur/
3. Version en ligne de l'outil Trameur
4. Selon la notion proposée par André Salem (1994), l'écart réduit produit une valeur numérique qui mesure le caractère non aléatoire d'une observation.
5. Les impacts de la pollution de l'air sont variés, ici, nous nous concentrons sur les impacts de la pollution de l'air sur la santé.

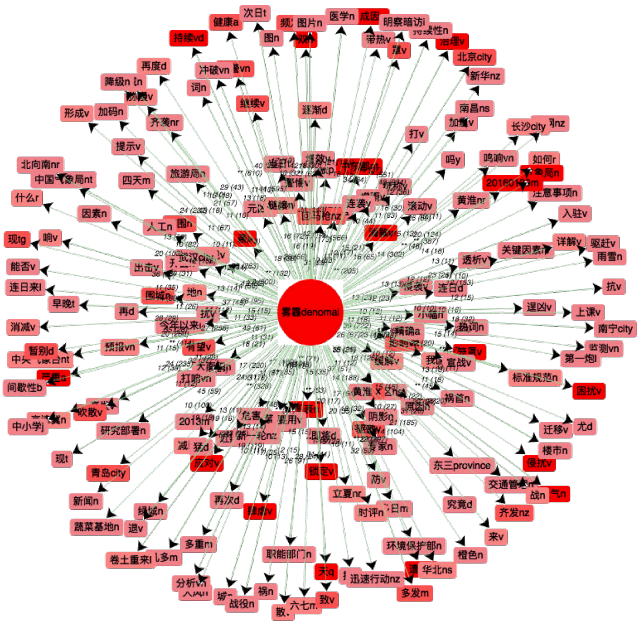


FIG. 5.1 – Cooccurrents de 雾霾 dans les textes du genre Ins

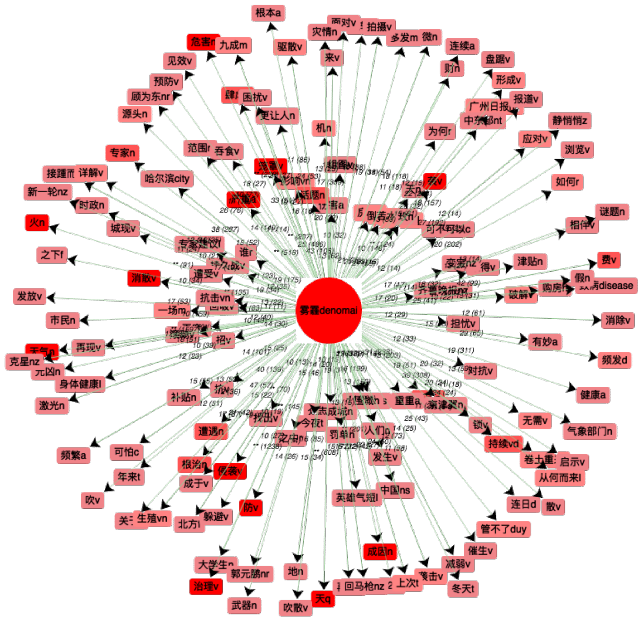


FIG. 5.2 – Cooccurrents de 雾霾 dans les textes du genre InsM

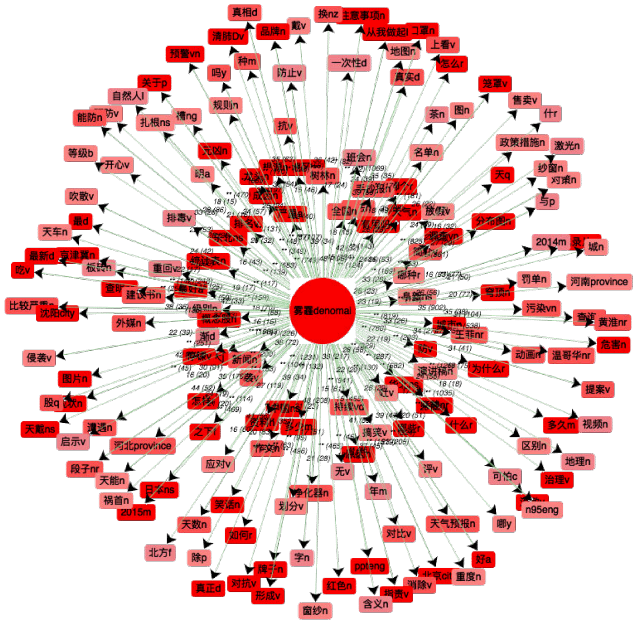


FIG. 5.3 – Cooccurrents de 雾霾 dans les textes du genre InfM

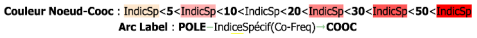


FIG. 5.4 – Consignes des réseaux sémantiques

FIG. 5.5 – Cooccurrents de 雾霾 dans les textes du genre Profane

5.3.1 *Présentation des thèmes identifiés dans les quatre genres de sous-corpus*

Les trois thèmes « causes de la pollution de l'air », « impacts de la pollution de l'air » et « mesures préventive contre la pollution de l'air » sont constitués respectivement par l'isotopie la /pollution de l'air/ en cooccurrence avec les mots associés aux sèmes /cause/, /impact/ et /mesure/ (voir ci-après). La lecture de la concordance et des segments répétés des mots-pivots du sens de /cause/, /impact/ et /mesure/ nous permet de relever les tournures préférées des textes Ins, InsM et InfM. Nous constatons qu'on utilise davantage de phrases ou propositions syntaxiquement complètes (et non pas des phrases averbales) quand on présente les causes, les impacts et les mesures relatifs au *wumai*[6]. Il s'agit principalement d'une phrase à la structure SVO (sujet-verbe-objet), où la /pollution de l'air/ constitue en général le sujet, les sèmes de /cause/, /impact/ et /mesure/ sont tantôt actualisés par les verbes tantôt par les noms comme objet. Par exemple, « 雾霾来源于 xxx 和 xxx (le brouillard de pollution est causé par xxx et xxx)», « xxx, et xxx 构成雾霾的主要来源 (xxx, et xxx forment les causes principales du brouillard de pollution)» ; « 雾霾天气晨练可以导致呼吸系统疾病 (Le sport matinal à l'extérieur peut causer les maladies respiratoires)» ; 雾霾问题的解决方法 (les solutions au problèmes de *wumai*), etc. Les trois genres (Ins, InsM et InfM) de sous-corpus privilégient chacun les mots (sur-employés) (cf. graphe 5.6 Distribution des mots du sème /cause/, /impact/ et /mesure/ dans les quatre sous-corpus) partageant la signification /cause/, /impact/ et /mesure/ et confinant à la synonymie, pour présenter les trois thèmes (cf. tableau).

À l'inverse, le genre Profane présente les trois thèmes avec des mots plus directs, qui sont spécialisés dans chaque domaine. Par exemple, les terminologies chimiques du domaine d'industrie (« 有害物质 (agent nuisible), 二氧化硫 (dioxyde de soufre), 二氧化氮 (dioxyde d'azote), 二氯甲烷 (dioxyméthane)») sont utilisées pour présenter les causes ; les noms de maladies (肺癌 (cancer des poumons), 癌症 (cancer)) ou bien les terminologies médicales (致癌物 (cancérigène), 死亡率 (taux de décès)) pour indiquer les impacts du *wumai* sur la santé ; les noms propres des produits (空气净化器 (purificateur d'air), 防霾口罩 (masque

6. Cet aspect correspond aux résultats d'étude sur les caractéristiques d'emploi de conjonction de subordination, et les spécificités de l'utilisation des phrases longues. syntaxiques (cf. section 4.6.2.2 Analyse des conjonctions, et tableau 4.7 Comparaison de la longueur de phrase du Chapitre 4).

anti-smog)) permettant de lutter contre le *wumai*.

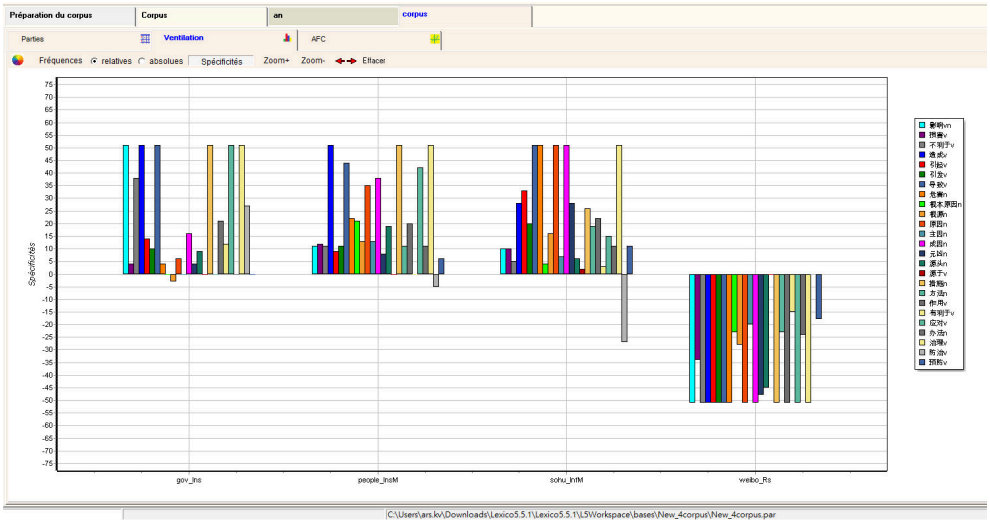


FIG. 5.6 – Distribution des mots du sème /cause/, /impact/ et /mesure/ dans les quatre sous-corpus

Le résultat de l'identification des thèmes principaux dans notre corpus confirme notre première hypothèse (cf. section 3 Hypothèse et problématique du travail de l'Introduction), selon laquelle les quatre genres des textes partagent les thèmes communs. Nous allons continuer d'exploiter la sémantique de chaque thème pour vérifier si les quatre genres de sous-corpus interprètent ces thèmes communs de manière distincte voire contrastive. Et ce, en tenant compte le privilège des mots, qui sont associés aux sèmes /cause/, /impact/ et /mesure/, récupérés dans chaque genre de textes, ainsi que les variables discriminantes que nous avons récupérées dans le chapitre précédent.

5.4 Étude sémantiques des trois thèmes

5.4.1 *Introduction*

Dans cette section, nous allons étudier la sémantique des trois thèmes relevés dans chaque sous-corpus. L'étude sémantique du thème s'effectue par deux procédés : les procédés de l'assimilation visant à relever la cohésion textuelle entre les lexiques et les thèmes en créant « un lien de similitude, d'homogénéité, d'harmonie sémantique » (PINCEMIN, 2011) ; les procédés de la dissimilation qui

TAB. 5.1 – Lexicalisation des trois thèmes dans les trois sous-corpus Ins, InsM et InfM

Sous-corpus	/causes/	/impacts/	/mesures/
Ins	成因 (cause), 原因 (raison)	影响 (influencer/influence), 不利于 (au détriment de), 造成 (causer), 导致 (conduire)	治理 (régulariser), 措施 (mesure), 防治 (prévenir), 应对 (affronter)
InsM	原因 (raison), 成因 (cause), 根本原因 (raison fondamentale), 源头 (origine)	造成 (causer), 导致 (conduire), 危害 (endommager)	治理 (régulariser), 措施 (mesure), 应对 (affronter)
InfM	元凶 (coupable), 成因 (cause), 原因 (raison)	引起 (provoquer), 导致 (conduire), 危害 (endommager)	治理 (régulariser), 措施 (mesure), 方法 (méthode), 作用 (affecter), 预防 (prévenir)

mettent en évidence les singularités et les caractéristiques présentes chez les uns mais absentes chez les autres. L'activité interprétative est un travail de mise en évidence des formes lexicales et des fonds sémantiques, qui dépend du contexte qui vient de son voisinage d'un passage, ou d'un autre passage. Afin de mieux décrire, interpréter et contraster les trois thèmes ainsi que la manière dont chaque thème est traitée, nous allons observer et analyser les cooccurrents ainsi que les segments répétés d'un thème.

5.5 Études sémantiques du thème 1 : Causes de la pollution de l'air

5.5.1 *Études sémantiques du thème 1 dans le sous-corpus Ins*

Explorons d'abord les segments répétés du thème 1 — causes de la pollution de l'air — calculés par l'outil Trameur [7] :

Tab. 5.2 – Segments répétés du thème 1 dans l'Ins et dans le InsM

Sous-corpus	Segments Répétés du thème 1 dans l'Ins	Traduction	Fréquence
Ins	雾霾成因	les causes du brouillard de pollution	47
InsM	雾霾的原因	idem	15
InsM	雾霾的主要原因	les causes principales du brouillard de pollution	14
Ins	雾霾的成因	idem	11
Ins/InsM	雾霾天气成因	les causes du temps *wumai*	11
Ins	分析即将出现的雾霾过程的成因及影响	analyser les causes et les influences de l'apparition du *wumai*	10

Dans le tableau ci-dessus, le segment 雾霾天气的成因 » (les causes du temps du brouillard de pollution), dont nous relevons une fréquence de 11, attire notre attention. Cette unité lexicale — « le temps de *wumai*» — associe le mot « 雾霾

7. De même que pour les résultats d'analyse du genre textuel, nous avons repéré lors de l'étude du thème 1 plusieurs points partagés par l'Ins et l'InsM, ce qui nous pousse à les considérer conjointement pour aborder les études suivantes.

(brouillard de pollution)» au vocable « 天气 » (temps, dans le sens la météo) ; ceci a pour effet de présenter la pollution de l'air comme un phénomène météorologique naturel. Il semble que les sous-corpus Ins et InsM tendent à confondre les deux notions qui sont pourtant distinctes. Pour vérifier cette hypothèse, nous allons calculer les cooccurrents du segment « 雾霾天气 (le temps *wumai*)» dans les deux genres de sous-corpus Ins et InsM. Et ce procédé est à observer et à évaluer les types de mots les plus caractérisants dans les contextes du segment-pivot.

En plus de l'occurrence « 天气现象 » (phénomène météorologique), nous constatons qu'un ensemble de mots se rapportant au /phénomène météorologique naturel/ : 轻雾 (brume légère), 薄雾 (brume), 降雪 (chute de neige), 降雨 (chute de pluie) apparaissent parmi les cooccurrents du « 雾霾天气 (temps *wumai*)». De plus, un faisceau d'occurrences actualisent trois sèmes différents : 1) le sème /météorologie/ : 气象局 (office météorologique), 中央气象台 (station centrale de météo), 气象专家 (expert météorologue) ; 2) le sème /géographie/ : 西北部 (région de l'ouest), 华北 (nord de Chine), 四川盆地 (bassin de Sichuan) ; 3) le sème /temps/ : 夜间 (nuit), 白天 (jour). Même s'il s'agit de trois sèmes différents, les deux derniers groupes de mots partagent la même classe sémantique de la dimension //nature//. Ainsi, les résultats d'analyses des cooccurrents du segment « 雾霾天气 (temps *wumai*)» confirment notre intuition : «**les textes institutionnels cherchent à désigner la pollution *wumai* comme un phénomène naturel**».

En nous basant sur cette tonalité principale, nous entreprenons les études analytiques et interprétatives de le premier thème « 雾霾原因 (causes du brouillard de pollution)» des deux sous-corpus institutionnels.

Outre le « temps *wumai*» actualisé par le sème /天气/ (temps naturel), nous avons aussi récupéré du vocabulaire au sens de /phénomène météorologique naturel/ dans le champs lexical du thème 1. L'assemblage de ces éléments lexicaux, inscrits dans des domaines variés (de temps, de lieux, d'organisme, etc.), font penser au « bulletin météorologique diffusé par l'établissement public ». Le bulletin météorologique institutionnel contient davantage de mots ou expressions temporelles qui parlent du futur : 预计 (estimer), 预报 (prévoir), 将出现 (il apparaîtra), 未来三天 (dans les trois jours à venir), 明天 (demain). Il présente en outre des aspects compositionnels qu'on trouvera dans un vrai bulletin météorologique : 气象局 (office météorologique), 中央气象台 (le centre national de la météorologie),

气象专家 (météorologue), 夜间 (nuit), 白天 (jour), 西北部 (région de l'ouest), 华北 (nord de Chine), 四川盆地 (bassin de Sichuan), 能见度不足1000m (la visibilité est inférieure à 1000 mètres). En plus de ces mots qui s'inscrivent dans le /bulletin météorologique/, une proportion mineure de mots, par exemple 电厂 (centrale à charbon), 燃煤 (charbon), 万吨 (dix mille tonnes), sont présents dans le champs lexical du thème 1 des deux sous-corpus Ins et InsM pour introduire la /production industrielle/ comme « causes du brouillard de pollution ».

Pour conclure, il paraît que les textes institutionnels, dans son « bulletin météorologique donné par les institutions gouvernementales », met en avant les conditions météorologiques liées au climat et à la configuration du terrain qu'il identifie comme les causes principales du brouillard de pollution, et non les causes liées aux activités humaines.

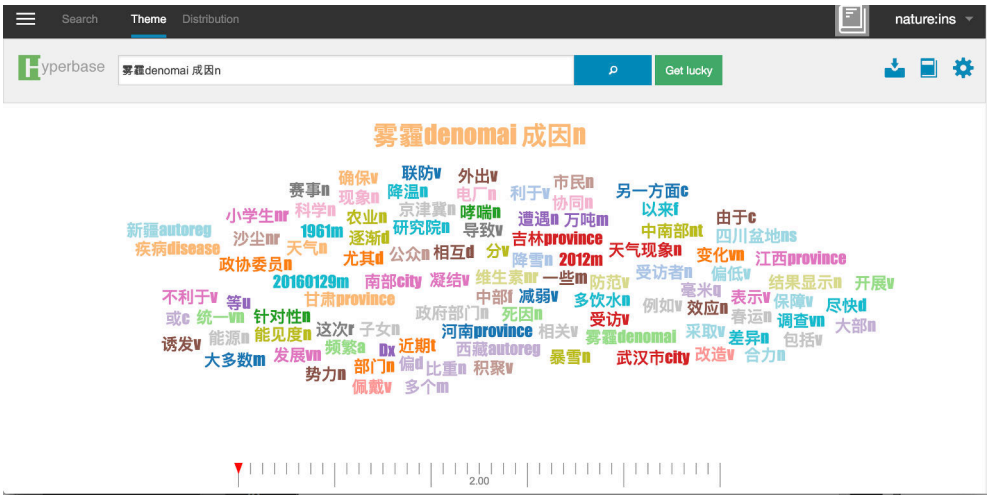


FIG. 5.7 – Cooccurrents du thème 1 du sous-corpus Ins

5.5.2 *Études sémantiques du thème 1 dans le sous-corpus InsM*

Après avoir étudié les mots figurant dans le réseau COOC du thème 1 des textes institutionnels-médiatiques, nous avons relevé les caractéristiques saillantes du sous-corpus InsM. Ces caractéristiques sont liées d'une part à sa nature institutionnelle, et d'autre part à sa fonction médiatique. En observant les cooccurrents, nous constatons que le sous-corpus InsM est semblable à l'Ins dans la mesure où il partage avec ce dernier un certains nombre de traits (cf. 4.6.2.7 Analyse des noms) qui sont inscrits dans un champ lexical typique du genre Ins. **1**) Nous

avons capté les marqueurs « institutionnels », par exemple, les noms d'état (欧洲国家 (les pays européens)), les noms de l'établissement public (气象部门 (les départements météorologiques), 环保部 (le ministère de la Protection environnementale)) et les titres de profession (人大代表 (représentant du peuple), 市长 (maire), 气象专家 (expert météorologue)) ; **2**) la tonalité injonctive est traduite par les verbes modaux exprimant le sens de 'falloir, devoir, exiger', tels que 明确要求 (demander explicitement que) ; **3**) le style emphatique est mis en évidence par les adverbes modaux, tels que 坚决 (fermement), 真正 (vraiment), 到底 (jusqu'au bout), 从根本上 (sur le fond), 明确 (explicitement), 极端 (extrêmement), 甚至 (voire), 十分复杂 (extrêmement compliqué) ; **4**) son trait abstrait est marqué par les lexiques généraux, tels que 人类 (l'être humain), 群众 (la grande masse), 全民 (tout le monde). En plus de ces traits qui caractérisent sa nature « institutionnelle », nous avons aussi noté que sa fonction de « média » est assumée grâce aux éléments suivants : **a**) la citation du nom des personnes (par exemple dans le cas d'un météorologue 马学款 ou d'un journaliste 白岩松) ; **b**) le toponyme : 伦敦 (Londres), 巴黎 (Paris)[8] ; **c**) le nom propre : 纽约时报 (le New York Times) ; **d**) le connecteur : 然而 (pourtant) ; **e**) l'indication générale de plusieurs causes qui sont liées aux activités humaines, qui sont potentiellement constituées par l'automobile, la production industrielle, la combustion du charbon et de paille aux champs : « 机动车污染 (pollution automobile)», 工业生产 (production industrielle), 烟煤 (charbon) et 秸秆 (paille).

En mettant ces cooccurrents en ordre, nous pouvons déduire la sémantique du thème 1 du sous-corpus InsM comme : des experts ou des journalistes (cf. **a**) parlent des causes de *wumai* en Chine (cf. **e**) en comparant la situation de la Chine avec celle des autres pays (cf. **b**, **c**). À travers ce document, ils lancent un appel urgent (cf. **3**) à la grande masse (cf. **4**) pour la faire réagir ensemble. En même temps, comme il s'agit des textes institutionnels, on met l'accent sur la volonté (cf. **2**, **3**) du gouvernement d'améliorer la situation du *wumai* en Chine

5.5.3 *Études sémantiques du thème 1 dans le sous-corpus InfM*

Les textes InfM expliquent les causes du *wumai* avec différents mots. Ces mots se présentent tantôt sous forme de nom, surtout des terminologies spécifiques,

8. Le sous-corpus InsM se sert de ces toponymes pour contraster les situations du *wumai* dans ces villes avec celle de la Chine dans son ensemble.

tantôt sous forme de verbe ou de groupe verbal. Ils sont listés ci-dessous :

- **Nom** : 柴油 (diesel), 柴油机 (moteur diesel), 煤 (charbon), 燃煤 (charbon), 锅炉 (chaudière), 电厂 (centrale), 燃料 (matières combustibles) ;
- **Terminologie spécifique** : 硝 (nitre), 硫 (soufre) ;
- **Verbe** : 炼 (raffiner), 排放 (émettre) ;
- **Groupe verbal** : 烧煤 (brûler les charbons).

À la différence de l'InsM, qui compare les causes du *wumai* en Chine avec celles des autres pays (cf. section 5.5.2 Études sémantiques du thème 1 dans le sous-corpus InsM), l'InfM se concentre plutôt sur les villes et les régions chinoises. En observant les segments répétés du thème 1 dans leurs contextes textuels du sous-corpus InfM, nous arrivons à résumer les causes du problème *wumai* qui sont propres à chaque ville ou à chaque région chinoise. Voici deux exemples concrets extraits de textes InfM :

Tab. 5.3 – InfM : Causes du brouillard de pollution par région

Ville/Région	Causes propres
北京雾霾污染的真正原因 (les vraies causes du brouillard de pollution de Beijing)	汽车尾气 (les gaz d'échappement des automobiles)
河北雾霾严重的原因 (les raisons pour lesquelles que le brouillard de pollution est grave)	工业排放 (émission industrielle)
华北地区雾霾原因 (les causes du brouillard de pollution dans la région du nord)	煤炭燃烧 (combustion des charbons)

Comme le montre le tableau ci-dessus, dans l'InfM, à côté du nom d'une ville ou région se trouve une cause particulière de cette ville ou région. Par exemple, la cause du *wumai* de la région du nord (华北地区) est la combustion du charbon (煤炭燃烧) ; c'est l'émission industrielle qui est à l'origine de la pollution dans la province du Hebei (河北) ; tandis qu'à Beijing, c'est l'émission des automobiles qui est la plus dénoncée. Ce qui nous explique d'ailleurs la distribution régionale des différents types de maladies présentés dans la section 1.4.2.2 Distribution régionale de problèmes de santé en Chine du Chapitre 1. Parmi les cooccurrents

obtenus du thème 1, ce sont l'adverbe argumentatif 或许 (probablement) et le nom polémique — 争议 (la controverse) qui sont les plus représentatifs du caractère médiatique du genre InfM. Nous pouvons conclure ainsi que, au lieu de discuter de manière générale les causes globales de la pollution de l'air, le genre InfM examine plutôt chaque ville/région pour étudier et affiner les causes spécifiques. Autrement dit, l'InfM se caractérise par la concrétisation, la spécialisation et l'affinement de ses analyses.

5.5.4 *Études sémantiques du thème 1 dans le sous-corpus Profane*

Par rapport au sous-corpus InfM, qui analyse les causes du *wumai* dans les grandes villes et sur des périodes de temps diverses, le sous-corpus Profane analyse les causes de manière plus détaillées : il étudie non seulement les causes dans les grandes villes, mais aussi en temps réel [9] dans les villes de petites tailles.

Prenons des exemples concrets issus de *weibo* pour expliquer cette particularité du genre Profane. Dans les exemples suivants, on publie en temps réel la qualité de l'air de tout type d'endroits (**weibo1** et **weibo4** dans le tableau suivant : il arrive qu'on affiche l'indice de l'IQA d'un quartier, voire d'une rue), les informations concernant la qualité de l'air sont mises à jour de manière régulière (**weibo2** : l'IQA du matin diffère de celle du soir.). 8 termes correspondant à 8 molécules chimiques sont utilisés dans les *weibo* pour indiquer la qualité de l'air dans différentes régions en temps réel : l'AQI (Indice de Qualité de l'Air), le PM2,5, le PM10, le O3, le SO2, le NO2, le NO et le CO. Selon les normes définies par l'Organisation Mondiale de la Santé (OMS) dans son document *Lignes directrices OMS relatives à la qualité de l'air : particules, ozone, dioxyde d'azote et dioxyde de soufre* (2005), quatre indices de polluants sont indispensables quand on examine la qualité de l'air : les particules, l'ozone, le dioxyde d'azote et le dioxyde de soufre. Selon l'OMS [10], ces polluants sont étroitement liés aux activités humaines, par exemple : l'émission industrielle, l'émission des automobiles (cf. figure 1.6 Relation entre les agents nuisibles et les types de maladies causés [11]) et la combustion de charbons ou des déchets agricoles [12]. Ainsi, en précisant les

9. Intervalle du temps change en fonction de la taille citadine, cela peut varier de toutes les 60 minutes à toutes les 6 heures.

10. Source d'information : https://www.who.int/topics/air_pollution/fr/, consulté le 22 juin 2019.

12. Source d'information : https://www.build-green.fr/

composants des polluants, le sous-corpus Profane nous explique directement les causes du brouillard de pollution. Autrement dit, les textes profanes présentent les activités humaines comme causes principales du *wumai*, ces agents nuisibles sont accompagnés d'indications quantitatives. En plus de la publication des statistiques sur l'IQA, on donne des conseils à la population lorsque la situation de la pollution pourrait être dangereuse pour les groupes sensibles (cf. **weibo1** et **weibo3** dans le tableau 5.5.4 Études sémantiques du thème 1 dans le sous-corpus Profane). Et cela introduit en effet le deuxième (cf. section 5.6 Études sémantiques du thème 2 : Impacts de la pollution de l'air sur la santé) et le troisième thème (cf. section 5.7 Études sémantiques du thème 3 : Mesures préventives contre la pollution de l'air) que nous allons détailler dans les parties suivantes.

TAB. 5.4 – Exemple de quatre *weibo* sur l'annonce en temps réel de l'IQA de quatre villes

weibo1	***weibo* original**	# 清远空气污染指数 # 为 64，空气质量状况为良，细颗粒物 PM2o5 浓度为 28 微克/立方米，可吸入颗粒物 PM10 浓度为 56 微克/立方米，二氧化硫浓度为 12 微克/立方米，二氧化氮浓度为 26 微克/立方米，一氧化碳浓度为 0.8 毫克/立方米，臭氧最大 8 小时浓度为 116 微克/立方米.，其中环保局站点 AQI 为 67，凤城街办 AQI 为 65, 极少数敏感人群应减少户外活动。了解清远空气质量实时数据请关注
	Traduction en français	#indice de la qualité de l'air de Qingyuan# est qualifiée de bonne avec l'indice 64. La concentration de PM2,5 est 28mg/m3, celle de PM10 est 56 mg/m3, celle de SO2 est de 12 mg/m3, celle de NO2 est de 26 mg/m3, celle de NO est de 0.8 mg/m3, celle de O3 est de 116 mg/m3 pendant 8 heures. L'IQA de la station de contrôle est de 67, et celle de la rue de Fengchengjieban est de 65. Il faut réduire les activités en extérieurs pour les groupes sensibles. Pour plus d'informations sur l'IQA de Qingyuan, veuillez suivre notre compte.
weibo2	***weibo* original**	# 上海空气质量 #【今早空气质量良，实时指数 55】来看下今早空气质量。早 7 时，全市 PM2o5 平均浓度为 27.3，最高的青浦淀山湖（淀峰渔民村 1 号）39，最低的浦东张江（祖冲之路 295 号）20。实时空气质量指数 55，良。逐小时更新数据请点上海市空气质量实时发布系统

`pollution-atmospherique-les-principaux-polluants/`. Consulté en novembre 2018.

	Traduction en français	#La Qualité de l'air de Shanghai# 【Bon pour ce matin, IQA 55】Voici la qualité de l'air de ce matin. À 7h du matin, l'indice moyen de PM2,5 est de 27.3, variant de 39 pour le plus élevé au lac de Qingpudianshan (1 Village Dingfeng Pêcheur) à 20 pour le plus bas à Pudong Zhangjiang (25 Boulevard Zuchongzhi). L'IQA actuelle est qualifié de bon avec un indice 55. Pour plus d'informations sur la qualité de l'air de Shanghai, veuillez consulter le système qui met à jour les informations toutes les heures.
weibo3	***weibo* original**	02 月 01 日 13 时上海市青浦区实时空气质量平均指数为 166，中度污染，首要污染物为 PM2o5，进一步加剧易感人群症状，可能对健康人群心脏、呼吸系统有影响，儿童、老年人及心脏病、呼吸系统疾病患者避免长时间、高强度的户外锻炼，一般人群适量减少户外运动。
weibo4	**Traduction en français**	À 13h le 1 février, pour le quartier Qingpu de Shanghai, l'indice moyen de l'IQA est de 166 , la qualité de l'air médiocre, le PM2,5 constitue le polluant principal. L'air pollué présente des risques pour le groupe sensible, et pourrait nuire au système respiratoire et cardiopathique du groupe normal. Pour les enfants, les personnes âgées, les cardiaques ou les gens qui souffrent de maladies respiratoires, il faut éviter les exercices physiques prolongés et de forte intensité en extérieur, et il faut réduire les exercices extérieurs pour les autres.
	***weibo* original**	# 空气质量 #2018 年 6 月 3 日 13 时，天津全市 AQI 平均指数 104，PM2o5 指数 38，PM10 指数 67，O3 指数 207，SO2 指数 7，NO2 指数 14，CO 指数 0.7。空气质量级别为轻度污染，首要污染物 O3。如需查询各区、各点位实时更新数据，请点击 → 天津市环境空气质量 GIS 发布 天津环保发布
	Traduction en français	#Qualité de l'air# 13h le 3 juin 2018, Tianjin, l'indice moyen de IQA : 104, l'indice de PM2,5 : 38, l'indice de PM10 : 67, l'indice d'O3 : 207, l'indice de SO2 : 7, l'indice de NO2 : 14, l'indice de CO : 0.7. Si vous voulez consulter l'information les plus à jour sur les indice des polluants de quartier ou chaque zone, veuillez cliquer → 天津市环境空气质量 GIS 发布 天津环保发布

5.5.5 *Conclusion des études sémantiques du thème 1 : Causes de la pollution de l'air*

Du genre Ins au genre Profane, la description du thème 1 « causes de la pollution de l'air » passe d'un genre textuel imprécis et général, à un un genre textuel précis et concret. Et ce, que ce soit au niveau de l'emploi des mots dans

l'idée de /causes/ (synonymes *vs* terminologies chimiques), de la précision spatio-temporelle ou de la conceptualisation du thème 1.

Voici les caractéristiques communes des deux genres institutionnels concernant le premier thème « causes de pollution de l'air ».

L'Ins cherche à estomper la frontière entre ce qui relève de la pollution de l'air, phénomène étroitement lié aux activités humaines, et ce qui relève du phénomène naturel météorologique. Cela leur permettrait d'imputer les causes du *wumai* d'une région aux conditions naturelles de celle-ci, comme la configuration du terrain ou le climat, plutôt qu'aux activités humaines locales. Les deux sous-corpus se contentent d'indiquer les causes de *wumai* d'une manière générale, sans entrer dans les détails ni fournir de statistiques scientifiques ; Ils cherchent également à détourner l'attention des lecteurs en consacrant une grande partie de l'information à la discussion des problèmes du brouillard de pollution existant dans d'autres villes ou pays. Finalement, les deux font une grande attention à ce que soit mise en relief la fermeté du gouvernement pour lutter contre le brouillard de pollution.

Les textes des genres InfM et Profane témoignent d'un intérêt porté aux causes locales et d'une certaine précision scientifique. À partir de l'InfM, on commence à subdiviser et à spécifier, à l'aide de terminologies techniques, les causes en fonction des lieux considérés, selon qu'il s'agit d'une ville, d'une région. Dans le sous-corpus Profane, les types d'endroit se multiplient encore, il peut s'agir d'un quartier précis à l'intérieur d'une ville, ou encore d'une rue. Par ailleurs, les agents nuisibles changent en temps réel, ils sont même présentés avec des mesures quantitatives.

5.6 Études sémantiques du thème 2 : Impacts de la pollution de l'air sur la santé

5.6.1 *Introduction*

Lorsque nous explorons la sémantique du thème « impacts de la pollution de l'air sur la santé » dans les quatre sous-corpus, nous avons obtenu deux catégories de mots qui sont corrélés de manière étroite avec le segment pivot : 1) **les catégories de maladies causées par le brouillard de pollution**, 2) **les publics sensibles impactés par le *wumai***. Toutefois, bien que ces deux

champs lexicaux soient présents dans l'ensemble des textes, nous observons une variation dans le contenu des catégories de maladies et de publics impactés. Dans cette section, nous explorons la variation caractéristique de chaque genre textuel à l'égard du deuxième thème : « impacts de la pollution de l'air sur la santé ». Nous allons scinder les quatre genres de sous-corpus en deux groupes dans la mesure où chaque binôme partage des caractéristiques communes : le premier groupe est composé des sous-corpus Ins et InsM, et le deuxième comprend les sous-corpus InfM et Profane.

5.6.2 *Variation des types de maladies*

Dans le champ lexical du thème 2, nous avons recensé au total dix catégories de maladies (cf. liste *infra*) figurant dans l'ensemble du corpus. Parmi les types de maladies récupérés, les textes du genre Ins et InsM contiennent quatre types de maladies courantes et fréquentes causées par le *wumai* : maladie pulmonaire, cancer pulmonaire, maladie respiratoire, maladie cardio-cérébro-vasculaire. Par contraste, les catégories de maladies discutées dans les textes informels et profanes sont beaucoup plus variées. En plus des quatre catégories mentionnées ci-dessus, nous y trouvons des maladies spécifiques qui sont peu communes, tels que 皮肤病 (dermatose), 高血压 (maladie hypertensive), 糖尿病 (diabète), 佝偻病 (rachitisme), 孤独症 (autisme), 血管炎 (vascularite), 心理健康 (santé psychologique).

Concernant le cancer, un phénomène a suscité notre intérêt : le cancer (癌症) et le cancer des poumons (肺癌), qui étaient très spécifiques (sur-emploi) au commencement de l'apparition du *wumai* en Chine, sont devenus nonspécifiques (sous-emploi) dans les textes institutionnels au fur et à mesure de l'aggravation du brouillard de pollution à partir de l'année 2013. Cependant, c'est l'inverse qui se produit dans les sous-corpus InfM et Profane (cf. Figure 5.8 Distribution et évolution annuelle des mots relatifs au cancer dans les quatre sous-corpus). Si nous combinons ce phénomène avec les catégories de maladies exposées et avec la durée des maladies graves (par exemple le cancer) discutées dans les textes institutionnels, nous pouvons constater que ces derniers ont tendance à diminuer les variétés des catégories de maladies et à abréger la durée des impacts du cancer. Cela permet de garder l'homogénéité et la stabilité des textes Ins dans le temps. Par contraste, les textes informels, notamment les textes profanes s'efforcent d'exposer de manière la plus exhaustive les catégories de maladies.

Plus les maladies sont graves, plus elles deviennent saillantes dans les sous-corpus InfM et Profane. La spécificité de la rapidité de la mise à jour de l'information des sous-corpus InfM et Profane est ainsi est de nouveau mentionnée à ce propos.

Sur les types de maladies causées par le brouillard de pollution, les deux sous-corpus institutionnels sont homogènes en présentant les mêmes types de maladies courantes. Alors que les sous-corpus InfM et Profane défendent leur diversité, spécificité en abordant des maladies spécifiques (dermatose, hypertension, maladie professionnelle, etc.) ; ils se démarquent des Ins et InsM par le fait que les premiers introduisent des maladies qui ne sont guère discutées (santé psychologique, autisme) dans les textes institutionnels. Les terminologies médicales dans les textes informels-médiatiques et profanes (死亡率 (taux de la mortalité), 发病率 (indice de morbidité), 致癌物 (substance cancérigène)) rapprochent InfM du Profane. Dans la section 4.6.2.6 Analyse des collocations et des terminologies du Chapitre 4, nous avons vu que l'emploi d'une terminologie de spécialité, en l'occurrence médicale, apporte un supplément de scientificité aux textes.

5.6.3 *Variation des publics sensibles*

Le deuxième groupe des cooccurrents du thème 2 est constitué par les publics sensibles au *wumai*.Il s'agit de trois types de publics ainsi que d'une population plus large. Ils figurent dans la liste suivante :

- les personnes âgées : 老人 (personnes âgées), 中老年人 (adultes d'âge moyen et les personnes âgées) ;
- les enfants et les bébés : 儿童 (enfant), 孩子 (gamin), 小儿 (petit enfant), 婴儿 (bébé), 婴幼儿 (bébé et le nouveau né), 宝宝 (bébé) ;
- les femmes : 女士 (dame), 女人 (femme), 女生 (fille), 妈妈 (maman), 孕妇 (femme enceinte) ;
- le peuple : 群众 (grande masse), 居民 (habitant), 民众 (public), 市民 (citoyen), 人民 (people), 人们 (gens), 人人 (tout le monde), 人类 (être humain) ;

L'AFC ci-contre a pour variables les unités lexicales correspondant aux publics sensibles concernés (voir figure 5.9 Distribution des publics victimes du *wumai* dans les quatre sous-corpus). Nous distinguons quatre groupes de publics : le peuple en général, les personnes âgées, les femmes et les enfants, qui sont corrélés avec trois groupes de sous-corpus (le genre InsM et InfM sont dans un

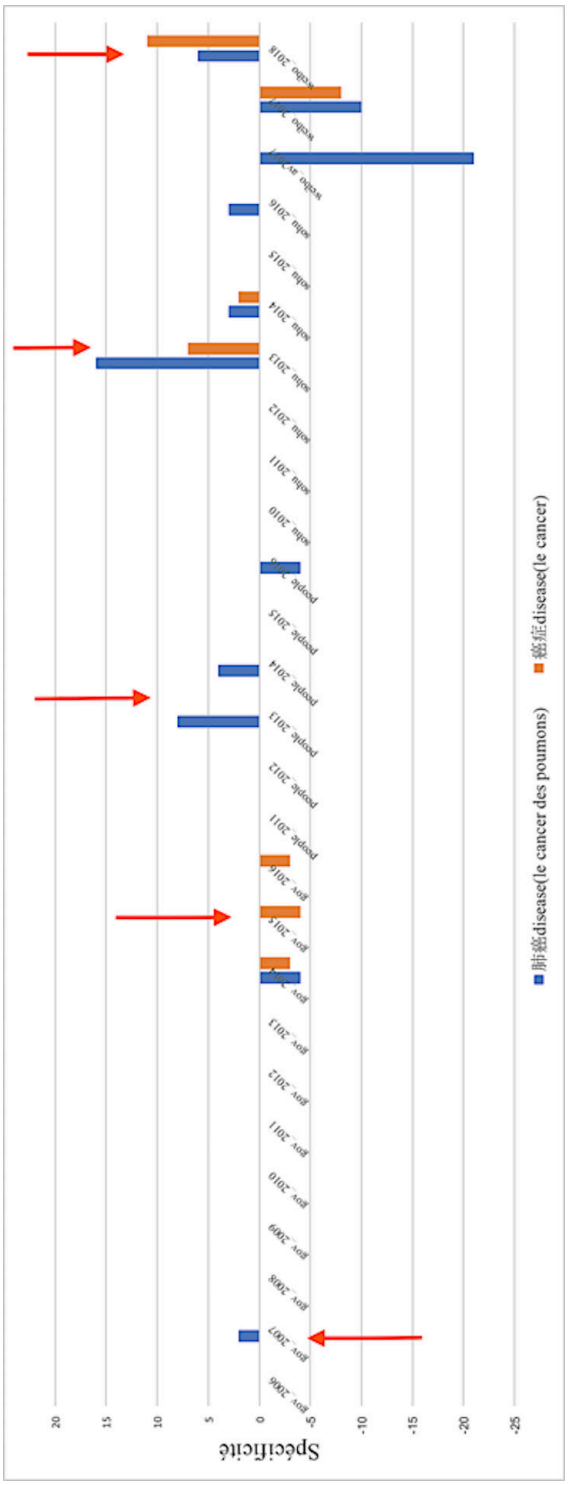


FIG. 5.8 – Distribution et évolution annuelle des mots relatifs au cancer dans les quatre sous-corpus

même groupe). La plupart des unités lexicales relatives aux victimes présentées dans le sous-corpus institutionnel constituent un type général : 居民 (habitants), 市民 (citoyens), 病人 (patients) ; par contraste, le Profane s'intéresse à deux sous-groupes : les femmes (et les femmes enceintes) et les enfants. Les mots associés à ces sous-groupes sont, respectivement, 孕妇 (femme(s) enceinte(s)), 女人 (femme), 女生 (fille), 妈妈 (maman) ; et 儿童 (enfant), 婴儿 (bébé).

InsM et InfM, qui se trouvent entre Ins et Profane (cf. 5.9 Distribution des publics victimes du *wumai* dans les quatre sous-corpus), ont des caractéristiques intermédiaires. Avec les occurrences de mots comme 人民 (population), 中老年人 (adulte d'âge moyen et personnes âgées), l'InsM se rapproche de son équivalent Ins, et l'InfM rejoint le Profane sur les deux groupes de publics spéciaux avec 女士 (dame), 孩子 (gamin), 小儿 (petit enfant), 小学生 (écolier). Ces résultats d'études portant sur le rapprochement entre les occurrences de « groupes personnels » et les genres de sous-corpus correspondent bien à la distribution des pronoms personnels dans les quatre sous-corpus que nous avons analysés dans la section 4.6.2.5 Analyse des pronoms personnels, où « elle et elles » sont deux pronoms personnels spécifiques du genre Profane. Par conséquent, nous pouvons constater qu'au niveau de la présentation des publics sensibles impactés par le *wumai*, une tendance «**du plus général et abstrait au plus spécifique et concret**» se vérifie du genre Ins au genre Profane, en passant par l'InsM et l'InfM.

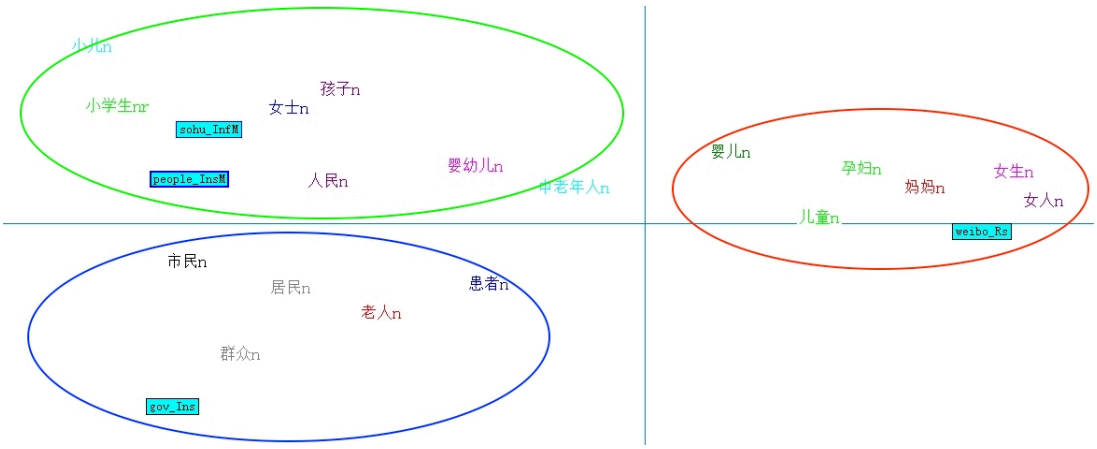


FIG. 5.9 – Distribution des publics victimes du *wumai* dans les quatre sous-corpus

5.6.4 *Conclusion des études sémantiques du thème 2 : Impacts de la pollution de l'air sur la santé*

À travers les résultats d'analyse du thème 2 sur les deux groupes de cooccurrents : les types de maladies causées par le *wumai* et les publics sensibles du *wumai*, le genre Ins s'oppose complètement au genre Profane, et les genres InsM et InfM occupent une place intermédiaire entre ces deux extrémités. Si le genre Ins présente ces caractéristiques d'**homogénéité**, d'**uniformité**, de **stabilité**, de **généralité** et d'**abstraction**, le genre Profane, à l'opposé de son « rival », se caractérise pour sa part par la **diversité**, la **variation**, la **spécificité**, la **précision**, la **concrétisation** et la **complétude**, que ce soit en matière de l'exposé des types de maladies, de l'évolution temporelle de ces derniers, ou encore des publics sensibles au *wumai*.

5.7 Études sémantiques du thème 3 : Mesures préventives contre la pollution de l'air

5.7.1 *Introduction*

L'étude sémantique du thème 3 sera abordée dans les sections à venir en suivant le même procédé : nous divisons les quatre genres des sous-corpus dans deux groupes séparés, le groupe 1 est composé d'Ins et InsM, le groupe 2 d'InfM et Profane. Et ce groupement en binôme est justifié par leurs caractéristiques proches.

5.7.2 *Études sémantiques du thème 3 dans les sous-corpus Ins et InsM*

5.7.2.1 Mise en avant des mesures préventives adoptées par le gouvernement

Nous explorons le thème 3 en combinant les mots associés au sème /mesure/ (治理 (résoudre), 措施 (mesure), 防治 (prévenir), 应对 (affronter) et à l'isotopie /pollution de l'air/ (雾霾, 空气污染, 大气污染, etc.). À travers les cooccurrents récupérés, nous concluons comme suit les mesures proposées par les institutions gouvernementales pour lutter contre le *wumai* :

– réduction de l'émission industrielle : 减排 (réduire l'émission), 限产 (limiter la capacité productive), 停产 (mise en arrêt de la production), 关厂

(fermeture des usines), 淘汰 (éliminer/élimination) ;
- utilisation de l'énergie verte ou renouvelable : 绿色能源, 清洁 ;
- législation : 环保法 (lois sur la protection de l'environnement), 立法 (législation)
- contrôle de la circulation des automobiles : 单双号 (impaire-paire plaque d'immatriculation), 限行 (contrôle de la circulation des véhicules).

Ces mesures partagent ont pour point commun de s'intégrer dans des réformes politiques établies par les institutions gouvernementales à court ou à long terme. Cet aspect, conformément aux caractéristiques du genre Ins et InsM (cf. sections 4.6.2.6 Analyse des collocations et des terminologies et 4.6.2.9 Analyse de la nominalisation du Chapitre 4), est confirmé par les mots suivants qui apparaissent dans le champ lexical du thème 3.

Au niveau dialectique, nous avons repéré des mots emphatiques tels que 加大 (élargir), 加强 (renforcer), 加快 (accélérer), 强调 (souligner), ainsi que des verbes modaux 要 (falloir), 需 (devoir), 必需 (devoir), 必要 (nécessiter). Ces champs sémantiques sont associés à des lexiques qui impliquent le futur (未来十年 (dans les dix ans à venir)) pour exposer ce que l'on envisage comme mesures préventives contre le *wumai* ; ou encore le passé (近些年 (les dernières années)) pour exposer les mesures déjà adoptées visant à lutter contre le *wumai*. Voici la figure suivante qui présente la concordance de ces mots.

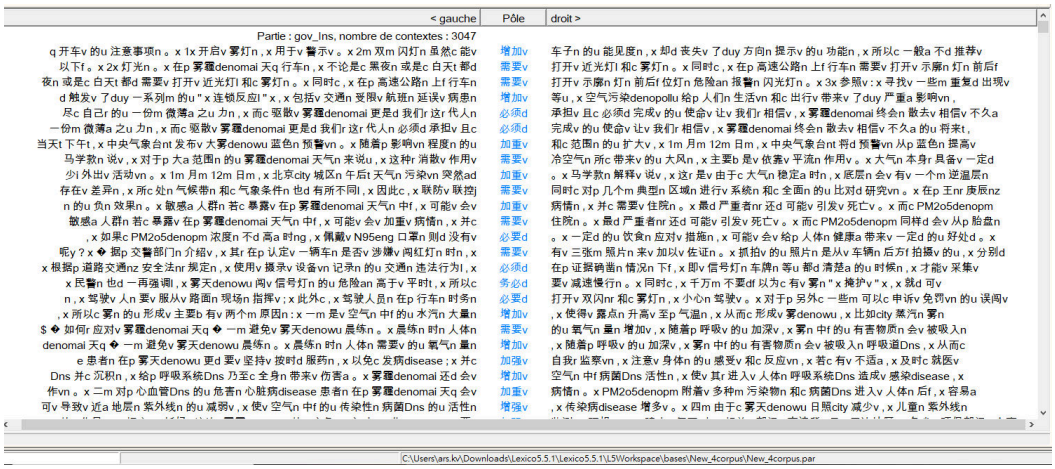


FIG. 5.10 – Concordance des mots emphatiques dans les textes institutionnels

Au niveau dialogique, beaucoup d'entités nommées sont employées dans la

présentation du thème 3 pour mettre en avant le rôle de l'État : l'État (国家), les établissements gouvernementaux (环保部门, 国务院) et les titres professionnels 总理 (premier ministre), 人大代表 (représentant du peuple), 督查人员 (les superviseurs).

De ce point de vue, le sous-corpus institutionnel met en évidence les mesures préventives adoptées par le gouvernement afin de montrer qu'il a assumé ses responsabilités et que son rôle a été actif et positif (应急, 加快, 加强, 加大, 解决, 治理, 宣战) dans la lutte contre la pollution de l'air (改革, 预案, 持久战).

Fig. 5.11 – Cooccurrents du thème 3 dans GOV

5.7.2.2 Souffler le *wumai* par le vent

Dans les cooccurrents de « mesures préventives contre la pollution de l'air », le verbe 吹散 (être soufflé par le vent) attire notre attention. Inscrit dans la condition préalable où le *wumai* est considéré par Ins et InsM comme un phénomène naturel météorologique 雾霾天气 (le temps *wumai*) (voir section 5.5.1 Études sémantiques du thème 1 dans le sous-corpus Ins), cette mesure concrète correspond à ce trait du genre Ins : on compte sur le vent pour disperser le *wumai*. Voici deux exemples représentatifs respectivement extraits du sous-corpus Ins et InsM : Exemple 1) 一股冷空气带来的偏北风，吹散笼罩我国中东部的雾霾 (Le vent froid du nord a soufflé le *wumai* qui a embrumé la région centrale et celle de l'est de la Chine) ; Exemple 2) 北京将出现三四级北风，有望吹散雾霾 (Le vent de force 3-4 venant du nord soufflera le *wumai*). La figure suivante présente la concordance du verbe 吹散 (être soufflé par le vent) dans les textes Ins et InsM.

partie gauche	pivot	partie droite	texte
地uv ,x 但c 由于c 路径n 偏东n ,x 不能v 彻底ad	吹散v	雾霾denomai ,x 只d 起到v 缓解v 作用v 。x 1921m 日m	corpus:gov
蒸发v ; x 二m 是v 风速n 增大v ,x 将雾n	吹散v	或c 抬升v 成云nr ; x 再有v 就是d 湍流n 混合vn	corpus:gov
范围n 雾霾denomai 天气n 逐渐d 减弱v 冷空气n 即将d " x	吹散v	" x 雾霾denomai 山东province 继续v 发布v 雾霾denomai 橙色n 预警vn	corpus:gov
范围n 雾霾denomai 天气n 逐渐d 减弱v 冷空气n 即将d " x	吹散v	" x 雾霾denomai	corpus:gov
,x 有利于v 户外活动n 和c 交通n 出行v 。x 冷空气n 虽c	吹散v	了duy 雾霾denomai ,x 但c 同时c 也d 带来v 了duy 大风n	corpus:gov
,x 北京city 出现v 5m 级q 左右m 大风n 。x 大风n	吹散v	了duy 雾霾denomai ,x 同时c 也d 卷起v 近a 地面n 的u	corpus:gov
上面f 而c 起v 雾n ,x 并且c 不易a 被p 风n	吹散v	。x " x 孙冷nr 说v ,x " x 因此c	corpus:gov
北京city 将d 出现v 三m 四级m 的u 北风n ,x 有望v	吹散v	雾霾denomai 改善v 空气n 状况n ,x 但c 在此之前i ,x 整体n	corpus:gov
北京city 将d 出现v 三m 四级m 的u 北风n ,x 有望v	吹散v	雾霾denomai 改善v 空气n 状况n ,x 但c 在此之前i ,x 整体n	corpus:gov
北京city 将d 出现v 四五级m 的u 北风n ,x 届时d 有望v	吹散v	雾霾denomai ,x 明显改善nr 空气质量n 。x 气象局n 冷空气n 携v	corpus:gov
北京city 将d 出现v 四五级m 的u 北风n ,x 届时d 有望v	吹散v	雾霾denomai ,x 明显改善nr 空气质量n 。x 1m 月m 23m 日m	corpus:gov
健康a 不d 应v 成为v 等待v 中f 的u 牺牲品n 。x	吹散v	雾霾denomai ,x 有所作为i 的u 第一步m 是v 打破v 沉默a ,x	corpus:gov
,x 开诚布公n 。x 愿v 真抓实干i 化为v 长效a 动风n ,x	吹散v	我们r 身边s 和c 心头s 的u 雾霾denomai 。x 气象局n	corpus:gov
冷空气n 即将d " x	吹散v	" x 雾霾denomai 冷空气n 即将d " x 吹散v "	corpus:gov
x 吹散v " x 雾霾denomai 冷空气n 即将d " x	吹散v	" x 雾霾denomai 冷空气n 即将d " x 吹散v	corpus:gov
吹散v " x 雾霾denomai 冷空气n 即将d " x	吹散v	" x 雾霾denomai 2013m 年m 01m 月m 31m 日m	corpus:gov
影响vn 将d 逐渐d 减轻v 。x 冷空气n 即将d " x	吹散v	" x 雾霾denomai 科学n 生活vn 戴v 口罩n 能否v	corpus:gov
范围n 雾霾denomai 天气n 逐渐d 减弱v 冷空气n 即将d " x	吹散v	" x 雾霾denomai 中东部nt 将d 有v 雨雪n 降温n 天气n	corpus:gov
江西province 降温n 大风n	吹散v	雾霾denomai 2013m 年m 02m 月m 19m 日m 14m	corpus:gov
" x 入v 鼻ng 。x 9m 日m ,x 大风n	吹散v	京城ns 雾霾denomai ,x 但c 带来v 了duy 同样d 恼人a 的u	corpus:gov
降雪n 天气n ,x 连日d 的u 雾霾denomai 也d 被p 大风n	吹散v	。x 吕维兰nr 用p 真心d 爱心n 细心n 恒心n 帮助v	corpus:gov
冯琦nr 相关v 链接n 山东province 青岛city 冷空气n 带来v 降雪n	吹散v	雾霾denomai 石家庄city 冷风吹nr 长春city 雾霾denomai 消散v	corpus:gov
几多m 风雨n 山东province 青岛city 冷空气n 带来v 降雪n	吹散v	雾霾denomai	corpus:gov
年m 中国ns 天气n 山东province 青岛city 冷空气n 带来v 降雪n	吹散v	雾霾denomai 石家庄city 冷风吹nr 雾霾denomai 散v 长春city 雾霾denomai	corpus:gov
风够v 大n ,x 那么r 污染物n 就d 可以c 很快d 被p	吹散v	。x 但是c 风n 不能v 一直d 刮v ,x 冷空气n 活动vn	corpus:gov
比较d 大a ,x 是v 3m 级q 东风n ,x 有利于v	吹散v	空气n 中f 粉尘n 等u 颗粒物n ; x 加上v 太阳n	corpus:gov

FIG. 5.12 – Concordance du 吹散 (souffler par le vent) dans GOV

5.7.2.3 Chacun sa responsabilité

En explorant les segments répétés du thème 3 dans les textes institutionnels-médiatiques, le segment « 政府不治理雾霾的原因 » (les raisons pour lesquelles le gouvernement ne prend pas lui-même de mesures pour contenir le brouillard de pollution) attire notre attention. Afin de comprendre pourquoi ce propos a été introduit dans les textes institutionnels-médiatiques, nous nous sommes reportés à son contexte textuel, et avons trouvé des phrases où figure la phrase qui contient le segment : «我为治理雾霾做了哪些 (Qu'est-ce que j'ai fait pour résoudre le brouillard de pollution)», « 治理雾霾要从孩子抓起 (Pour résoudre le problème de *wumai*, il faut commencer par sensibiliser les enfants)», ou encore« 治理雾霾需从我做起 (Chacun sa propre responsabilité pour résoudre le brouillard de pollution)». Le pronom personnel « je » figure dans ces phrases. Toutefois, il faut noter qu'il s'agit ici d'un « je » générique, qui veut dire « n'importe qui ». Le « je » constitue en fait un pronom impersonnel. Ce qui correspond bien à nos résultats d'analyses sur les caractéristiques spécifiques du genre InsM au niveau de l'utilisation des pronoms personnels (cf. section 4.6.2.5 Analyse des pronoms personnels du Chapitre 4). De plus, on associe les verbes modaux (要 et 需) dans le sens de /falloir/ pour renforcer le ton injonctif dans le but de faire agir tout le

monde, et de faire entrer dans la tête l'idée que la responsabilité de « résoudre le *wumai*» incombe à chaque membre de la communauté, et ce depuis l'enfance (孩子).

5.7.3 *Études sémantiques du thème 3 dans les sous-corpus InfM et Profane*

Par rapport aux propositions globales des institutions, les mesures préventives discutées dans les textes informels-médiatiques et profanes sont plus concrètes et pratiques. En observant les cooccurrents du segment « mesures préventives de la pollution de l'air », nous avons répéré deux types de mots partagés dans les *weibo* et les textes informels-médiatiques : 1) des publicités postées par des vendeurs, (ce qui est catégorisé par (Yang, Yang et Zhou, 2015) dans leur étude comme « avantage aperçu », cf. section 2.4.4 Travaux combinant la SI et la textométrie du Chapitre 2), qui cherchent à écouler des produits tels que des masques ou des purificateurs d'air en vantant leurs propriétés ou fonctionnalités protectrices ; parmi celles-ci les terminologies spécifiques et les noms propres comme N95[13], 活性炭 (charbon actif), 防尘 (pare-poussière), 医用 (médical), 纸质 (en papier), 一次性 (jetable) pour les masques, et le 杀菌 (stérilisant), HEPA (High Efficiency Particulate Air)[14], 抗病毒 (anti-viral) pour les purificateurs d'air ; 2) des messages publiés par des utilisateurs de WEIBO, qui proposent des recettes de gastro-thérapie[15] permettant de réduire les effets négatifs de la pollution de l'air, tels que 清肺 (détoxifier les poumons), 排毒 (détoxiquer), 补钙 (le calcium), 雪梨 (la poire ; ou encore des méthodes pratiques pour renforcer le système immunitaire, tel que 开窗通风 (ouvrir la fenêtre), 锻炼 (faire des exercices), 跑步 (faire du jogging).

D'ailleurs, la présence de nombreux signes typographiques ou icônes d'émotion, des mots d'interjections (啊 (oh là là)) ou des néologismes (神器 (objet magique),

13. Selon la Food and Drug Administration des États-Unishttps://www.fda.gov/medical-devices/personal-protective-equipment-infection-control/masks-and-n95-respirators. Le respirateur N95 est un appareil de protection utilisé pour filtrer des particules en suspension dans l'air. Et l'appellation « N95 » signifie que le respirateur peut bloquer au moins 95% des particules de diamètre inférieur ou égal à 0,3 µm.

14. La dénomination HEPA est appliquée à tout dispositif capable de filtrer, ici, il s'agit d'un filtre à air à particules aériennes à haute efficacité. Selon le site https://www.engineersedge.com/filtration/hepa_filter.htm, le filtre HEPA de l'air peut purifier l'air et retenir 99,97% des particules de diamètre inférieur ou égal à 0,3 µm.

15. Cf. La définition de la gastro-thérapie sur le sitehttps://www.jdbn.fr/la-gastro-therapie-vous-connaissez/. Consulté en décembre 2019.

萌物 (truc mignon) [16] dans les *weibo* permettent de :

- dynamiser les textes écrits : figure d'émotion ;
- renforcer les émotions : !!!, émoticônes et mots exclamatifs ;
- créer des sujets de discussion : # # ;
- transmettre des connaissances spécialisées dans certains domaines techniques ou scientifiques : terminologies techniques ou mots anglais.

Ces traits typographiques ou terminologiques peuvent non seulement inciter plus d'utilisateurs à participer à la discussion, et à diffuser des informations d'une large ampleur. Ils mettent en évidence les caractéristiques des manières de traitement du troisième thème du genre Profane : émotionnel, dynamique, innovatif, créatif, spécifique. Ces caractéristiques correspondent aux résultats d'études sur celles du genre Profane (cf. sections 4.5.3 Caractéristiques infratextuelles du Profane et 4.7.4 Caractéristiques intratextuelles du genre Profane) du Chapitre 4.

5.7.4 *Conclusion des études sémantiques du thème 3 : Impacts de la pollution de l'air sur la santé*

Nous pouvons interpréter la sémantique du thème « mesures préventives contre la pollution de l'air » traité dans les quatre genres de sous-corpus selon deux angles : dialectique et dialogique.

Pour les textes du genre Ins et InsM, au niveau dialectique, les mesures préventives sont abordées de manière globale (réduction de l'émission du charbon, restriction des véhicules automobiles, fermeture d'usines fortement polluantes, utilisation des énergies vertes ou renouvelables, etc.), et on les intègre soit dans les mesures contre le *wumai* à adopter en urgence dans le futur, soit dans les actions déjà accomplies par l'État dans le passé. Tous les thèmes qui sont liés à la pollution de l'air deviennent un problème stratégique pour le gouvernement. Au niveau dialogique, quand il s'agit de proposer des mesures préventives générales, le rôle de l'État, des établissements publics ou des gouvernants est mis en avant. Au contraire, lorsque des mesures préventives concrètes sont recommandées, on constate qu'au niveau dialectique le *wumai* est présenté comme un phénomène naturel, qui peut être « soufflé par le vent ». De même, au niveau dialogique,

16. Ici, en l'occurrence, ces deux néologismes sont utilisés pour qualifier soit le purificateur, soit les masques de visage.

au lieu de souligner la responsabilité de l'État, on met en avant celle de chaque membre de la société, y compris les enfants.

Quant aux textes du genre InfM et Profane, au niveau dialogique, les institutions ne sont plus au centre de l'action, les utilisateurs ou les lecteurs ordinaires occupent une place majeure pour proposer aux autres utilisateurs des mesures préventives courantes et faciles à adopter. Au niveau dialectique, grâce au dispositif interactif et aux caractéristiques innovantes, dynamiques et techniques du genre Profane, les informations circulent en tous sens. Elles peuvent être répandues de manière à attirer de nouveaux lecteurs ou utilisateurs, comme pour bénéficier à ceux-ci en leur faisant participer aux efforts émis pour lutter contre contre le *wumai*.

5.8 Conclusion des analyses sémantiques des trois thèmes

À l'aide de la méthode de SI, nous avons identifié trois thèmes principaux qui sont partagés par les quatre genres de textes. Cependant, ces quatre genres textuels se distinguent les uns des autres par la description et l'interprétation sémantique de chaque thème.

Les causes de la pollution de l'air (la topographie, le temps) et les mesures préventives proposées (soufflé par le vent) dérivent toutes du postulat que *wumai* est « un phénomène naturel météorologique ». Les textes institutionnels parlent des trois thèmes tantôt dans un **reportage météorologique diffusé par les institutions gouvernementales**, tantôt dans les « commentaires médiatiques donnés par les experts d'institutions », tantôt dans un « rapport de travail du gouvernement ». L'objectif est de faire connaître les nouvelles et des décisions de l'État dans le futur ou de déclarer les résultats déjà obtenus dans la lutte contre le brouillard de pollution. Et ce, avec 1) le transfert dialogique - de l'État/établissements publics/ministre/expert aux particuliers, enfant et toute la société ; 2) les tournures dialectiques au sens de /falloir/, /devoir/ pour présenter un ton emphatique et injonctif.

À travers les analyses des trois thèmes dans les sous-corpus institutionnels (Ins et InsM), nous avons constaté que tous les thèmes autour du problèmes de la pollution de l'air sont unifiés par les institutions dans leur idéologie politique. Leur objectif est de mettre en avant les actions de l'État quel que soit le thème

abordé. En plus, la centralisation des actions de l'État, la transformation de la pollution de l'air en phénomène naturel météorologique, ainsi que le fait de faire peser la responsabilité d'action sur chaque membre de la société produisent ensemble un effet latéral : détourne l'attention de la population de ce que l'État a fait ou va faire comme efforts afin de résoudre le problème de *wumai*. À la fin, nous avons détecté une certaine « anomalie » dans le sous-corpus Ins : plus la pollution de l'air est jugée grave, plus Ins a tendance à diminuer les variétés des catégories de maladies et à abréger la durée des impacts du cancer.

Le sous-corpus InfM assume son rôle médiatique, qui interprète les trois thèmes dans un style de « commentaire médiatique ». Il polémique sur les trois thèmes en comparant le *wumai* en Chine avec les situations de pollution dans d'autres pays (Londres, Paris, États-Unis), qui sont causées par les activités humaines. Le sous-corpus InfM s'approche du sous-corpus Profane en matière d'interprétation sémantique des trois thèmes. À partir de l'InfM, on commence à s'intéresser à une région ou un groupe spécifique qui est sensible au *wumai* (femme et enfant). Les causes, les impacts et les mesures commencent à se diversifier, se multiplier, se spécialiser et se concrétiser.

Le genre Profane intensifie les caractéristiques du genre InfM que ce soit au niveau de la présentation des trois thèmes ou bien au niveau de l'interprétation de ces derniers. Le genre Profane présente non seulement les causes du *wumai* en lieu et temps réel selon l'endroit, il indique les agents nuisibles à ce dernier sous forme des terminologies scientifiques (SO2, CO2, CO, O3, etc.) avec la mesure quantitative de l'indice de la qualité de l'air. Les textes profanes essaient d'exposer de la façon la plus exhaustive les catégories de maladies (le cancer, la maladie professionnelle, dermatose, etc.). En même temps, ils spécifient à la fois les catégories de maladies causées par la pollution de l'air, et les types de publics (femmes, femmes enceinte, enfants, bébé, etc.) qui sont sensibles au *wumai*. À la fin, les mesures préventives proposées par les utilisateurs sont simples, concrètes et pratiques à adopter. Le dispositif du genre Profane favorise la circulation et la transmission de ces informations.

Le genre Profane constitue un moyen de pression contre le gouvernement, car d'une part, il évoque des choses qui sont moins mentionnées dans les médias traditionnels, par exemple le cancer lié au *wumai*, et les indices explicites des agents nuisibles dans l'air pollué ; d'autre part, grâce à son dispositif interactif,

le genre Profane accorde une place plus importante aux utilisateurs ordinaires et leur offre une plateforme pour échanger, faire circuler voire faire enrichir les informations et les expériences pratiques personnelles, de manière que ces informations soient accessibles pour chaque utilisateur/lecteur. Ainsi, dans cette arène de communication, les informations circulent très rapidement, les utilisateurs peuvent trouver facilement des informations qui ne sont pas présentes dans les médias traditionnels.

Chapitre 6

Synthèse et résultats

6.1 Résultats d'étude

Dans les sections suivantes, nous allons présenter les résultats d'études en fonction du genre textuel auquel ils correspondent. La présentation sera axée sur trois paliers : palier macrosémantique, auquel nous présentons les caractéristiques du genre textuel étudiées à deux niveaux (infratextuel et intratextuel) ; paliers méso- et microsémantique, auxquels nous présentons les thèmes identifiés et leurs sémantiques interprétatives.

6.1.1 *Résultats d'études du sous-corpus institutionnel*

Comme ce qui a été avancé dans l'introduction, les textes institutionnels ont pour objectif principal de transmettre des décisions des hiérarchies, et de faire connaitre au grand public les affaires politiques nationales ou internationales. Leur mode de transmission des informations suit l'ordre univoque « de haut (décideurs) en bas (lecteurs)». Il forme un système d'exportation unidirectionnelle statique et distante afin de transmettre des informations homogènes (politiques) aux lecteurs dans une sphère uniformisée. Basé sur ce système spécifique au sous-corpus institutionnel, nous avons étudié les caractéristiques du genre institutionnel, identifié les thèmes relatifs au corpus du brouillard de pollution et interprété la sémantique des thèmes relevés.

6.1.2 *Genre institutionnel*

Au niveau macrosémantique, les caractéristiques du sous-corpus institutionnel sont récupérées à deux niveaux — infratextuel et intratextuel — peuvent être qualifiées avec des mots-clés suivants : identitaire, uniforme, injonctif (emphatique), général, distant, statique.

D'abord, la forte **identité** du sous-corpus institutionnel est mise en évidence par les éléments représentatifs de la Chine et du gouvernement chinois, tels que le logo-type en emblème national, le fond de couleur en rouge, et certains traits spécifiques d'institution, par exemple les grands portraits des dirigeants, la prédominance des entités nommées des noms de pays, des mentions de titres de professions, des établissements publics, etc.

Ensuite, ses caractères de l'**uniformité** et de l'**homogénéité** se traduisent par plusieurs aspects : en termes d'architecture du site, le sous-corpus institutionnel privilège la place centrale aux « unes » des dirigeants chinois, mais défavorise le cadre réservé aux commentaires ; en termes de circulation de l'information, toujours de haut en bas, et non l'inverse ; en termes de langage adopté, les textes institutionnels sont rattachés au discours énonciatifs, introduits avec des phrases longues ou subordonnées, qui sont unies par des conjonctions de subordination, des connecteurs logiques, et du vocabulaire descriptif ; en termes d'émetteur d'information, il s'agit de l'État, des établissements d'institution, des personnels gouvernementaux, etc. ; en termes de contenu, les informations politiques constituent la thématique majeure du sous-corpus institutionnel.

La tonalité **impérative** des textes institutionnels est marquée par la présence de nombreux termes impératifs et de verbes de modalité sur-employés, qui confèrent à ces textes un ton nettement injonctif et insistant. Nous y trouvons également des adverbes intensifieurs, de nombreuses négations et la figure du parallélisme.

L'utilisation intensive de termes qui marquent le passé et le futur, ainsi que l'absence de cadre réservé aux commentaires marquent ensemble son caractère **distant** entre les institutions et le monde réel, et entre les producteurs des articles et les lecteurs ; et son caractère **statique** est actualisé par l'utilisation des expressions figées, telles que certaines expressions toutes faites et les expressions de l'ordre ; ou par l'emploi des signes de ponctuation correspondant à la norme standard.

Le nombre d'occurrences élevé de termes exprimant des concepts abstraits (le futur, la conscience, le développement) ou généraux et synthétiques (habitant, patient, citoyen) sont une spécificité du sous-corpus institutionnel ; ces termes constituent une autre caractéristique typique — **abstrait** — du sous-corpus institutionnel.

6.1.3 *Sémantiques des trois thèmes dans le sous-corpus institutionnel*

Au niveau mésosémantique, en explorant les sèmes qui apparaissent de manière récurrente avec l'isotopie /pollution de l'air/ du corpus, nous avons récupéré au total trois thèmes : « causes de la pollution de l'air » (/cause/+/pollution de l'air/), « impacts de la pollution de l'air sur la santé » (/impact/+/pollution de l'air/) et « mesures préventives contre la pollution de l'air » (/mesure/+/pollution de l'air/). Ces trois thèmes sont partagés par les quatre genres de sous-corpus.

L'interprétation sémantique de ces trois thèmes a été effectuée à travers les cooccurrents de chaque thème. Ces cooccurrents sont actualisés par les lexiques au niveau microsémantique. Nous interprétons la sémantique de chaque thème selon deux angles : dialectique et dialogique.

Le sous-corpus institutionnel unifient tous les trois thèmes dans son idéologie politique. Ils appliquent une stratégie de déplacement : D'abord, la transformation des causes de la pollution de l'air. Dans l'analyse des causes de la pollution de l'air dans les textes institutionnels, on transforme la pollution atmosphérique en phénomène naturel météorologique (雾霾天气). Ainsi, les causes sont désormais attribuées plutôt aux conditions naturelles topographiques (多山 (montagne), 盆地 (bassin)) ou météorologiques (干旱 (sécheresse)，少雨 (manque de pluie)), qu'aux activités humaines.

Ensuite, la transformation des types de maladies. Dans l'étude du thème 2 « les impacts de la pollution de l'air sur la santé », nous notons qu'au fur et à mesure de l'aggravation de la pollution de l'air, les types de maladies sont pourtant passé de ceux plus graves (肺癌 (le cancer des poumons) et 癌症 (le cancer)) à ceux moins graves (肺炎 (pneumonie), 呼吸系统疾病 (les maladies respiratoires)).

À la fin, la transformation des mesures préventives contre la pollution. La présentation globale des mesures préventives contre la pollution est transformée en une louange des actions gouvernementales face au *wumai*. Dans lequel, le rôle du gouvernement est mis en avant au niveau dialogique. Dans ce cas-là, au niveau dialectique, les termes injonctifs ou emphatiques (négation, verbe modal) ainsi que la figure du parallélisme sont massivement utilisés pour organiser les informations à un rythme soutenu, et pour les transmettre de manière plus efficace. Cependant, lorsqu'il s'agit de discuter des mesures concrètes contre la pollution de l'air, les rôles dialogiques sont déplacés du gouvernement, des établissements

publics, du ministre, des experts, vers le « particuliers, enfants et toute la société ». Tout le monde est responsable (人人有责) pour lutter contre le brouillard de pollution : 治理雾霾需从我做起 (Chacun doit prendre la responsabilité pour régulariser le brouillard de pollution), 治理雾霾要从孩子抓起 (Il faut sensibiliser les enfants aux mesures préventives contre la pollution de l'air.).

D'une vue panoramique sur les caractéristiques du sous-corpus institutionnel, nous pouvons conclure que les textes institutionnels constituent un sous-corpus identitaire qui transmet d'un ton insistant les informations du gouvernement à la population, dont le contenu est homogène et la forme est statique. Au niveau dialogique, le rôle du gouvernement est mis en avant à travers le sur-emploi des termes relatifs aux établissements gouvernementaux et aux titres professionnels d'institution ; au niveau dialectique, son style injonctif et emphatique est accentué par les verbes modaux, le parallélisme, la négation, et certains adverbes, etc.

6.1.4 *Résultats d'études du sous-corpus profane*

Par contraste, le sous-corpus profane est polymorphe et polyphonique. Son caractère polymorphe est manifesté à la fois par sa large gamme de forme de présentation (blog, chat, réseau social) et par son contenu multidimensionnel (sport, faits divers, publicité, économie, politique, etc.) ; alors que son trait polyphonique est mis en évidence par la diversité d'énonciateurs : citoyen, expert, journalistique, officiel. La voix de toute classe sociale trouve une place dans cette arène de réalité mixte. Ces deux caractères du sous-corpus profane permettent d'une part d'accorder une place plus importante aux utilisateurs ; d'autre part, d'accélérer la circulation des informations. L'échange dynamique et interactif des utilisateurs est en faveur de l'hétérogénéité des informations, en cela que l'on peut échanger, faire circuler voire créer, compléter ou problématiser les informations dans une zone publique ou privée. Ainsi, le sous-corpus profane construit un système multidimensionnel de circulation de l'information hétérogène dans un ordre du type « importation - exportation - importation ».

6.1.5 *Genre profane*

Au niveau macrosémantique,le sous-corpus profane offre une grande diversité architecturale, thématique et procédurale que ce soit au niveau infratextuel ou intratextuel.

Au niveau infratextuel, les caractères **interactif** et **communicationnel** résultent de la mise en relief du cadre dédié aux commentaires du site WEIBO, qui favorise la communication mutuelle entre les utilisateurs. Ce qui garantit dans certaine mesure la **rapidité** de la mise à jour des information étant donné son dispositif « user-generated-content ». Ses particularités de la **multimodalité**, du **dynamisme** se traduisent par le réaménagement et l'hybridation des textes écrits, des clips-audios, clip-vidéos, et des images animées, des émoticônes et des ponctuations inventées (!+ ou !?+). Sa **multidimensionnalité** réside dans les thématiques très variés qui peuvent aller des nouvelles politiques jusqu'aux divertissements.

Au niveau intratextuel, le sous-corpus profane suit principalement l'ordre de l'*exposé* et s'inscrit dans un style plutôt oral : l'utilisation dominante des 1ère et 2ème personnes du singulier et du pluriel, de l'indicatif présent et d'un vocabulaire simple. Son caractère **expressif** se construit à travers des néologismes (囧), des émoticônes, des interjections (啊), l'onomatopée, l'utilisation non conforme d'une ponctuation expressive (!!!,?!), des signes techniques (# et @), et des métaphores.

D'ailleurs, le sur-emploi de la 3ème personne du singulier et du pluriel, ainsi l'apparition fréquente de la catégorie explicite des noms féminins et des noms de personne : 女人 (femme), 女生 (fille), 妈妈 (maman), 母亲 (mère), 孕妇 (femme enceinte), 小孩 (gamin), 儿童 (enfant), 婴儿 (bébé), montrent que le sous-corpus profane accorde une attention particulière aux groupes sensibles composés des femmes et enfants. Tout ceci rend le sous-corpus profane plus concret et précis.

6.1.6 *Sémantiques des trois thèmes dans le sous-corpus profane*

A partir de mêmes procédés et thèmes, nous interprétons la sémantique de chaque thème au niveau méso- et microsémantique. Le sous-corpus profane traite les trois thèmes dans de multiples dimensions et de manière détaillée et ciblée.

Le sous-corpus profane, quand il présente les causes de la pollution de l'air (thème 1), emprunte directement des terminologies chimiques (SO2, CO, CO2, NO, O3) pour indiquer les agents polluants. De plus, par rapport aux textes institutionnels qui illustrent globalement les causes du *wumai* dans l'ensemble de la Chine, les textes profanes annoncent de manière circonstanciée les agents polluants de l'air en lieu et en temps réel selon les villes/régions, voire les quartiers d'une ville. Ce qui relèvent en fait plusieurs types de sources potentiellement

responsables de la pollution de l'air.

Exposition des impacts par la pollution de l'air (thème 2) est spécialisée dans le domaine de la médecine : plusieurs catégories de maladies et symptômes sont présentées dans les textes profanes, du plus grave comme le 癌症 (cancer), 肺癌 (cancer des poumons) au moins grave tel que 打喷嚏 (éternuements) et 咳嗽 (tousser), et du plus quotidien comme par exemple 鼻炎 (rhinite), 感冒 (rhume) au moins fréquent tel que 孤独症 (autisme) et 皮肤病 (dermatose), 眼病 (ophtalmopathie). De même, les publics sensibles sont divisés de manière très détaillée, notamment concernant les femmes et les enfants, selon toutes les tranches d'âge : 婴儿 (bébé), 儿童 (enfant), 女孩 (fille), 母亲 (maman).

Proposition des mesures préventives contre la pollution de l'air (thème 3) est spécialisée dans le domaine de la publicité et de la gastro-thérapie : de nombreux utilisateurs-vendeurs proposent leurs produits anti-smog avec les *weibo*, par exemple le masque ou le purificateurs d'air. Ils se servent des terminologies techniques, telles que N95, l'HEPA, et 活性炭 (charbon actif), 防尘 (pare-poussière), 一次性 (jetable), 杀菌 (stérilisant), et des néologismes, tels que 神器 (objets magiques), 萌物 (truc mignon) pour promouvoir l'efficacité de leurs marchandises. Dans la présentation de ces produits, on utilise massivement des mots d'emprunt (caractères japonais ou mots anglais) pour moderniser leur slogan publicitaire. De plus, des signes sémiotiques (émoticônes, ponctuation exclamative (!!!), # ou @), des mots interjections et des onomatopées (啊 (oh là là)), sont davantage présents dans les publicités pour dynamiser le contenu de la publicité, et pour attirer l'attention des lecteurs, qui sont pour eux de potentiels clients.

En plus des publicités, de nombreux *weibo* produits par des utilisateurs ordinaires proposentt des recette de gastro-thérapie permettant de réduire les effets négatifs de la pollution de l'air, tels que 雪梨 (poire) et 木耳 (champignon noir) ; ou encore des méthodes pratiques pour renforcer le système immunitaire, tel que 补钙 (enrichir le calcium), 开窗通风 (ouvrir la fenêtre), 锻炼 (faire des exercices), 跑步 (faire du jogging).

6.1.7 *Résultats d'étude des sous-corpus institutionnel-médiatique et informel-médiatique*

Nous allons présenter conjointement les résultats d'études des sous-corpus institutionnel-médiatique et informel-médiatique. Et ce, pour deux raisons. Pre-

mièrement, malgré leur différence de nature (institutionnel vs informel), leur rôle commun médiatique fait que la production de l'information soit multimodale, notamment l'intégration de blog et forum, qui laisse la possibilité aux lecteurs de participer à la production de l'information dans les commentaires ; deuxièmement, le centre d'intérêt des deux sous-corpus tend à être multidimensionnel pour capter le plus grand nombre possible d'audiences. Ainsi, le système de circulation de l'information des deux sous-corpus est plutôt mix d'« exportation » et d'« importation ».

En raison de leur caractère mélangé, ces deux genres se trouvent dans une place intermédiaire, ils se servent de lien entre le genre institutionnel (purement institutionnel) et le profane (purement informel) en partageant les caractéristiques de ces deux derniers.

6.1.8 *Genre institutionnel-médiatique*

Au palier macrosémantique, le sous-corpus institutionnel-médiatique partage d'une part les caractères **identitaire**, **uniforme** et **homogène** du sous-corpus institutionnel en matière infratextuelle (identité politique présentée par les logotype en emblème, du fond de couleur en rouge, les grands portraits des dirigeants). Au niveau intratextuel, il partage les caractères **général**, **abstrait**, **injonctif** (noms de pays, titres de professions, établissements publics, expressions de l'ordre, verbe modal, adverbes emphatiques, termes général ou abstrait, etc.) de son équivalent institutionnel ; d'autre part, la **multimodalité** (cadre de commentaire, hyperlien vers d'autres sites), la **multidimensionnalité** (multiples rubriques thématiques) et l'**argumentation** (point d'interrogation, noms interrogatifs) du sous-corpus institutionnel-médiatique l'apparente au discours médiatique.

6.1.9 *Sémantiques des trois thèmes dans le sous-corpus institutionnel-médiatique*

Dans l'étude contrastive de la sémantique des trois thèmes aux niveaux micro- et mésosémantique, nous constatons que le sous-corpus institutionnel-médiatique traite les trois thèmes presque de la même manière que le sous-corpus institutionnel. Au niveau dialogique, on observe une utilisation importante de la 3ème personne du singulier ainsi qu'un emploi massif des entités nommées, telles que

les noms de personne, les noms d'état, les noms d'établissements, les toponymes pour mettre en avant le rôle de l'État. Au niveau dialectique, la saillance de la figure du parallélisme, des verbes modaux et des termes emphatiques permet de valoriser les actions méritoires de l'État quand il s'agit de la proposition des mesures préventives contre la pollution de l'air. Les deux sous-corpus se rapprochent l'un de l'autre en ce qui concerne la transformation de la pollution de l'air en phénomène naturel météorologique et la présentation globale des causes et des impacts du *wumai*.

6.1.10 *Genre informel-médiatique*

Quant au sous-corpus informel-médiatique, de par sa nature informelle, il rejoint le sous-corpus profane sur plusieurs aspects : au niveau infratextuel, malgré ces divergences fondamentales sur l'organisation du site, le sous-corpus informel-médiatique commence à proposer le dispositif de blog, qui permet aux lecteurs de laisser des commentaires avec la possibilité de les agrémenter d'émoticônes. Cet aspect lui confère un caractère multimodal et lui ajoute un rôle communicatif. Au niveau intratextuel, la multiplication de l'usage de signes techniques, des terminologies scientifiques, des mots d'emprunts, des émoticônes, de la ponctuation expressive (! et ?) permet de solidariser le sous-corpus informel-médiatique au sous-corpus profane par ses caractères **spécifique** et **expressif**.

6.1.11 *Sémantiques des trois thèmes dans le sous-corpus informel-médiatique*

De même, le traitement des trois thèmes du sous-corpus informel-médiatique ressemble celui du sous-corpus profane. Le sous-corpus informel-médiatique introduit un aspect scientifique avec l'usage de signes techniques, l'emploi de l'énumération, les terminologies techniques et les mots d'emprunt. En même temps, il assume son rôle **polémique** et de **problématisation** à travers la modalité interrogative manifestée par l'utilisation abondante des pronoms interrogatifs. Par ailleurs, il explore les trois thèmes dans un style de commentaire médiatique. Le sous-corpus informel-médiatique discute de ce sujet en exposant de manière exhaustive les types de maladies potentiellement causées par ce dernier. À partir du sous-corpus informel-médiatique, on commence à s'intéresser les causes du *wumai* à une ville ou une région et à analyser les spécificités de cette der-

nière en matière de pollution atmosphérique. De plus, il commence à prêter une attention particulière au groupe de femme et enfant. En même temps, les mesures préventives sont devenues plus concrètes (跑步 (jogging), 口罩 (masque), 空气净化器 (purificateur d'air)). Tous ces paramètres apparente le sous-corpus informel-médiatique au profane.

6.1.12 *Résumé*

En résumé, du genre institutionnel au genre profane, en passant par les genres institutionnel-médiatique et informel-médiatique, nous observons une transition d'un modèle de diffusion d'information « de haut en bas » à un modèle circulaire, de la généralité à la spécificité, de l'homogénéité à la diversité, de l'absence d'interaction à l'interaction, de l'abstraction à la concrétisation, de la stratégie à la pratique, du statisme au dynamisme. Et ce, quel que soit au niveau des caractéristiques du genre textuel des quatre sous-corpus ou au niveau de la sémantique des trois thèmes de ces derniers. Le graphe suivant montre la relation différentielle entre les quatre genres de sous-corpus.

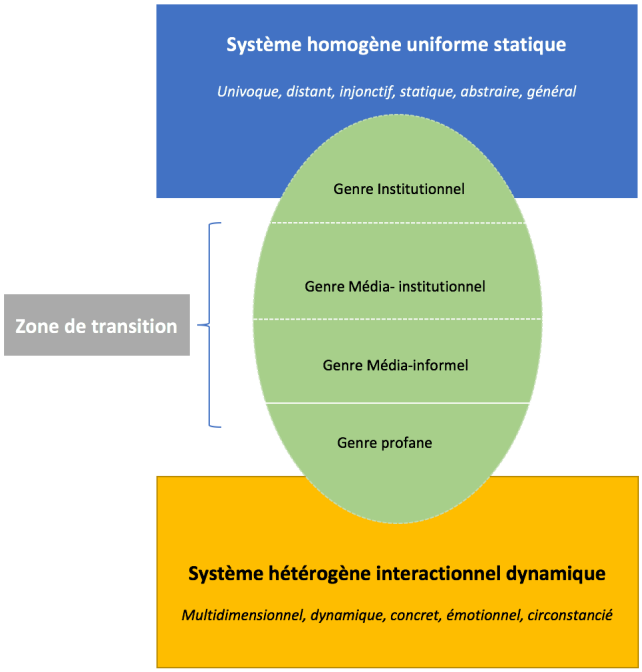


FIG. 6.1 – Relation entre les quatre genres de sous-corpus

Conclusion générale

Dans le cadre de notre recherche, nous avons articulé les études qualitatives de la SI avec celles quantitatives statistiques en adoptant la méthode textométrique. Nous objectif était d'explorer les quatre genres de sous-corpus au sujet du brouillard de pollution en Chine pour relever les thèmes principaux des quatre sous-corpus, ce qui permet ensuite d'interpréter la sémantique de chaque thème. Notre hypothèse était que les quatre genres de sous-corpus partagent des thèmes communs, mais ils se distinguent les uns des autres sur l'interprétation sémantique de chaque thème. Et ce, en tenant compte des caractéristiques du genre textuel auquel est associé chaque sous-corpus.

1 Contexte

De 2008 à aujourd'hui, l'inquiétude du public face au problème de brouillard de pollution n'a pas faibli ; au contraire, il devient un sujet très accrocheur et suscite un intérêt particulier des institutions gouvernementales ainsi que de la population chinoise. Ainsi, diverses voix provenant des médias traditionnels et des réseaux sociaux émergent pour discuter de ce problème écologique, qui touche des domaines très variés : de l'écologie à l'économie, de la politique à la vie quotidienne. Par conséquent, une grande quantité de données textuelles numériques, sous forme d'articles journalistiques et de *weibo*, sont disponibles et accessibles sur internet.

À l'aide des outils de *crawling* et des scripts R, nous avons récupéré au total 2556 articles journalistiques et 118,898 messages *weibo* avec cinq mots-clés : 雾霾 (brouillard de pollution), 霾 (smog), 空气污染 (pollution de l'air), 大气污染 (pollution atmosphérique), PM2.5 (Particular Matter Ø 2.5 μm) et PM10 (Particular Matter Ø 10 μm), qui partagent le sens de la /pollution de l'air/. Après une série de traitements de nettoyage, de segmentation, d'annotation, de normalisation, de balisage, de catégorisation et d'organisation, nous avons obtenu des textes

comptant 4,622,566 occurrences de forme et 117,000 formes (mots différents), qui couvrent la période de 2006 jusqu'à fin de 2018. Ces textes sont catégorisés, selon leur nature et caractéristiques discursives, en quatre genres textuels : institutionnel, institutionnel-médiatique, informel-médiatique et profane. Nous avons ajouté des annotations dans l'ensemble du corpus avec trois dictionnaires développés par nous-mêmes : « Dictionnaire des toponymes chinois », « Dictionnaire des maladies et symptômes » et « Dictionnaire de la dénomination de *wumai*». À partir de ces informations de métadonnées, nous avons ensuite indexé et balisé notre corpus à l'aide des logiciels techniques pour le rendre importable et analysable dans les outils textométriques.

2 Méthodologie

Notre méthodologie de travail est composée de la sémantique interprétative (SI) de Rastier et de la méthode textométrique. Dans l'intention d'étudier les thèmes principaux et d'interpréter la sémantique de ces derniers, nous avons fait usage opportuniste de la SI et exposé certaines propositions fondamentales qui sont adéquates à notre recherche :

- Principe général : « le global détermine le local, le sens est contextuel » ;
- Concepts fondamentaux de la théorie SI :
  - Trois paliers sémantiques : palier macrosémantique, palier mésosémantique et palier microsémantique ;
  - Genre textuel ;
  - Sème, Isotopie et Thème ;
  - Quatre composantes sémantiques : dialectique, dialogique, tactique et thématique.

De plus, nous avons adopté certains types de calculs de la textométrie ainsi que des outils qui peuvent nous aider à mieux utiliser des concepts fondamentaux de la SI : 1) le calcul des spécificités pour observer la distribution des termes à l'aide de la fonction « Distribution » d'Hyperbase ; 2) le calcul des « segments répétés » autour d'un mot-pivot grâce à la fonction « Segments Répétés » de Lexico5 dans le but de relever les expressions figées ; 3) l'AFC (Fonction « AFC » de TXM ou de Lexico5) pour étudier le rapport entre des termes et un genre de sous-corpus ; 4) le calcul des cooccurrences réalisé par la fonction « COOC » d'iTrameur et le « Nuage de mots » de Hyperbase pour relever des thèmes et étudier la sémantique

de ces derniers, et 5) la fonctionnalité « Concordance » de Lexico5 permettant de consulter les contextes d'apparition des termes ou mots-pivots choisis.

En nous basant sur le principe fondamental de la SI, nous avons effectué deux types d'analyses assistées par des outils textométriques. Le premier consiste à relever au palier macrosémantique les caractéristiques de chaque genre textuel à deux niveaux : 1) L'exploitation des caractéristiques infratextuelles, qui est effectuée à travers les spécificités de l'architecture des textes, stylistiques et de l'emplacement des éléments compositionnels configurés dans les sites web traditionnels et dans le réseau social, par exemple les logotypes, la couleur de fond, les images, les hyperliens externes, etc. ; 2) L'investigation des particularités intratextuelles, qui est réalisée à la lumière des études dialectiques et dialogiques des variables discriminantes lexicales, sémiotiques, modales, rhétoriques et syntaxiques de chaque genre textuel de sous-corpus. En ce qui concerne le deuxième type d'analyses, nous avons détecté, à travers les sèmes isotopants et leurs co-occurrents, à partir des palier micro- et mésosémantique, les thèmes principaux portés par chaque genre de sous-corpus : « les causes de la pollution de l'air », « les impacts de la pollution de l'air sur la santé » et « les mesures préventives contre la pollution de l'air ». Dans un second temps, les thèmes identifiés ont été analysés et interprétés à l'aide des composantes sémantiques et en fonction du genre textuel dans lequel ils s'inscrivent.

3 Contribution et originalité du travail

La contribution et l'originalité de cette thèse s'avèrent multiples :

1. La première contribution du travail présent réside dans la constitution du corpus. Allant de la récupération des données textuelles sur les différentes plateformes jusqu'à l'organisation du corpus, en passant par la segmentation, l'annotation, le nettoyage, le balisage, nous avons renseigné de manière explicite dans chaque étape les outils ainsi que les procédés. Notre objectif est de proposer une méthode reproductible pour créer et traiter un corpus numérique. Ce corpus écologique relatif au brouillard de pollution peut être ré-utilisé pour effectuer de diverses recherches. Et les multiples partitions (par genre/par an/par région) du corpus permettent de varier les types d'études sur ce dernier ;
2. Pour annoter notre corpus, nous avons créé trois dictionnaires : « Dic-

tionnaire toponyme des villes et régions de la Chine », « Dictionnaire des maladies et symptômes » ainsi que « Dictionnaire de la dénomination du *wumai*». Ces trois dictionnaires sont réutilisables ;

3. Afin d'explorer les caractéristiques du genre textuel au niveau intratextuel, nous avons repéré dans notre corpus des variables lexicales (pronom personnel, pronom interrogatif, néologisme, conjonction de subordination et de coordination, expression toute faite, verbe modal, verbe nominalisé, mot peu commun, mot d'emprunt fréquent), sémiotiques (signe mathématique, ponctuation d'exclamation, ponctuation d'interrogation, ponctuation inventée, ponctuation de numérotation, certaines émoticônes), modales (terme de négation, adverbe emphatique, temps descriptif, temps du passé, temps du futur, interjection, onomatopée) et rhétoriques (expressions de l'ordre en parallélisme), et les avons classés dans des listes séparées. Ces listes peuvent fournir aux chercheurs des pistes de réflexion dans leur étude du corpus ;
4. Au niveau de la méthodologie, en plus de la méthode pour créer et constituer un corpus, nous avons également mis en œuvre un dispositif expérimental de la SI au recours de la textométrie. Ce dispositif met à profit de certains principes théoriques relatifs à la SI (trois paliers sémantiques, quatre composantes sémantiques, genre textuel, isotopie, thème), des calculs statistiques (fréquence, spécificité, cooccurrence, segments répétés, AFC) et des outils textométriques (Lexico5, Trameur, Gromoteur, Hyperbase, TXM, JIEBA). De multiples exploitations outillées du corpus permettent de caractériser un genre textuel, d'identifier de manière semi-automatique des thèmes dans un corpus numérique, et d'interpréter la sémantique de chaque thème dans chaque genre du sous-corpus à travers ses caractéristiques du genre textuel. Cette méthode mélangée constitue également une proposition originale.

4 Perspectives

Les études que nous avons effectuées sur notre corpus relatif au *wumai* sont d'ordre synchronique. Toutefois, les thèmes ainsi que leurs interprétations sémantiques peuvent évoluer avec le temps et la localité dans un même genre de

sous-corpus [1], sans parler des divergences voire des distinctions spatio-temporelles parmi les textes du genre différent. Ainsi, nous envisageons, dans les recherches futures, de réaliser le même type d'analyses sémantiques que celui réalisé dans le présent travail mais de manière diachronique.

La méthodologie reproductible que nous avons proposée dans cette étude permet de faciliter notre travail à partir de l'ajout de nouveaux textes jusqu'à l'analyse sémantique. Ainsi, nous nous appuierons sur notre méthodologie actuelle pour entreprendre la future recherche.

Dans l'intention d'effectuer des études plus à jour, nous compléterons notre corpus en ajoutant des textes des genres institutionnels (datés de 2017 à fin 2020), institutionnels-médiatiques (datés de 2006 à 2010 et de 2017 à 2020) et informels-médiatiques (datés de 2006 à 2009 et de 2017 à 2020), et les *weibo* (datés de 2006 à 2013 et de 2019 à 2020). Cela étant fait, nous reproduirons toutes les procédures de traitements du corpus sur les nouveaux morceaux des textes et de *weibo* : nettoyage, segmentation, annotation, balisage et les fusionner dans chaque sous-corpus correspondant.

Comme nous avons déjà partitionné notre corpus par année et par région [2]. Nous profiterons de cette partition spatio-temporelle pour effectuer notre future étude, qui sera divisée en trois grandes parties :

- Analyses de l'évolution annuelle (de 2006 à 2019) des trois anciens thèmes dans chaque genre de sous-corpus : par exemple, l'évolution annuelle du thème « cause de la pollution de l'air » dans le sous-corpus institutionnel de 2006 à 2020 ;
- Repérage et interprétation sémantique de nouveaux thèmes dans l'ensemble du corpus par an et par genre de sous-corpus : par exemple, l'évolution annuelle de nouveau thème dans le sous-corpus profane de 2006 à 2020 ;
- Repérage et interprétation sémantique de nouveaux thèmes dans l'ensemble du corpus par région : par exemple, comparaison de nouveaux thèmes entre Hebei et Tianjin.

Les résultats des futures recherches enrichiront nos résultats actuels. Nous espérons que nos expériences et notre proposition de méthodologie pourra aider

1. D'après Rastier (2008), « l'interprétation doit se profiler au fil du temps et des transformations ».

2. Nous choisirons les villes les plus représentatives, qui sont en proie à la pollution de l'air, par exemple.

d'autres chercheurs dans leurs propres recherches de l'analyse sémantique des données textuelles numériques.

Annexe

1 Choix des sources de données de l'indice de qualité de l'air (AQI)

Deux organismes ont publié l'indice de la qualité de l'air (AQI) en Chine sur des plateformes dédiées :

1. Le Centre national de surveillance de l'environnement de la Chine (que nous noterons CH), rattaché au Ministère de l'Écologie et de l'Environnement de la Chine. Les relevés de cet organisme sont basés sur sept indices : AQI, PM2,5, PM10, SQ2, CO, NO2, O3 ; ils ont été effectués de décembre 2013 à septembre 2018. Toutes ces données sont accessibles sur le site de AQISTUDY (中国空气质量在线监测分析平台, plateforme en ligne de détection et d'analyse de l'IQA de Chine)[1] sous forme de fichiers CSV et de graphes visuels.
2. En 2008, l'Ambassade des États-Unis en Chine (noté US) a lancé un programme de surveillance de la qualité de l'air en Chine (http://stateair.net/. Consulté en février 2019.). Le taux de particule fine PM2,5 y a été publié toutes les heures jusqu'à juin 2017 pour plusieurs grandes métropoles chinoises : Beijing, Shanghai, Chengdu, Guangzhou, Shenyang. Les données sont stockées dans des fichiers au format CSV et accessibles sur le site officiel de l'Ambassade des États-Unis en Chine.

Nous allons effectuer une comparaison des relevés de ces deux organismes afin de sélectionner un de ces jeux de données comme source principale pour notre travail. Pour cela, nous avons choisi de confronter les relevés de l'indice de PM2,5 sur la période de 2013 à 2017. Dans la mesure où Beijing est l'une des villes les plus touchées par la pollution atmosphérique, nous avons choisi de comparer les relevés dans cette ville.

Nous avons d'abord calculé la valeur moyenne annuelle de l'indice PM2,5 à

1. https://www.aqistudy.cn/historydata/about.php. Consulté en novembre 2018.

Beijing de 2013 à 2017 pour chacune des deux sources de données (CH et US). Le graphe obtenu Figure 1 Comparaison de l'indice moyen de PM2,5 selon le AQISTUDY [2]) montre qu'à part une petite divergence en 2013 et en 2014 [3], les résultats sont globalement identiques. Nous avons décidé d'utiliser les données fournies par le site de AQISTUDY [4], car cette plateforme nous offre la possibilité de générer des graphes permettant d'observer et d'étudier la distribution spatiale (dans les 31 régions chinoises [5]) et l'évolution temporelle (de 2013 à la fin 2018) de la pollution de l'air dans l'ensemble de la Chine. Il est également possible de contraster les villes sélectionnées (en l'occurrence Beijing et Shanghai) sur une période prédéfinie (2014 - 2018).

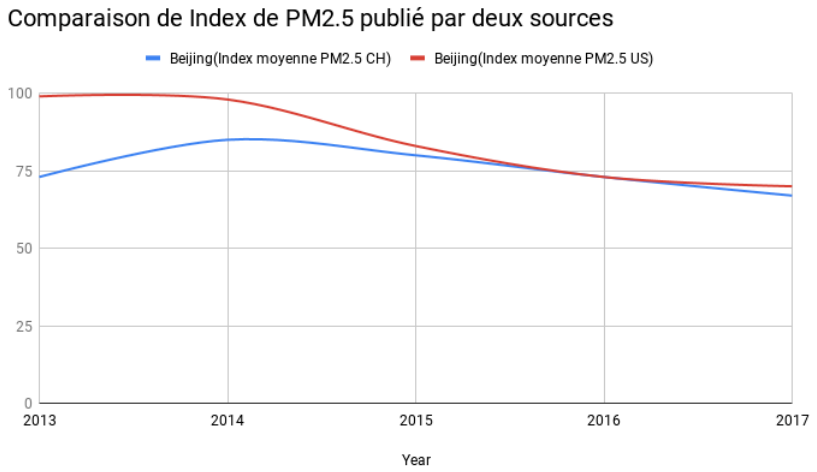


FIG. 1 – Comparaison de l'indice moyen de PM2,5 selon le AQISTUDY [6]

3. Les valeurs moyennes de PM2,5 en 2013 sont respectivement 73 pour CH et 99 pour US, 85 pour CH et 98 pour US en 2014

4. https://www.aqistudy.cn/historydata/about.php.

5. Les 31 régions sont composées de 22 provinces, 4 municipalités, 5 régions autonomes. Faute de données accessibles, la province de Taiwan et les 2 villes administrativement spéciales ne sont pas prises en compte. voir les Divisions Administratives et Disputes Territoriales de Chine dans l'Annexe.

6. Source : https://www.aqistudy.cn/historydata/about.php.

2 Tableau des Formes à demi et pleine chasse

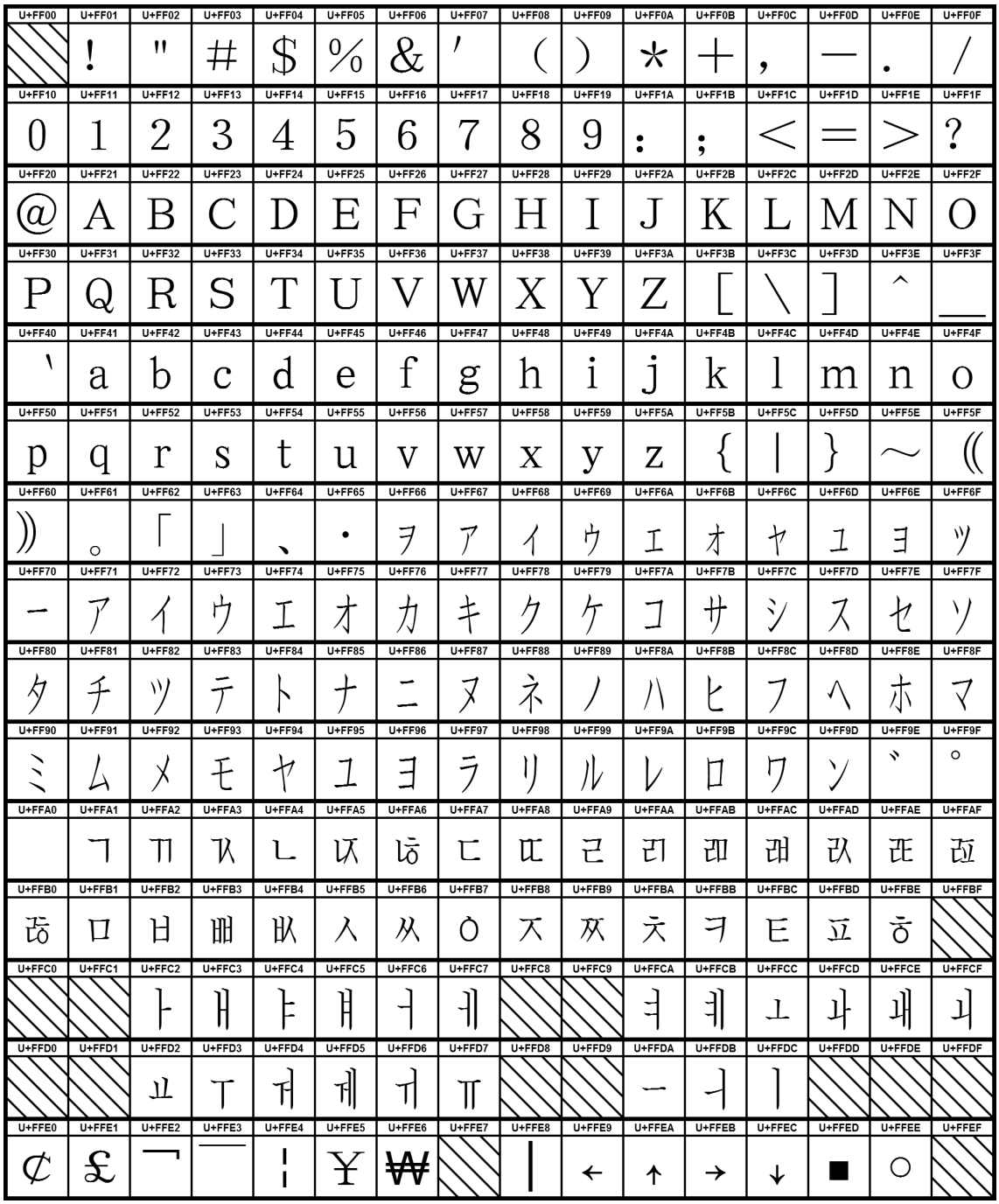


FIG. 2 – Tableau des Formes à demi et pleine chasse

3 Exemples de requête dans BCC

TAB. 1 – Exemples de requête dans BCC

检索式 **(Expression régulière)**	含义 **(Définition)**
高大的 n (grand n)	高大的 + 名词 (n) : grand+n
v 了 v (v particule v)	动词 (v) + 了一 + 动词 (v) : verbe(v)+particule 了 +verbe(v)
见 * 面 (voir*face)	见后面离合出现面 : voir est suivi de face
洗. 澡 (laver.douche)	洗后面隔一个字后接澡 : le verbe « laver est suivie d'un nom douche »
../v	二字动词 un verbe composé de deux mots
我./c 你	我和你之间用单字连词连接 : un seul connecteur relie « je » et « tu »
v[下去出来上来进去]	动词 (v) 后面连接下去、出来、上来、进去其中任何一个词。中括号内用空格隔离 : verbe est suivi des mots 下去、出来、上来、进去, ces mots sont séparés par un espace simple
把 *v 下去 w	把离合出现" 动词 (v)+ 下去" 并出现句尾 : mettre la combinaison « verbe+ 下去 » à la fin de la phrase
跑./v	以跑为首的双音节动词 : un verbe commencé par 跑
w 讨论 [n a v]	讨论后接名词或形容词或动词，并且讨论是句首 : 讨论 est suivi d'un nom ou d'un adjectif ou d'un verbe

4 Comparaison des signes typographiques en graphèmes pleine chasse et en demi chasse

Tab. 2 – Comparaison des graphèmes pleine chasse et demi chasse

Le début du tableau				
ASCII	Pleine chasse	Unicode	Demi chasse	Unicode
0x20	” ”	U+3000	” ”	U+0020
0x21	！	U+FF01	!	U+0021
0x22	"	U+FF02	”	U+0022
0x23	#	U+FF03	#	U+0023
0x24	$	U+FF04	$	U+0024
0x25	%	U+FF05	%	U+0025
0x26	&	U+FF06	&	U+0026
0x27	′	U+FF07	’	U+0027
0x28	（	U+FF08	(	U+0028
0x29	）	U+FF09	)	U+0029
0x2A	*	U+FF0A	*	U+002A
0x2B	+	U+FF0B	+	U+002B
0x2C	,	U+FF0C	,	U+002C
0x2D	—	U+FF0D	-	U+002D
0x2E	.	U+FF0E	.	U+002E
0x2F	/	U+FF0F	/	U+002F
0x30	0	U+FF10	0	U+0030
0x31	1	U+FF11	1	U+0031
0x32	2	U+FF12	2	U+0032
0x33	3	U+FF13	3	U+0033
0x34	4	U+FF14	4	U+0034
0x35	5	U+FF15	5	U+0035
0x36	6	U+FF16	6	U+0036
0x37	7	U+FF17	7	U+0037
0x38	8	U+FF18	8	U+0038
0x39	9	U+FF19	9	U+0039
0x3A	：	U+FF1A	:	U+003A

Continuation du tableau				
ASCII	Pleine chasse	Unicode	Demi chasse	Unicode
0x3B	;	U+FF1B	;	U+003B
0x3C	<	U+FF1C	<	U+003C
0x3D	=	U+FF1D	=	U+003D
0x3E	>	U+FF1E	>	U+003E
0x3F	?	U+FF1F	?	U+003F
0x40	@	U+FF20	@	U+0040
0x41	A	U+FF21	A	U+0041
0x42	B	U+FF22	B	U+0042
0x43	C	U+FF23	C	U+0043
0x44	D	U+FF24	D	U+0044
0x45	E	U+FF25	E	U+0045
0x46	F	U+FF26	F	U+0046
0x47	G	U+FF27	G	U+0047
0x48	H	U+FF28	H	U+0048
0x49	I	U+FF29	I	U+0049
0x4A	J	U+FF2A	J	U+004A
0x4B	K	U+FF2B	K	U+004B
0x4C	L	U+FF2C	L	U+004C
0x4D	M	U+FF2D	M	U+004D
0x4E	N	U+FF2E	N	U+004E
0x4F	O	U+FF2F	O	U+004F
0x50	P	U+FF30	P	U+0050
0x51	Q	U+FF31	Q	U+0051
0x52	R	U+FF32	R	U+0052
0x53	S	U+FF33	S	U+0053
0x54	T	U+FF34	T	U+0054
0x55	U	U+FF35	U	U+0055
0x56	V	U+FF36	V	U+0056
0x57	W	U+FF37	W	U+0057
0x58	X	U+FF38	X	U+0058
0x59	Y	U+FF39	Y	U+0059
0x5A	Z	U+FF3A	Z	U+005A

Continuation du tableau				
ASCII	Pleine chasse	Unicode	Demi chasse	Unicode
0x5B	［	U+FF3B	[	U+005B
0x5C	＼	U+FF3C	\	U+005C
0x5D	］	U+FF3D	]	U+005D
0x5E	＾	U+FF3E	^	U+005E
0x5F	＿	U+FF3F	_	U+005F
0x60	｀	U+FF40	‘	U+0060
0x61	ａ	U+FF41	a	U+0061
0x62	ｂ	U+FF42	b	U+0062
0x63	ｃ	U+FF43	c	U+0063
0x64	ｄ	U+FF44	d	U+0064
0x65	ｅ	U+FF45	e	U+0065
0x66	ｆ	U+FF46	f	U+0066
0x67	ｇ	U+FF47	g	U+0067
0x68	ｈ	U+FF48	h	U+0068
0x69	ｉ	U+FF49	i	U+0069
0x6A	ｊ	U+FF4A	j	U+006A
0x6B	ｋ	U+FF4B	k	U+006B
0x6C	ｌ	U+FF4C	l	U+006C
0x6D	ｍ	U+FF4D	m	U+006D
0x6E	ｎ	U+FF4E	n	U+006E
0x6F	ｏ	U+FF4F	o	U+006F
0x70	ｐ	U+FF50	p	U+0070
0x71	ｑ	U+FF51	q	U+0071
0x72	ｒ	U+FF52	r	U+0072
0x73	ｓ	U+FF53	s	U+0073
0x74	ｔ	U+FF54	t	U+0074
0x75	ｕ	U+FF55	u	U+0075
0x76	ｖ	U+FF56	v	U+0076
0x77	ｗ	U+FF57	w	U+0077
0x78	ｘ	U+FF58	x	U+0078
0x79	ｙ	U+FF59	y	U+0079
0x7A	ｚ	U+FF5A	z	U+007A

Continuation du tableau				
ASCII	Pleine chasse	Unicode	Demi chasse	Unicode
0x7B	｛	U+FF5B	{	U+007B
0x7C	｜	U+FF5C	|	U+007C
0x7D	｝	U+FF5D	}	U+007D
0x7E	～	U+FF5E	~	U+007E
La fin du tableau				

5 Fonctionnalités principales de l'Hyperbase

Tab. 3 – Fonctionnalités de l'Hyperbase

Fonctions documentaires	**Fonctions statistiques**
Retour au texte plein ou lemmatisé pour une lecture naturelle du corpus	Calcul des spécificités et graphes de distribution des unités linguistiques du corpus
Navigation hypertextuelle dans le corpus par mots-clefs	Indices de richesse lexicale et d'accroissement du vocabulaire
Recherche et tri des contextes et des concordances d'une unité	Traitement et représentation factoriels de matrices lexicales ou grammaticales complexes dans la lignée des travaux de Jean-Paul Benzécri
Index et dictionnaires des formes, des lemmes, des codes et des fréquences	Calcul de distances entre textes, classification et représentation arborées
	Extraction des phrases typiques et des segments répétés
	Calcul et représentations des cooccurrences et réseaux thématiques

6 Calcul des spécificités

Pour mesurer les spécificités dans un corpus, on doit faire appel à la méthode de Lafon (1980). Cette méthode permet de mesurer les variations de la fréquence dans un corpus découpé en parties et, en fonction d'un seuil choisi par l'analyste, il indique si la fréquence locale observée dans telle ou telle partie peut-être considérée comme normale ou non en fonction de la fréquence totale de l'ensemble du corpus. Dans ce dernier cas, P. Lafon propose de baptiser cette forme « spécifique » (de la partie considérée). Le calcul des spécificités est basé sur un modèle hypergéométrique, qui permet de comparer les quatre nombres :

- k : longueur total du corpus ;
- k_i :fréquence de la fome i dans le corpus entier ;
- k_j : longueur de la partie j ;
- k_{ij} : fréquence de la forme i dans la partie j ;

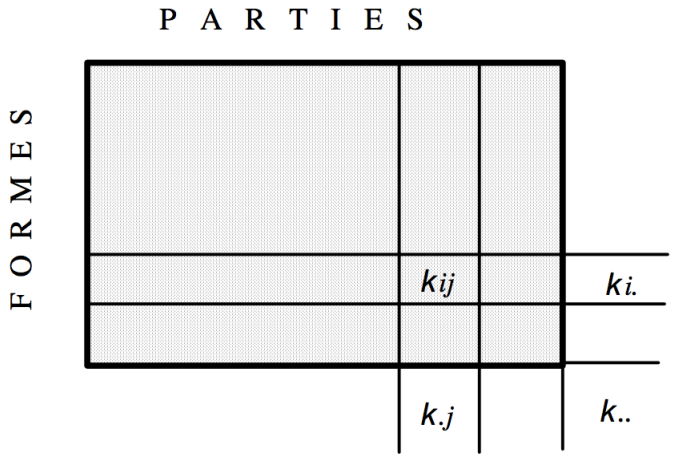


Fig. 3 – Paramètres du calcul des spécificités (Lebart et Salem, 1994)

> « Pour porter un jugement sur k_{ij}, le calcul de la probabilité tient compte des trois autres nombres (k, k_j et k_{ij}). Deux autres paramètres doivent être préalablement définis pour le calcul des spécificités : la valeur du seuil de spécificité et la fréquence minimale de la forme[7].

7. Dans l'outil textométrique *Lexico3*, le seuil de probabilité du calcul est par défaut fixé à 5% avec un fréquence des formes supérieures à 10.

Cette méthode de la mesure des variations de la fréquence permet de déceler trois types de formes :

- Si la probabilité de k_{ij} est supérieur au seuil fixé, la forme i est jugée sur-employée, c'est-à-dire que sa fréquence dans la partie j est « anormalement élevée », la forme i est la *spécificité positive*[8] dans la partie j ;
- Si la probabilité de k_{ij} est inférieure au seuil fixé, la forme i est jugée sous-employée, c'est-à-dire que sa fréquence dans la partie j est « anormalement faible », la forme i est la *spécificité négative* dans la partie j ;
- Si la probabilité de k_{ij} est égale au seuil calculé, la forme i dans la partie j est dite « banale ».

»(WU, 2016)

8. On note que l'indice de spécificité contient deux indicateurs : un signe + ou un signe - indiquant un sur-emploi ou un sous-emploi dans la partie considérée du corpus, suivi d'une valeur a indiquant la probabilité d'un écart de répartition de l'ordre de 10^{-x} supérieur ou égale à la valeur constatée. Dans l'outil textométrique *Lexico3*, si la valeur de spécificité est supérieure à 50 ou inférieure à -50, l'indice est noté «+*** » ou «-*** »

7 AFC (Analyse factorielle des correspondances)

L'analyse factorielle des correspondances (AFC) est une méthode statistique développée par Benzécri (1973, 1984, 1993). L'AFC fournit une typographie (selon l'indice de spécificité) basée sur les différentes parties des textes qui permet de contraster ou de rapprocher des différents regroupements des unités dans l'ensemble des données textuelles. « L'AFC s'appuie sur des décomptes d'occurrences des unités textuelles sous forme de tableaux à double entrée (appelés aussi tableaux croisés ou tableaux de contingence) où les lignes représentent, par exemple, les différentes formes graphiques (i) et les collonnes, les différentes parties (j) du corpus. Le croisement de la ligne et de la colonne du tableau...correspond au nombre de la forme i dans la partie j, ce nombre noté comme K (i, j)».

8 Dictionnaire des maladies et symptômes

8.1 *Maladies pulmonaires*

Tab. 4 – Maladies pulmonaires

Forme	Traduction	Fréquence
肺 Dns	poumons	981
肺癌 disease	cancer du poumon	692
清肺 Dv	nettoyer les poumons	381
肺泡 Dns	alvéoles	353
肺部 Dns	poumons	277
润肺 Dv	humecter les poumons	206
心肺 Dns	cardiopulmonaire	164
肺炎 disease	pneumonie	144
肺病 disease	maladie pulmonaire	143
肺脏 Dns	poumons	59
护肺 Dv	protéger les pulmons	40
肺气肿 disease	emphysème	38
养肺 Dv	protéger les pulmons	31
尘肺病 disease	pneumoconiose	26
伤肺 Dv	blesser les poumons	26
肺叶 n	lobe pulmonaire	24
肺结核 disease	tuberculose	11
肺心病 disease	cor pulmonale	8
肺炎球菌 Dns	pneumocoque	2

8.2 *Cancers*

Tab. 5 – Cancer

Forme	Traduction	Fréquence
肺癌 disease	cancer du poumon	692
癌症 disease	cancer	334
致癌 Dv	cancérigène	165
致癌物 Dns	cancérigène	161
抗癌 Dv	anti-cancer	23
肝癌 disease	cancer du foie	21
致癌物质 Dns	cancérigène	21
胃癌 disease	cancer gastrique	18
治癌 Dv	guérir le cancer	15
癌 disease	cancer	13
癌细胞 Dns	cellule cancéreuse	13
癌变 Dv	se cancériser	10
患癌 Dv	attraper un cancer	8
肾癌 disease	cancer du rein	6
膀胱癌 disease	cancer de la vessie	6
乳腺癌 disease	cancer du sein	6
癌症病人 Dns	patient cancéreux	4
胰腺癌 disease	cancer du pancréas	4
致癌性 Dns	cancérogénicité	6
食管癌 disease	cancer de l'oesophage	4
鼻咽癌 disease	carcinome nasopharyngé	4
宫颈癌 disease	cancer du col utérin	3
直肠癌 disease	cancer rectal	3
皮肤癌 disease	cancer de la peau	2
食道癌 disease	cancer de l'oesophage	2
口腔癌 disease	cancer buccal	1
喉癌 disease	cancer du larynx	1
贲门癌 disease	cancer du cardia	1

8.3 *Maladies respiratoires*

TAB. 6 – Maladies respiratoires

Forme	Traduction	Fréquence
呼吸系统 Dns	système respiratoire	993
呼吸道 Dns	voies respiratoires	918
哮喘病 disease	asthme	516
感冒 disease	rhume	510
支气管炎 disease	bronchite	167
呼吸科 Dns	département de respiratoire	110
气管炎 disease	trachéite	72
上呼吸道 Dns	appareil respiratoire supérieur	24
呼吸器官 Dns	organes respiratoires	11

8.4 *Symptômes*

Tab. 7 – Symptômes

Forme	Traduction	Fréquence
咳嗽 symptom	toux	344
炎症 symptom	inflammation	160
痛 symptom	douleur	145
疼 symptom	blessé	113
发炎 symptom	inflammation	82
咳 symptom	toux	73
头痛 symptom	maux de tête	58
呼吸困难 symptom	difficulté à respirer	47
头疼 symptom	maux de tête	39
上呼吸道感染 symptom	infection des voies respiratoires supérieures	39
皮肤过敏 symptom	irritation cutanée	36
疼痛 symptom	douleur	33
打喷嚏 symptom	éternuer	25
刺痛 symptom	piqûre	21
流鼻涕 symptom	nez qui coule	21
鼻塞 symptom	nez bouché	21
呛咳 symptom	toux	17
痛经 symptom	dysménorrhée	17
心绞痛 symptom	angine de poitrine	15
咽痛 symptom	maux de gorge	14
咳痰 symptom	toux grasse	13
嗓子疼 symptom	maux de gorge	13
咽炎 disease	pharyngite	12
阵痛 symptom	douleur expulsive	11
咳喘 symptomz	toux et essoufflement	10
肚子疼 symptom	douleurs à l'estomac	10
干咳 symptom	toux sèche	9
喉咙痛 symptom	maux de gorge	8
牙疼 symptom	mal aux dents	8
肿痛 symptom	douleur	8
流鼻血 symptom	saignement de nez	7
酸痛 symptom	douleur	6
胃痛 symptom	maux d'estomac	5
胸痛 symptom	douleur thoracique	5
腿疼 symptom	douleur aux jambes	5
呼吸衰竭 symptom	insuffisance respiratoire	5
咽干 symptom	gorge sèche	4
喉痛 symptom	maux de gorge	4
背疼 symptom	maux de dos	4
咽喉痛 symptom	maux de gorge	3
痛风 disease	goutte	3
腹痛 symptom	douleur abdominale	3
偏头疼 symptom	migraine	2
牙痛 symptom	mal aux dents	2

8.5 *Inflammation*

Tab. 8 – Inflammation

Forme	Traduction	Fréquence
鼻炎 disease	rhinite	210
肺炎 disease	pneumonie	142
咽炎 disease	pharyngite	46
消炎 Dv	anti-inflammatoire	39
结膜炎 disease	conjonctivite	28
咽喉炎 disease	inflammation de la gorge	23
鼻窦炎 disease	sinusite	8
关节炎 disease	arthrite	9
毛囊炎 disease	folliculite	6
角膜炎 disease	kératite	6
喉炎 disease	laryngite	4
胃炎 disease	gastrite	4
龟头炎 disease	Balanite	4
心肌炎 disease	myocardite	3
肝炎 disease	hépatite	3
肩周炎 disease	périarthrite	3
乙型肝炎 disease	hépatite B	2
睾丸炎 disease	orchite	2
视网膜炎 disease	rétinite	2
脊柱炎 disease	spondylarthrite	2
血管炎 disease	vascularite	2
结肠炎 disease	colite	1
肠胃炎 disease	gastroentérite	1
胆囊炎 disease	cholécystite	1
脑膜炎 disease	méningite	1
腱鞘炎 disease	ténosynovite	1
阴道炎 disease	vaginite	1
风湿性关节炎 disease	polyarthrite rhumatoïde	1
骨关节炎 disease	arthrose	1

8.6 *Maladies cardiovasculaires*

Tab. 9 – Maladies cardiovasculaires

Forme	Traduction	Fréquence
心脏病 disease	maladie cardiaque	628
心血管 Dns	cardiovasculaire	239
心脏 Dns	coeur	154
心脑血管 Dns	cardiovasculaire et cérébrovasculaire	119
血管 Dns	vaisseau sanguin	74
心脏 disease	coeur	71
心血管病 disease	maladie cardiovasculaire	46
心脑血管病 disease	maladie cardiovasculaire	17
毛细血管 Dns	capillaires	10
脑血管 Dns	cérébrovasculaire	10
血管炎 disease	vascularite	2

8.7 *Dermatose*

Tab. 10 – Dermatose

Forme	Traduction	Fréquence
皮肤 Dns	peau	1183
皮肤病 disease	dermatose	18
皮炎 disease	dermatite	15
皮肤科 n	dermatologie	8
皮肤癌 disease	cancer de la peau	2
皮肤感染 symptom	infections cutanées	1
皮肤性病 disease	maladie cutanée	1

8.8 *Ophtalmologie*

Tab. 11 – Ophtalmologie

Forme	Traduction	Fréquece
眼部 Dns	yeux	56
眼科 Dns	ophtalmologie	23
红眼病 disease	maladie des yeux rouges	12
眼药水 Dns	collyre	9
眼角膜 Dns	cornée transparente	8
眼病 disease	maladie oculaire	6
干眼症 disease	sécheresse oculaire	4
眼药 Dns	collyre	4

8.9 *Trouble mental*

Tab. 12 – Trouble mental

Forme	Traduction	Fréquence
心理 Dns	mentalité	119
心理健康 Dns	santé mentale	18
心理学 Dns	psychologie	8
心理障碍 Dns	trouble mental	8
心理准备 Dns	préparation psychologique	6
心理咨询 Dns	counselling	4
心理疾病 disease	maladie mentale	1

8.10 *ORl (oto-rhino-laryngologie)*

Tab. 13 – ORl (oto-rhino-laryngologie)

Forme	Traduction	Fréquence
鼻腔 Dns	cavité nasale	263
止咳 Dv	toux	91
喉咙 Dns	gorge	67
洗鼻 Dv	lavage nasal	61
咽喉 Dns	gorge	59
刺鼻 Dv	pique le nez	58
润喉 Dv	humecter la gorge	47
鼻黏膜 Dns	muqueuse nasale	33
咽 Dns	pharynx	31
耳鼻喉 Dns	ORl (oto-rhino-laryngologie)	25
擤鼻涕 Dv	moucher le nez	17
鼻部 Dns	nez	13
喉 Dns	gorge	11
呛鼻 Dv	irriter le nez	7
耳鼻喉科 Dns	ORl	6
鼻窦 Dns	sinus	5
鼻息肉 Dns	polypes nasaux	3

9 Dictionnaire de dénomination de *wumai*

9.1 *Brume (雾) étiquetés avec denowu*

Tab. 14 – Brume (雾)

Forme	Traduction	Fréquence
大雾 denowu	brouillard intense	2765
雾 denowu	brume	1433
雾天 denowu	temps de brouillard	974
烟雾 denowu	fumée	510
轻雾 denowu	brume légère	357
雾气 denowu	vapeur	310
浓雾 denowu	brouillard épais	307
水雾 denowu	embruns	231
白雾 denowu	buée blanche	137
云雾 denowu	nuage et brume	112
毒雾 denowu	brume empoisonnée	36
雾霭 denowu	brouillard	32
薄雾 denowu	brouillasse	31
黄雾 denowu	brouillard brunâtre	26
雨雾 denowu	pluie et brume	23

9.2 *Smog (霾) étiquetés avec denomai*

Tab. 15 – Smog (霾)

Forme	Traduction	Fréquence
雾霾 denomai	brouillard de pollution	30322
霾 denomai	smog	5857
灰霾 denomai	smog poussiéreux	803
阴霾 denomai	smog sombre	193
重霾 denomai	smog intense	88
尘霾 denomai	smog poussiéreux	48
烟霾 denomai	smog âcre	38
强霾 denomai	smog grave	28
毒霾 denomai	smog empoisonné	21
轻霾 denomai	smog léger	9
大霾 denomai	brouillard de pollution	8
京霾 denomai	smog de Beijing	5
湿霾 denomai	smog humide	3
黄霾 denomai	smog brunâtre	3
冬霾 denomai	brouillard de pollution hivernal	2
冀霾 denomai	smog de Hebei	2
干霾 denomai	smog sec	2
辽霾 denomai	smog de Liaoning	1
厚霾 denomai	smog épais	1
豫霾 denomai	smog de Henan	1
晋霾 denomai	smog de Shanxi	1
晨霾 denomai	smog du matin	1

9.3 *Pollution de l'air (空气污染) étiquetés avec denopollu*

Tab. 16 – Pollution de l'air (空气污染)

Forme	Traduction	Fréquence
大气污染 denopollu	pollution atmosphérique	4070
空气污染 denopollu	pollution de l'air	4509

9.4 *Particule ultrafine étiquetée avec denopm*

Tab. 17 – Particule ultrafine

Forme	Fréquence
PM2,5	35888
PM10	4990

10 Liste des différentes subdivisions de la structure territoriale de la Chine

Tab. 18 – Tableau de subdivision de la structure territoriale de la Chine[9]

Niveau	**Subdivision**	**POS tag topo-nyme**
Niveau provincial (省级行政区, 33)	Provinces (省, shěng, 23) , Municipalités (直辖市, zhíxiáshì, 4) Régions autonomes (自治区, zìzhìqū, 5) Régions administratives spéciales (特别行政区, tèbié xíngzhèngqū, 2)	province autoreg city
Niveau préfectoral (地级行政区, 333)	Préfectures (地区, dìqū, 17) réfectures autonomes (自治州, zìzhìzhōu, 30) Villes-préfectures (地级市, dìjíshì, 283) Ligues (盟, méng, 3)	city
Niveau districtal (县级行政区, 2 862)	Xian (县, xiàn, 1 464) Xian autonomes (自治县, zìzhìxiàn, 117) Villes-districts (县级市, xiànjíshì, 374) Districts (市辖区, shìxiáqū, 852) Bannières (旗, qí, 49) Bannières autonomes (自治旗, zìzhìqí, 3) Zones forestières (林区, línqū, 1) Districts spéciaux (特区, tèqū, 2)	city
Niveau cantonal (乡级行政区, 41 636)	Cantons (乡, xiāng, 14 677) Cantons ethniques (民族乡, mínzúxiāng, 1 092) Bourg (镇, zhèn, 19 522) Sous-districts (街道办事处, jiēdàobànshìchù, 6 152) Offices publics de district (区公所, qūgōngsuǒ, 11) Sumu (苏木, sūmù, 181) Sum ethnique (民族苏木, mínzúsūmù, 1)	city
Niveau communal (村和委员会)	Communautés résidentielles (社区居民委员会, jūmínwěiyuánhùi, 80 717) Villages (村民委员会, cūnmínwěiyuánhùi, 623 669) groupes de villages (村民小组, cūnmínxiǎozǔ) Villages administratifs (行政村, xíngzhèngcūn) Villages naturels (自然村, zìráncūn)	city

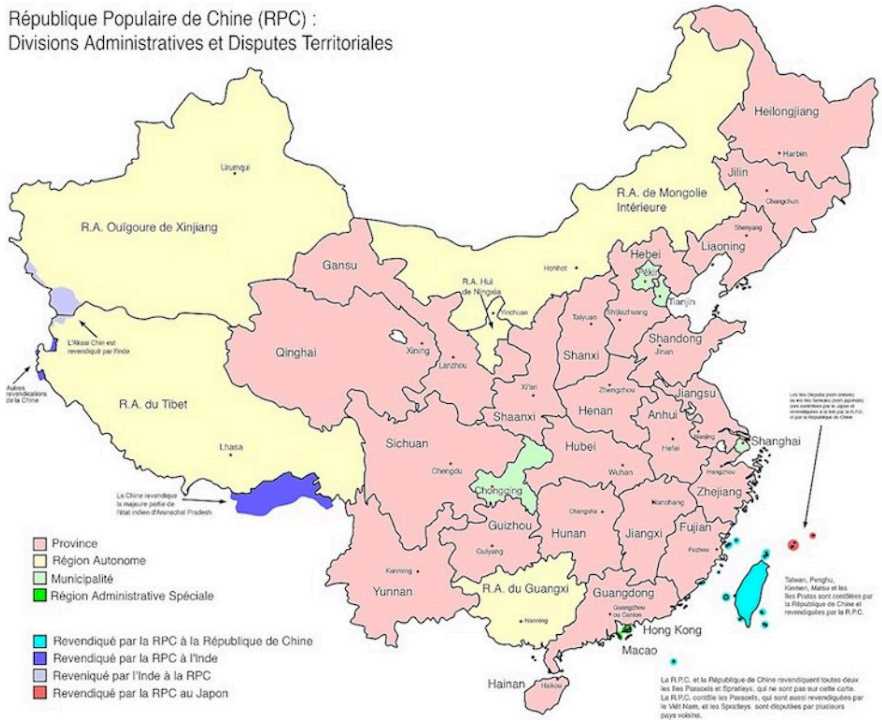


FIG. 4 – Divisions Administratives et Disputes Territoriales de Chine

9. https ://commons.wikimedia.org/wiki/File :RP_Chine_administrative2.jpg.

11 Liste des variables intratextuelles

11.1 *Conjonctions*

Tab. 19 – Conjonctions

Le début du tableau	
Terme	**Traduction**
	et
和 c	
但 c	mais
或 c	ou
而 c	et
不是 c	pas
还是 c	ou bien
如果 c	si
同时 c	en attendant
因为 c	parce que
以及 c	et
但是 c	mais
由于 c	en raison de
所以 c	donc
虽然 c	bien que
不过 c	mais
只有 c	Seulement
只是 c	juste
因此 c	ainsi
不仅 c	non seulement
及时 c	à temps
而且 c	et
或者 c	ou
如此 c	comme cela
此外 c	en plus
另外 c	par ailleurs
只要 c	tant que
而是 c	mais

Continuation du tableau	
Terme	**Traduction**
还要 c	encore
然后 c	alors
可是 c	toutefois
总是 c	toujours
既 c	puisque
是因为 c	parce que
若 c	si
并且 c	et
然而 c	cependant
即使 c	même si
从而 c	donc
不管 c	quoi que
以为 c	considérer
可怕 c	effrayant
尽管 c	malgré
而言 c	à propos de
不如 c	mieux vaut
于是 c	alors
反而 c	en revanche
无论 c	quoi qu'il en soit
并非 c	n'est pas
既然 c	puisque
此时 c	en ce moment
因 c	parce que
另一方面 c	d'autre part
否则 c	sinon
无论是 c	quel que soit
即便 c	même si
虽 c	bien que
以外 c	à part
或是 c	ou bien
只 c	seulement

Continuation du tableau	
Terme	**Traduction**
要是 c	si
何时 c	quand
与此同时 c	en même temps
不怕 c	ne pas avoir peur
所致 c	causée par
由此 c	donc
不但 c	non seulement
由 c	par
La fin du tableau	

11.2 *Termes de négation*

Tab. 20 – Termes de négation

Le début du tableau	
Termes de négation	**Traduction**
	Ne pas
不 d	
不再 d	Ne plus
不 d 需要 v pas besoin	
不 d 需 v Pas besoin	
不必 d	non nécessaire
不许 d	interdire
不 d 应当 v	Ne pas falloir
不 d 应该 v	Ne devrait pas être
没 d	Non
没有 v	Non
无 d	absence
莫 d	Non
非 d	Non
绝非 d	pas du tout
禁忌 v	Tabou
禁止 v	Interdire
防止 v	Prévenir
难以 d	Difficile
忽视 v	Négligence
放弃 v	Abandonner
拒绝 v	Rejeter
杜绝 v	Mettre fin à
La fin du tableau	

11.3 *Expressions proverbiales*

Tab. 21 – Expressions proverbiales

Le début du tableau	
Terme	**Traduction**
	par rapport à la même période
同期相比 i	
罪魁祸首 i	le coupable en chef
腾云驾雾 i	marcher rapidement sur les nuages
息息相关 i	relation très étroite
一目了然 i	s'en rendre compte d'un coup d'œil
刻不容缓 i	qui ne souffre pas point de retard
日常生活 i	vie quotidienne
突出重点 i	souligner l'importance capitale
贯彻落实 i	mettre en œuvre
安全隐患 i	risque de sécurité
全面实现 i	mettre pleinement en œuvre
日益严重 i	de plus en plus grave
若隐若现 i	être vaguement visible
迫在眉睫 i	imminent
防寒保暖 i	protection contre le froid et chaleur
前所未有 i	sans précédent
源源不断 i	en continu
当务之急 i	l'affaire la plus urgente
始于足下 i	un chemin de mille lieues commence toujours par un premier pas
众所周知 i	il est de notoriété publique que
风姿绰约 i	apparence de charme et de personnalité
晴空万里 i	ciel dégagé
为富不仁 i	cupide et cruel
晶莹剔透 i	claire comme du cristal
雾里看花 i	regarde les fleurs dans le brouillard
栩栩如生 i	être palpitant de vie
焕然一新 i	tout neuf
不知不觉 i	Inconsciemment

Continuation du tableau	
Terme	**Traduction**
一蹴而就 i	obtenir un résultat du jour au lendemain
秋高气爽 i	ciel dégagé et air vivifiant de l'automne
不可思议 i	Incroyable
独善其身 i	faire attention à sa propre morale sans tenir compte de l'idée des autres
触目惊心 i	choquant
齐心协力 i	unir les forces de tous
对症下药 i	administrer le médicament selon la maladie
积极参与 i	participer activement à
翩翩起舞 i	danser avec légèreté
立竿见影 i	obtenir de l'efficacité sur-le-champ
迫不及待 i	ne peux pas attendre
五花八门 i	de tout acabit
雪上加霜 i	une catastrophe après l'autre
来之不易 i	se procurer avec difficulté
势在必行 i	Impératif
习以为常 i	habituer très vite
一朝一夕 i	du jour au lendemain
因地制宜 i	selon les conditions locales
莫名其妙 i	sans rime ni raison
层出不穷 i	se reproduire sans fin
无孔不入 i	s'infiltrer partout
鱼龙混杂 i	mélange de bien et de mal
La fin du tableau	

11.4 *Rhétoriques Parallélisme*

- **Original** : 新水平、新境界、新举措、新发展、新突破、新成绩、新成效、新方法、新成果、新形势、新要求、新期待、新关系、新体制、新机制、新知识、新本领、新进展、新实践、新风貌、新事物、新高度；
- **Traduction** : Nouveaux niveaux, Nouveaux domaines, nouvelles initiatives, Nouveaux développements, nouvelles percées, nouvelles réalisations, nouvelles réalisations, nouvelles méthodes, nouvelles réalisations, nouvelles situations, nouvelles exigences, nouvelles attentes, nouvelles relations, Nouveaux systèmes, Nouveaux mécanismes, nouvelles connaissances, nouvelles compétences , de Nouveaux progrès, de nouvelles pratiques, de Nouveaux styles, de nouvelles choses, de Nouveaux sommets ;
- **Original** : 重要性、紧迫性、坚定性、民族性、全局性、前瞻性、战略性、积极性、创造性、复杂性、艰巨性、计划性、敏锐性、有效性；
- **Traduction** : Importance, urgence, fermeté, nationalité, globalité, prospective, stratégie, motivation, créativité, complexité, sensibilité, planification, perspicacité, efficacité ;
- **Original** : 法制化、规范化、制度化、程序化、集约化、正常化、智能化、优质化、常态化、科学化、年轻化、知识化、专业化、系统性、时效性；
- **Traduction** : Légalisation, normalisation, institutionnalisation, procéduralisation, intensification, normalisation, intelligence, qualification, normalisation, scientification, professionnalisation et systématisation ;
- **Original** : 是历史的必然、现实的选择、未来的方向。
- **Traduction** : (C'est)la nécessité de l'histoire, le choix de la réalité et la direction de l'avenir ;
- **Original** : 抓住机遇，应对挑战：量力而行，尽力而为 ;
- **Traduction** : Saisir l'opportunité et relever le défi : Faites ce que vous pouvez, faites de votre mieux ;

11.5 *Termes de temps descriptif*

Tab. 22 – Termes de temps descriptif

Terme	**Traduction**
腊八 t	fête de Laba
腊八节 t	fête de Laba
腊月 t	le douxième mois du calendrier lunaire
腊月初 t	fébut du douzième mois lunaire
节假日 t	jours fériés
节前 t	avant les vacances
节后 t	après les vacances
秋 t	automne
秋冬 t	automne et hiver
秋冬季 t	automne et hiver
秋后 t	après l'automne
秋夜 t	nuit d'automne
秋天 t	automne
秋季 t	automne
秋日 t	dans les jours d'automne
秋暮 t	fin de l'automne
秋月 t	mois d'automne
秋末 t	fin de l'automne
秋末冬 t	fin de l'automne hiver
正月 t	le premier mois selon calendrier lunaire
正月初 t	au début du premier mois selon le calendrier lunaire
正月初一 t	le premier jour du premier mois selon le calendrier lunaire
深冬 t	hiver profond
清秋 t	automne frais
此冬 t	cet hiver
今冬 t	cet hiver
冬初 t	début de l'hiver
冬夜 t	nuit d'hiver
冬天 t	hiver
冬季 t	hiver
冬日 t	hiver
冬至 t	premier jour de l'hiver

11.6 *Termes du temps du présent*

Tab. 23 – Termes du temps du présent

Terme	**Traduction**
现今 t	de nos jours
现在 t	maintenant
现如今 t	de nos jours
现年 t	année en cours
现日 t	jour actuel
现时 t	actuellement
现阶段 t	étape actuelle
現在 t	maintenant
現時 t	actuellement
當下 t	de nos jours
當代 t	contemporaine
當前 t	à présent
當季 t	saison actuelle
而今 t	maintenant
目前 t	actuellement
今 t	aujourd'hui
今个 t	aujourd'hui
今儿 t	aujourd'hui
今儿个 t	aujourd'hui
今夕 t	ce soir
今天 t	aujourd'hui
今年 t	cette année
当下 t	pour le moment
当今 t	à l'instant
当代 t	contemporaine
当前 t	pour le moment

11.7 *Termes du temps de l'imparfait*

Tab. 24 – Termes du temps de l'imparfait

Terme	**Traduction**
近一年来 t	depuis l'année dernière
近三年 t	depuis ces trois dernières années
近些天 t	depuis ces derniers jours
近些年 t	depuis ces dernières années
近些年来 t	depuis ces dernières années
近代 t	dans le temps moderne
近午 t	vers midi
近半年 t	depuis ces six derniers mois
近年来 t	depuis ces dernières années
近日 t	depuis ces derniers jours
近期 t	depuis ces derniers temps
近来 t	récemment
十年间 t	pendant dix ans

11.8 *Termes du temps du passé*

Tab. 25 – Terme du temps du passé

Le début du tableau	
Terme	**Traduction**
	le passé
过去 t	
翌日 t	jour suivant
西元 t	ère chrétienne
西周 t	dynastie des Zhou occidentaux
解放前 t	avant la libération
解放后 t	après la libération
秦代 t	dynastie Qin
秦时 t	à l'époque des Qin
當年 t	dans la même année
當時 t	à l'époque
汉代 t	dynastie des Han
汉朝 t	dynastie des Han
汉末 t	à la fin de dynastie des Han
清代 t	dynastie des Qing
清朝 t	dynastie des Qing
清末 t	à la fin de la dynastie des Qing
上一年 t	l'année dernière
上个世纪 t	siècle dernier
上个星期 t	la semaine dernière
上个月 t	le mois dernier
上代 t	génération précédente
上元节 t	festival de Shangyuan
上冻前 t	avant la congélation
上前 t	à l'avant
上午 t	matin
上半 t	moitié supérieure
上半夜 t	au milieu de la nuit

Continuation du tableau	
Terme	**Traduction**
上半年 t	première moitié de l'année
上古时代 t	préhistoire
上周 t	la semaine dernière
上年 t	l'année dernière
上旬 t	la première moitié du mois
上月 t	le mois dernier
上月底 t	à la fin du mois dernier
上次 t	dernière fois
不久前 t	récemment
世代 t	génération
从前 t	auparavant
刚才 t	toute à l'heure
前一天 t	le jour d'avant
前一晚 t	la nuit précédente
前三 t	trois premiers
前三天 t	les trois premiers jours
前三季 t	les trois premières saisons
前世 t	vie passée
前五 t	top cinq
前些天 t	jours précédents
前些年 t	années précédentes
前半夜 t	la première moitié de la nuit
前夜 t	la nuit précédente
前天 t	avant-hier
前年 t	l'année précédente
半年前 t	six mois avant
去年初 t	au début de l'année dernière
去年底 t	à la fin de l'année dernière
去年末 t	à la fin de l'année dernière
后来 t	plus tard
当天 t	le jour même
当时 t	à l'époque

Continuation du tableau	
Terme	**Traduction**
往年 t	les années précédentes
往日 t	le passé
往昔 t	le passé
早前 t	auparavant
昨一晚 t	la nuit dernière
昨儿 t	hier
昨儿个 t	hier
昨夜 t	la nuit dernière
昨天 t	hier
昨天上午 t	hier matin
昨天下午 t	hier après-midi
昨天中午 t	hier midi
昨天夜里 t	la nuit dernière
昨天晚上 t	la nuit dernière
昨日 t	hier
昨晚 t	la nuit dernière
昨晨 t	hier matin
La fin du tableau	

11.9 *Termes du temps du futur proche*

Tab. 26 – Termes du temps du futur proche

Terme	**Traduction**
转眼间 t	en un instant
转瞬 t	en un instant
转瞬间 t	en un instant
瞬息 t	instantané
瞬时 t	instantané
瞬间 t	instantané
不日 t	sous peu
今冬明春 t	cet hiver et le printemps prochain
今夏 t	cet été
今夜 t	cette nuit
今年冬天 t	cet hiver
今年年底 t	à la fin de cette année
今年底 t	à la fin de cette année
今年春节 t	la fête du printemps de cette année
明后天 t	demain ou après demain
明夜 t	nuit lumineuse
明天 t	demain
明年 t	l'année prochaine
明日 t	demain
明早 t	demain matin
明晚 t	demain soir
本周 t	cette semaine
本季 t	cette saison
本底 t	fond
本日 t	aujourd'hui
本月 t	ce mois

11.10 *Termes du temps du futur*

Tab. 27 – Termes du temps du futur

Terme	**Traduction**
未來 t	avenir
下一代 t	prochaine génération
下一阶段 t	prochaine étape
下个星期 t	la semaine prochaine
下个月 t	le mois prochain
下代 t	prochaine génération
下個月 t	le mois prochain
下午 t	après-midi
下半夜 t	après minuit
下半年 t	deuxième moitié de l'année
下半月 t	la seconde moitié du mois
下周 t	la semaine prochaine
下旬 t	la deuxième moitié du mois
下时 t	quand
下星期 t	la semaine prochaine
下晚 t	nuit prochaine
下月 t	le mois prochain
下期 t	prochaine période
下次 t	la prochaine fois
下段 t	partie inférieure
今后 t	à l'avenir
从今 t	à partir d'aujourd'hui
往后 t	plusard
本世纪 t	ce siècle
本世纪内 t	dans ce siècle
本世纪末 t	à la fin du siècle

11.11 *Termes emphatiques*

Tab. 28 – Verbes nominalisés

Le début du tableau	
Terme	**Traduction**
	Augmenter
增加	
加强	Renforcer
加大	Augmenter
加重	Aggraver
更加	Encore plus
加快	Accélérez
加剧	Intensifier
加	Ajouter
增强	Améliorer
强化	Renforcer
强度	Intensité
强大	Puissant
强调	Accent
强力	Fort
坚强	Fort
强制	Forcer
明显增强	Significativement amélioré
强制性	Obligatoire
强效	Puissant
强化措施	Mesures de renforcement
强行	De force
特别强调	Accent particulier
不断加强	Continuez à vous renforcer
强制措施	Mesures coercitives
继续加强	Continuer à renforcer
强烈要求	Demande fortement
强烈建议	Hautement recommandé
强制执行	Appliquer

Continuation du tableau	
Terme	**Traduction**
必须	Doit
必要	Nécessaire
必备	Doit
必然	Inévitable
务必	Doit
必不可少	essentiel
必将	Will
势必	Lié à
势在必行	Impératif
必需	Requis
必定	Doit
不必要	Pas nécessaire
有法必依	Il doit y avoir une loi
必备品	Incontournable
必要性	Nécessité
必要条件	Condition nécessaire
执法必严	L'application doit être stricte
违者必究	Les délinquants feront l'objet d'une enquête
必然选择	Choix inévitable
必要措施	Mesures nécessaires
违法必究	Illégal doit faire l'objet d'une enquête
必然规律	Inévitable
必定会	Sera certainement
必然结果	Résultat inévitable
必然联系	Nécessairement connecté
需要	Besoin
需	Besoin
无需	Pas besoin
需注意	Besoin de faire attention
急需	Besoin urgent

Continuation du tableau	
Terme	**Traduction**
迫切需要	Besoin urgent
亟需	Très nécessaire
急切需要	Besoin désespéré
急需解决	Besoin urgent de résoudre
能源需求	Demande d'énergie
La fin du tableau	

11.12 *Verbes nominalisés*

TAB. 29 – Verbes nominalisés

Le début du tableau	
Terme	**Traduction**
	pollution
污染 vn	
影响 vn	influence
预警 vn	avertissement
发展 vn	développement
工作 vn	travail
应急 vn	urgence
生活 vn	vie
研究 vn	recherche
监测 vn	surveille
预计 vn	prévision
启动 vn	démarrage
预报 vn	prévision
活动 vn	activité
建设 vn	construction
生产 vn	production
分析 vn	analyse
升级 vn	mise à jour
行动 vn	action
管理 vn	gestion
旅行 vn	voyage
运动 vn	sport
调查 vn	enquête
消费 vn	consommation
调整 vn	réajustement
变化 vn	changement
免费 vn	offre
预测 vn	prévision
设计 vn	désign

Continuation du tableau	
Terme	**Traduction**
检查 vn	contrôle
改革 vn	réforme
投资 vn	investissement
监管 vn	réglementation
检测 vn	détectiion
监督 vn	supervision
宣传 vn	propagande
La fin du tableau	

11.13 *Adverbes*

Tab. 30 – Adverbes

Le début du tableau	
Adverbe	**Traduction**
	également
也 d	
都 d	tous
就 d	juste
还 d	aussi
又 d	encore une fois
更 d	plus
很 d	très
最 d	le plus
持续 vd	continuum
已经 d	déjà
再 d	de nouveau
已 d	déjà
较 d	plus
太 d	trop
一定 d	certainement
才 d	seulement
却 d	mais
仍 d	encore
特别 d	spécialement
只 d	seulement
比较 d	relativement
必须 d	obligatoirement
一直 d	toujours
非常 d	extrêmement
甚至 d	voire
直接 ad	directement
高速 d	rapidement
其实 d	effectivement

Continuation du tableau	
Adverbe	**Traduction**
不少 d	beaucoup
逐渐 d	progressivement
越来越 d	de plus en plus
尤其 d	surtout
约 d	environ
自然 d	naturellement
越 d	plus
长期 d	à long terme
真正 d	réellement
终于 d	finalement
不断 d	constamment
均 d	en moyenne
仅 d	seulement
正常 d	normalement
进一步 d	d'avantage
La fin du tableau	

11.14 *Autres variables*

Tab. 31 – Autres variables

Le début du tableau							
Exclamatifs	Onomatopées	Mots peu communs	Chiffres	Signes typologiques	Toponymes	Signes mathématiques	Mots étrangers
乌 y	嗒 o	灞 x	000m	x	中国 ns	x	一 x
一样 y	咯 o	幫 x	001m	x	城市 ns	x	〆 x
一般 y	喵 o	狈 x	003m	x	青岛市 ns	x	丶 x
不 y	啪啪 o	餅 x	005m	x	华北 ns	x	丷 x
不成 y	嘟嘟 o	亳 x	009m	x	中度 ns	x	々 x
不过 y	咔 o	卟 x	00m	x	山东省 ns	±x	x
么 y	嘤 o	艸 x	015m	x	北京市 ns	＞x	x
乌乎 y	咚 o	誠 x	01m	x	鲁西北 ns	x	x
乌戏 y	咩 o	亍 x	0222m	x	美国 ns	＜x	x
乖乖 y	呜 o	處 x	02m	x	上海市 ns	%x	x
也 y	叽 叽 喳 喳 o	詞 x	039m	x	鲁南 ns	%x	厶 x
也好 y	哒 o	哃 x	03m	¨x	周边地区 ns		丅 x
也罢 y	嘎 o	哒 x	04m	ˋx	江南 ns		く x
也那 y	咕咚 o	噠 x	05m	ˊx	日本 ns		巜 x
了 y	喔 o	汏 x	06m	˚x	东北 ns		勹 x
了得 y	唧唧 o	擋 x	07m	x	城六区 ns		x
于乎 y	啧啧 o	昀 x	08m	—x	台风 ns		x

Continuation du tableau							
Exclamatifs	Onomatopées	Mots peu communs	Chiffres	Signes typologiques	Toponymes	Signes mathématiques	Mots étrangers
于嗟 y	羋 o	铞 x	0921m	−x	伦敦 ns		x
于戏 y	叮当 o	調 x	09m	x	上市 ns		x
于皇 y	叭 o	仃 x	0m	−x	华南 ns		x
于铔 y	哈 o	娲 x	0 m	x	蓝牙 ns		x
价 y	啧 o	讀 x	10000m	Øx	英国 ns		x
伙颐 y	嘎嘎 o	隊 x	1000m	∞x	鲁西南 ns		x
似地 y	滴答 o	莪 x	1008m	⋯x	美丽 ns		x
便了 y	嘣 o	洩 x	100m	x	太阳 ns		ん x
偌 y	铴锵 o	餓 x	1010m	x	成都市 ns		ン x
兮 y	咣 o	発 x	1012m	? x	东北地区 ns		x
再说 y	嗷嗷 o	髮 x	101m	; x	海洋 ns		を x
则 y	噗 o	啡 x	1020m	: x	西南地区 ns		わ x
别忙 y	淅沥 o	汐 x	102m	/ x	京城 ns		ワ x
叱 y	叮叮当当 o	x	103m	。x	西北 ns		ロ x
叹辞 y	叽叽 o	噶 x	104m	. x	中至 ns		れ x
吁 y	呼噜 o	尬 x	105m	、x	北风 ns		レ x
吁嗟 y	咔嚓 o	苷 x	106m	，x	东南 ns		る x
吓 y	咪咪 o	顧 x	107m	*x	长三角 ns		ル x
吗 y	哗啦啦 o	館 x	108m		四川盆地 ns		り x
否 y	呜 o	還 x	1097m		天津市 ns		リ x
吧 y	嘻 o	咁 x	109m		美 ns		ら x

Continuation du tableau							
Exclamatifs	Onomatopées	Mots peu communs	Chiffres	Signes typologiques	Toponymes	Signes mathématiques	Mots étrangers
吭唷 y	扑通 o	楁 x	10m		江汉 ns		ラ x
呀 ey	咕噜 o	啈 x	1100m	x	舟山 ns		よ x
呀呀呼 y	咪咕 o	鴻 x	1103m	" x	华北地区 ns		ょ x
呃 y	哗啦 o	逅 x	110m	* x	中 ns		ヨ x
呃嚽 y	嗖嗖 o	鎚 x	1112m	) x	韩国 ns		ゆ x
呐 ey	嗵 o	話 x	111m	x	河北省 ns		ュ x
呔 y	噼里啪啦 o	懷 x	112m	（x	台 ns		や x
呕 y	锵锵 o	徨 x	11392m	x			ヤ x
呗 y	阿咚 o	虺 x	113m	& x	洛杉矶 ns		ゃ x
呜 y	乒乒乓乓 o	匯 x	114m	x	欧洲 ns		も x
呜呼 y	乓 o	灬 x	115m	～ x			モ x
呢 y	叽 o	圾 x	116m	} x	西北地区 ns		め x
呣 y	呱 o	乩 x	1177m	|x			メ x
呦 ey	呱唧 o	擠 x	117m	丨 x	绍兴 ns		む x
呵 y	呼啦啦 o	蛱 x	1180m	{x	山区 ns		ム x
呵呵 y	咕噜噜 o	撿 x	118m	x	松江区 ns		み x
呶 y	咚咚 o	鹹 x	119m	‘x	南昌 ns		ミ x
呸 y	咪嗦 o	見 x	11m	x̂	美白 ns		ま x
咄 y	咭 o	艦 x	1200m	] x	北京地区 ns		マ x
咄咄 y	咯咯 o	腳 x	120m	\x	杭州市 ns		ぽ x

Continuation du tableau							
Exclamatifs	Onomatopées	Mots peu communs	Chiffres	Signes typologiques	Toponymes	Signes mathématiques	Mots étrangers
咍 y	哇唔 o	尐 x	121m	＼ x	广州市 ns		ポ x
咍吁 y	哇嗷 o	婕 x	122m	［x	法国 ns		ボ x
咦 y	哼哧 o	槿 x	12345m	x	巴黎 ns		ほ x
咧 y	嘟 o	勁 x	12369m	x	湘潭 ns		へ x
咨 y	滴哈 o	啾 x	123m	” x	京东 ns		ぷ x
咨虐 y	叭叭 o	據 x	124m	“x	上路 ns		ブ x
咯 y	叮叮 o	狷 x	125m	’x	浦东 ns		フ x
咳 y	叮咚 o	嘅 x	126m	‘x	都市 ns		ビ x
哇 y	吧唧 o	課 x	127m	‖x	黄山 ns		ひ x
哈 y	吭嗤 o	蔻 x	128m	゜x	印度 ns		ヒ x
哈呀 ey	吱 o	徠 x	129m	ˇx	抚顺 ns		ぱ x
哈哈 y	吱吱 o	瀾 x	12m	x	西南 ns		パ x
哉 y	唄唄 o	啷 x	1300m	´x			ば x
哎 y	呱呱 o	仂 x	130m	x	德国 ns		バ x
哎也 y	呼哧 o	淚 x	131m	x	江苏 ns		は x
哎呀 ey	呼 哧 呼哧 o	莉 x	132m	〆x	无锡市 ns		ハ x
哎哈 y	咔哒 o	麗 x	133m		青春 ns		の x
哎哟 ey	咕咕 o	蓮 x	134m	〗x	长江 ns		ノ x
哑 y	咕 噜 咕噜 o	啉 x	135m	〖x			ね x
哟 ey	咪哒 o	劉 x	136m	〉x	南京市 ns		ぬ x
哦 y	咪滴 o	珑 x	1379m	€x	北方地区 ns		ヌ x
哦呵 y	咯吱 o	胧 x	137m	〔x			に x
哦嗬 y	咯啷 o	龍 x	138m	】x	深度 ns		ニ x

Continuation du tableau							
Exclamatifs	Onomatopées	Mots peu communs	Chiffres	Signes typologiques	Toponymes	Signes mathématiques	Mots étrangers
哩 y	哇啦 o	萎 x	1398m	【x	钟南山 ns		な x
哪 y	哈哒 o	嘍 x	139m	』x	东西 ns		ど x
哼 y	哈哒啦 o	栌 x	13m	『x	宁波市 ns		ド x
哼唷 y	哼哈 o	猡 x	1400m	」x	日照市 ns		と x
唉 y	唧 o	嗎 x	140m	「x	京 ns		ト x
唏 y	啪 o	馬 x	141m	￥x	包邮 ns		で x
唗 y	啪啦 o	買 x	14260m	¥x	朝阳 ns		デ x
唦 y	喃 o	壳 x	142m	》x			て x
唷喂 y	喃喃 o	熳 x	1436m	《x	南 ns		テ x
唻 y	喳 o	氓 x	143m	〉x	北 ns		ヅ x
啊 ey	喳喳 o	硿 x	144m	〈x	天津地区 ns		つ x
啊呀 ey	嗖 o	貓 x	145m	‾‾ x	巢湖市 ns		ツ x
啊哈 y	嗡 o	眊 x	146m	x	山西省 ns		っ x
啊哟 ey	嗷 o	麼 x	147m	x			ッ x
啊唷 y	嘎吱 o	檬 x	1480m	x	哈市 ns		ヂ x
啦 y	嘟噜 o	蘼 x	148m	»x	桐庐县 ns		ち x
啧 y	噜啦噜 o	缈 x	1499m				チ x
啧啧 y	噜噜 o	緲 x	149m		山林 ns		だ x
La fin du tableau							

12 Concordance de 心理健康 (santé psychologique) du sous-corpus GOV

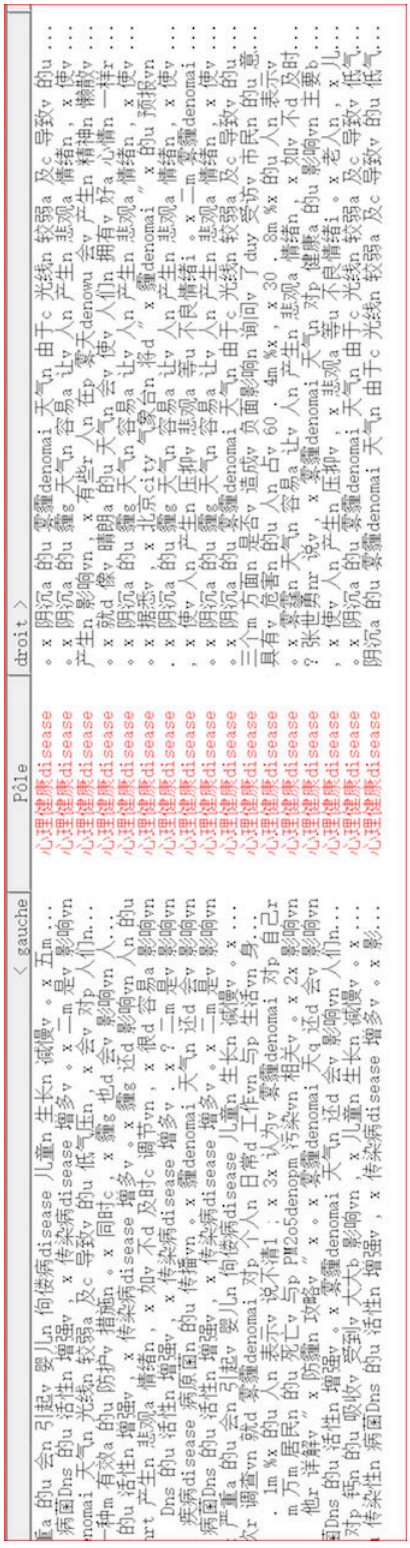


FIG. 5 – Concordance de 心理健康 (santé psychologique) du sous-corpus GOV

13 Liste de COOC de 雾霾 du GOV

Tab. 32 – COOC de 雾霾 (brouillard de pollution) du GOV

Le début du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
	日 m	Jour	4412	4288	**	374
雾霾 denomai						
雾霾 denomai	天气 n	Météo	3550	3457	49	487
雾霾 denomai	中东部 nt	Centre Est	955	944	24	161
雾霾 denomai	北京 city	La ville de Beijing	2419	2325	21	340
雾霾 denomai	地区 n	Zone	2260	2174	20	314
雾霾 denomai	预警 vn	Avertissement	1355	1320	20	260
雾霾 denomai	冷空气 n	Air froid	682	676	20	164
雾霾 denomai	南部 city	Ville du sud	1026	1005	19	110
雾霾 denomai	中央气象台 nt	Station météo centrale	595	591	19	124
雾霾 denomai	华北 ns	Chine du Nord	973	953	18	152
雾霾 denomai	08m	8	691	680	16	74
雾霾 denomai	北部 f	North	1208	1173	16	108
雾霾 denomai	气象局 n	Bureau météorologique	678	668	16	122
雾霾 denomai	大雾 denowu	Brouillard épais	815	798	15	146
雾霾 denomai	笼罩 v	Enveloppé	358	358	15	127
雾霾 denomai	新华社 nt	Agence de presse Xinhua	483	479	15	155
雾霾 denomai	部分 n	Partie	1169	1131	14	202
雾霾 denomai	黄淮 nr	Huang Huai	783	766	14	121
雾霾 denomai	中南部 nt	Centre sud	534	528	14	95
雾霾 denomai	东部 f	F orientale	1296	1249	13	130
雾霾 denomai	上午 t	Matin	277	277	12	114

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	夜间 t	Nuit	351	349	12	111
雾霾 denomai	白天 t	Jour	272	272	12	97
雾霾 denomai	持续 vd	Vd continu	803	778	11	243
雾霾 denomai	东北部 f	Nord-est	469	461	11	76
雾霾 denomai	月 m	Mois	2428	2296	10	393
雾霾 denomai	局地 n	Local	626	610	10	105
雾霾 denomai	发布 v	Publier	897	865	10	244
雾霾 denomai	西北部 f	Northwestern	373	368	10	72
雾霾 denomai	四川盆地 ns	Bassin du Sichuan	259	258	10	68
雾霾 denomai	重度 n	Sévère	505	492	9	156
雾霾 denomai	雨雪 n	Pluie et neige	360	355	9	75
雾霾 denomai	影响 vn	Influence	1781	1689	9	367
雾霾 denomai	预计 vn	Vn attendu	464	454	9	135
雾霾 denomai	不足 a	Moins qu'un	422	412	8	140
雾霾 denomai	西部 f	Ouest	771	743	8	110
雾霾 denomai	大部 n	La plupart	755	726	8	95
雾霾 denomai	黄色 n	Jaune	429	419	8	142
雾霾 denomai	小雨 nr	Légère pluie	191	191	8	53
雾霾 denomai	江南 ns	Jiangnan	537	521	8	93
雾霾 denomai	应对 v	Faire face	411	401	8	176
雾霾 denomai	东南部 f	Sud-est	439	428	8	81
雾霾 denomai	能见度 n	Visibilité	1019	974	8	217
雾霾 denomai	东北地区 ns	Nord-est	306	301	8	70
雾霾 denomai	23m	23m	233	230	7	89
雾霾 denomai	疾病 disease	Maladie	364	355	7	130
雾霾 denomai	日电 j	Électricité japonaise	243	240	7	128
雾霾 denomai	江淮 j	Jianghuai	412	400	7	83
雾霾 denomai	中度 n	Modéré	249	246	7	93

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	记者 n	Journaliste	1015	969	7	296
雾霾 denomai	预案 n	Plan	233	230	7	81
雾霾 denomai	污染 vn	Pollution	2258	2122	7	372
雾霾 denomai	预报 vn	Prévisions	536	517	7	154
雾霾 denomai	雨夹雪 l	Pluie et neige mêlées	268	264	7	55
雾霾 denomai	小到中雪 l	Petite à moyenne neige	230	227	7	52

La fin du tableau

14 Liste de COOC de 雾霾 du PEOPLE

Tab. 33 – COOC de 雾霾 (brouillard de pollution) du PEOPLE

Le début du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
	天气 n	Météo	2031	1958	**	684
雾霾 denomai						
雾霾 denomai	治理 v	Gouvernance	1604	1491	43	572
雾霾 denomai	空气 n	Air	1427	1295	25	521
雾霾 denomai	影响 vn	Influence	1004	924	23	459
雾霾 denomai	污染物 n	Matière polluante	707	660	21	299
雾霾 denomai	排放 v	Émission	886	813	20	308
雾霾 denomai	预警 vn	Avertissement	882	804	18	270
雾霾 denomai	天 q	Jour	613	571	18	356
雾霾 denomai	PM2o5denopm	PM2o5denopm	863	789	18	347
雾霾 denomai	疾病 disease	Maladie	365	349	17	162
雾霾 denomai	大气 n	L'atmosphère	535	500	17	240
雾霾 denomai	大气污染 denopollu	polluollutione l'air	493	462	16	227
雾霾 denomai	北京 city	Ville de Beijing	1637	1443	16	549
雾霾 denomai	污染 vn	Pollution	2547	2214	16	681
雾霾 denomai	严重 a	grave	907	818	15	464
雾霾 denomai	发布 v	Publier	544	505	15	290
雾霾 denomai	口罩 n	Masque	989	884	14	228
雾霾 denomai	应急 vn	Urgence	526	487	14	167
雾霾 denomai	环保部 n	Ministèree larotectione l'environnement	263	253	14	149
雾霾 denomai	健康 a	Santé	619	565	13	320
雾霾 denomai	扩散 v	Diffusion	256	245	13	142
雾霾 denomai	持续 vd	continu	501	462	13	277
雾霾 denomai	成因 n	Genèse	194	190	13	118

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	报道 v	Rapport	267	256	13	103
雾霾 denomai	范围 n	Gamme	272	260	13	188
雾霾 denomai	出现 v	Vpparaît	718	648	12	376
雾霾 denomai	形成 v	Formulaire	421	390	12	223
雾霾 denomai	重度 n	Sévère	402	374	12	198
雾霾 denomai	消散 v	Dissipation	147	145	12	89
雾霾 denomai	笼罩 v	Enveloppé	209	202	12	140
雾霾 denomai	颗粒物 n	Matièrearticulaire	352	330	12	162
雾霾 denomai	研究 vn	Recherche	425	394	12	193
雾霾 denomai	2014m	2014m	390	362	12	267
雾霾 denomai	中心 n	Centre	248	237	12	153
雾霾 denomai	对 p	Pour	1986	1714	11	711
雾霾 denomai	应对 v	Faire face	307	288	11	187
雾霾 denomai	部分 n	Partie	312	292	11	180
雾霾 denomai	治 v	Règle	320	299	11	168
雾霾 denomai	霾 denomai	Smog	516	469	11	217
雾霾 denomai	较 d	Plus	267	252	11	167
雾霾 denomai	区域 n	Zone	299	280	11	151
雾霾 denomai	室内 s	Intérieur	227	217	11	100
雾霾 denomai	浓度 n	Concentration	376	348	11	173
雾霾 denomai	危害 n	Danger	315	294	10	191
雾霾 denomai	防治 v	Prévention	281	264	10	141
雾霾 denomai	主要 b	Main	628	564	10	341
雾霾 denomai	铀 n	Uranium	132	130	10	11
雾霾 denomai	呼吸系统 Dns	Système respiratoirens	157	153	10	94
雾霾 denomai	空气质量 n	Qualitée l'air	838	742	10	316
雾霾 denomai	空气污染 denopollu	Polluollutione l'air	605	544	10	261
雾霾 denomai	相关 v	Concerner	409	376	10	248
雾霾 denomai	显示 v	Montrer	253	238	10	178

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	2013m	2013m	262	246	10	171
雾霾 denomai	监测 vn	Surveillance	284	265	10	146
雾霾 denomai	公众 n	Public	256	241	10	154
雾霾 denomai	指出 v	Indiquer	184	177	10	130
雾霾 denomai	来源 n	Source	202	193	10	110
雾霾 denomai	记者 n	Reporter	616	550	9	322
雾霾 denomai	伦敦 ns	Londress	210	198	9	58
雾霾 denomai	措施 n	Mesure	605	540	9	273
雾霾 denomai	2015m	2015m	346	318	9	237
雾霾 denomai	防 v	Anti-	146	141	9	76
雾霾 denomai	中央气象台 nt	Stationétéo centrale	125	122	9	76
雾霾 denomai	过程 n	Processus	347	318	9	204
雾霾 denomai	燃煤 n	Charbon	292	271	9	141
雾霾 denomai	呼吸道 Dns	Voix respiratoires	165	158	9	100
雾霾 denomai	天 n	Jour	290	269	9	133
雾霾 denomai	系统 n	Système	220	208	9	107
雾霾 denomai	市民 n	Citoyen	326	301	9	169
雾霾 denomai	认为 v	Pense	354	324	9	209
雾霾 denomai	洛杉矶 ns	Los angeless	111	109	9	26
雾霾 denomai	巴黎 ns	Pariss	107	106	9	18
雾霾 denomai	黄淮 nr	Huang Huair	99	98	9	37
雾霾 denomai	由于 c	En raisone	277	255	8	189
雾霾 denomai	进行 v	entreprendre	495	444	8	254
雾霾 denomai	导致 v	Conduire à	418	378	8	246
雾霾 denomai	遭遇 n	Rencontre	168	159	8	121
雾霾 denomai	启动 vn	Démarrer	306	281	8	155
雾霾 denomai	引起 v	Cause	147	141	8	114
雾霾 denomai	新闻 n	Nouvelles	260	241	8	177
雾霾 denomai	减少 v	Réduire	542	483	8	268
雾霾 denomai	分析 vn	Analyse	187	176	8	125

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	机制 n	Mécanisme	135	130	8	86
雾霾 denomai	预防 v	Prévention	122	118	8	78
雾霾 denomai	含量 n	Contenu	106	104	8	43
雾霾 denomai	10m	10m	833	730	8	395
雾霾 denomai	气溶胶 n	Aérosol	71	71	8	37
雾霾 denomai	南部 city	Ville du sud	108	106	8	49
雾霾 denomai	人体 n	Corps humain	246	229	8	133
雾霾 denomai	能见度 n	Visibilité	309	284	8	149
雾霾 denomai	控制 v	Contrôle	274	251	7	132
雾霾 denomai	平均 a	Moyenne	128	122	7	83
雾霾 denomai	人们 n	Personnes	375	339	7	231
雾霾 denomai	活动 vn	Activité	219	202	7	132
雾霾 denomai	我国 r	Chine	484	432	7	223
雾霾 denomai	煤炭 n	Charbon	162	153	7	78
雾霾 denomai	重 a	Lourd	644	568	7	247
雾霾 denomai	中国气象局 nt	Administration météorologique de Chine	62	62	7	53
雾霾 denomai	中东部 nt	Centre Est	119	114	7	73
雾霾 denomai	室外 s	Extérieur	129	124	7	58
雾霾 denomai	产生 n	Générer	382	345	7	219
雾霾 denomai	逐渐 d	Progressivement	150	142	7	108
雾霾 denomai	增多 v	Augmenter	96	94	7	70
雾霾 denomai	加重 v	Aggravée	139	133	7	100
雾霾 denomai	颗粒 n	Particule	201	188	7	109
雾霾 denomai	大雾 denowu	Brouillard épais	202	187	7	101
雾霾 denomai	内蒙古 autoreg	Autoregeongolie intérieure	84	83	7	33
雾霾 denomai	锻炼 v	Exercice physique	110	106	7	40
雾霾 denomai	天然 b	Naturel	54	54	6	23

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	较大 a	Plus grand	76	74	6	67
雾霾 denomai	直接 ad	directement	209	191	6	146
雾霾 denomai	患者 n	Patient	201	186	6	86
雾霾 denomai	程度 n	Degré	243	222	6	164
雾霾 denomai	医院 n	Hôpital	171	158	6	90
雾霾 denomai	近日 t	Récemment	97	93	6	85
雾霾 denomai	研究所 n	Institut	69	68	6	61
雾霾 denomai	首席 n	Chef	50	50	6	48
雾霾 denomai	时间 n	Temps	337	301	6	225
雾霾 denomai	及时 c	À temps	132	125	6	85
雾霾 denomai	吸入 v	Aspirer	63	62	6	44
雾霾 denomai	北京市 city	Ville de Beijing	452	400	6	183
雾霾 denomai	中南部 nt	Centre sud	89	86	6	47
雾霾 denomai	中科院 nt	Académie chinoise des Sciences	80	78	6	40
雾霾 denomai	媒体 n	Média	208	191	6	111
雾霾 denomai	氮氧化物 n	Oxyde d'azote	69	68	6	53
雾霾 denomai	净化器 n	Purificateur	229	208	6	80
雾霾 denomai	气象条件 n	Conditions météorologiques	130	122	6	84
雾霾 denomai	专家 n	Expert	460	407	6	268
雾霾 denomai	气候变化 n	Changement climatique	54	54	6	32
雾霾 denomai	机动车 n	Véhicule automobile	385	342	6	162
雾霾 denomai	水平 n	Niveau	207	189	6	107
雾霾 denomai	二氧化硫 nz	Dioxyde de soufre	74	72	6	55
雾霾 denomai	石家庄 city	Villee Shijiazhuang	64	63	6	38
雾霾 denomai	北京地区 ns	Régioneeijings	51	51	6	36

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	科普 nrt	Nrtopulaire	51	51	6	16
雾霾 denomai	上午 t	Matin	117	111	6	85
雾霾 denomai	河 北 province	Province Hebei	385	343	6	150
雾霾 denomai	运动 vn	Mouvement	122	115	6	63
雾霾 denomai	辐射 v	Rayonnement	65	64	6	22
雾霾 denomai	秸秆 n	Paille	243	222	6	90
雾霾 denomai	减弱 v	Affaiblir	115	110	6	73
雾霾 denomai	煤 n	Charbon	175	163	6	57
雾霾 denomai	强度 n	Intensité	68	67	6	52
雾霾 denomai	夜间 t	Nuit	151	142	6	92
雾霾 denomai	造成 v	Cause	534	471	6	311
雾霾 denomai	体检 v	Examen médical	57	57	6	8
雾霾 denomai	肆虐 v	Sans ménagement	46	46	5	41
雾霾 denomai	日数 n	Nombreeours	47	47	5	31
雾霾 denomai	督查 vn	Superviseur	47	47	5	29
雾霾 denomai	最新 d	Dernier	56	55	5	48
雾霾 denomai	部长 n	Ministre	71	69	5	45
雾霾 denomai	立方米 q	Mètre cube	190	173	5	88
雾霾 denomai	对此 d	Pour cela	108	102	5	97
雾霾 denomai	补贴 n	Subvention	167	153	5	58
雾霾 denomai	碳 n	Charbon	119	111	5	57
雾霾 denomai	沈阳 city	Ville de Shenyang	85	81	5	32
雾霾 denomai	成分 n	Ingrédient	106	100	5	61
雾霾 denomai	预计 vn	prévision	206	187	5	113
雾霾 denomai	南京 city	Ville de Nanjing	225	204	5	95
雾霾 denomai	放射性 n	Radioactif	45	45	5	9
雾霾 denomai	灰霾 denomai	Smog gris	172	158	5	76
雾霾 denomai	焚烧 v	brûler	151	139	5	72

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	停课 v	Suspension des cours	207	188	5	71
雾霾 denomai	防霾 n	anti-smog	200	182	5	86
雾霾 denomai	户外活动 n	Activités en plein air	98	93	5	60
雾霾 denomai	辽宁 province	Provinceu Liaoning	122	114	5	38
雾霾 denomai	细菌 n	Bactéries	40	40	5	29
雾霾 denomai	主任 b	Directeur	107	101	5	85
雾霾 denomai	开窗 n	Ouvrir les fenêtres	51	50	5	31
雾霾 denomai	频繁 a	Fréquent	70	68	5	62
雾霾 denomai	降水 n	Précipitations	87	83	5	38
雾霾 denomai	呼吸科 Dns	Service respiratoire	49	49	5	28
雾霾 denomai	人群 n	la foule	130	121	5	68
雾霾 denomai	晨练 n	Excercice du matin	40	40	5	28
雾霾 denomai	气象部门 n	Département de la météorologie	83	80	5	58
雾霾 denomai	节能 v	Économiser des énergies	111	105	5	66
雾霾 denomai	原因 n	la raison	386	342	5	251
雾霾 denomai	生殖 vn	Reproduction	42	42	5	18
雾霾 denomai	咳嗽 symptom	La toux	63	61	5	45

La fin du tableau

15 Liste de COOC de 雾霾 du SOHU

Tab. 34 – COOC de 雾霾 (brouillard de pollution) du SOHU

Le début du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
	天气 n	Météo	2220	2177	**	708
雾霾 denomai						
雾霾 denomai	口罩 n	Masque	1246	1183	33	286
雾霾 denomai	严重 a	Un sérieux	1064	1018	33	517
雾霾 denomai	天 q	Jour	753	732	32	396
雾霾 denomai	防 v	Anti-v	435	433	30	143
雾霾 denomai	预警 vn	Avertissement	1275	1204	30	282
雾霾 denomai	日 m	Jour	1888	1748	28	537
雾霾 denomai	治理 v	Gouvernance	871	835	28	273
雾霾 denomai	重度 n	Sévère	456	447	23	213
雾霾 denomai	排名 v	Rang	323	322	23	175
雾霾 denomai	污染 vn	Pollution	2618	2354	18	610
雾霾 denomai	城市 n	Ville	1260	1160	17	428
雾霾 denomai	空气 n	Air	1440	1319	17	520
雾霾 denomai	记者 n	Reporter	492	470	16	271
雾霾 denomai	关于 p	À propos de	492	470	16	345
雾霾 denomai	霾 denomai	Smog	496	471	14	185
雾霾 denomai	成因 n	Genèse	216	214	14	71
雾霾 denomai	地区 n	Zone	555	523	14	261
雾霾 denomai	危害 n	Danger	460	436	13	278
雾霾 denomai	北京 city	La ville de Beijing	2163	1936	13	635
雾霾 denomai	指数 n	Index	372	356	13	159
雾霾 denomai	华北 ns	Chine du Nord	347	334	13	124
雾霾 denomai	京津冀 n	Beijing-Tianjin-N	324	311	12	135
雾霾 denomai	原因 n	Raison	696	644	12	389
雾霾 denomai	净化器 n	Purificateur	342	327	12	110
雾霾 denomai	空气质量 n	Qualité de l'air	707	656	12	289
雾霾 denomai	人们 n	Personnes	279	268	11	182

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	笼罩 v	Envelopper	193	190	11	122
雾霾 denomai	2015m	2015m	462	432	10	308
雾霾 denomai	中度 n	Modéré	162	160	10	86
雾霾 denomai	出现 v	Apparaît	655	602	9	321
雾霾 denomai	形成 v	Former	487	452	9	290
雾霾 denomai	空气污染 denopollu	Pollution de l'air	519	480	9	265
雾霾 denomai	儿童 n	Enfant	174	170	9	85
雾霾 denomai	室内 s	Intérieur	229	220	9	123
雾霾 denomai	食物 n	Nourriture	198	192	9	80
雾霾 denomai	黄淮 nr	Huang Huai	154	151	9	52
雾霾 denomai	东北 ns	Nord-est	125	123	8	70
雾霾 denomai	应急 vn	Urgence	425	393	8	115
雾霾 denomai	影响 vn	Influence	839	760	8	399
雾霾 denomai	沈阳 city	La ville de Shenyang	140	137	8	73
雾霾 denomai	停课 v	suspension des cours	239	228	8	83
雾霾 denomai	清肺 Dv	Humecter les poumons	159	155	8	76
雾霾 denomai	应对 v	Faire face	199	189	7	123
雾霾 denomai	疾病 disease	Maladie	348	322	7	155
雾霾 denomai	红色 n	Rouge	527	483	7	148
雾霾 denomai	最新 d	Dernier	113	111	7	99
雾霾 denomai	12m	12m	629	572	7	310
雾霾 denomai	扩散 v	Diffusion	244	231	7	134
雾霾 denomai	面对 v	Face à	109	107	7	93
雾霾 denomai	持续 vd	continu	407	375	7	242
雾霾 denomai	宝宝 nr	Bébé	180	172	7	47
雾霾 denomai	月 m	Mois	1367	1211	7	542

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	图 n	Figure	134	130	7	87
雾霾 denomai	康熙 nrfg	Kangxi nrfg	87	87	7	15
雾霾 denomai	图片 n	Photo	168	161	7	121
雾霾 denomai	橙色 n	Orange	225	213	7	103
雾霾 denomai	气象台 n	Station météo	119	116	7	75
雾霾 denomai	能见度 n	Visibilité	309	289	7	154
雾霾 denomai	黄色 n	Jaune	191	182	7	90
雾霾 denomai	最好 a	Meilleur	198	189	7	134
雾霾 denomai	张召忠 nr	Zhang Zhaozhong	66	66	6	8
雾霾 denomai	戴 v	Porter	286	265	6	161
雾霾 denomai	遭遇 n	Rencontre	126	122	6	95
雾霾 denomai	北部 f	North	79	78	6	35
雾霾 denomai	预报 vn	Prévisions	130	125	6	71
雾霾 denomai	北方 f	North	97	95	6	65
雾霾 denomai	措施 n	Mesure	445	406	6	225
雾霾 denomai	哮喘 disease	Asthme	66	66	6	49
雾霾 denomai	中南部 nt	Centre sud	63	63	6	33
雾霾 denomai	乾隆 nr	Qianlong	69	69	6	12
雾霾 denomai	冷空气 n	Air froid	207	194	6	103
雾霾 denomai	吃 v	Manger	391	357	6	169
雾霾 denomai	17m	17m	112	109	6	76
雾霾 denomai	消散 v	Dissiper	127	122	6	89
雾霾 denomai	南京 city	La ville de Nanjing	92	90	6	55
雾霾 denomai	引发 v	Provoquer	118	115	6	95
雾霾 denomai	预案 n	Plan	218	204	6	63
雾霾 denomai	大雾 denowu	Brouillard épais	202	190	6	90
雾霾 denomai	清洗 v	Nettoyage	90	89	6	54
雾霾 denomai	减弱 v	Affaiblir	91	89	6	61

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	局地 n	Local	101	99	6	51
雾霾 denomai	江苏 province	Province du Jiangsu	122	118	6	48
雾霾 denomai	锻炼 v	Exercice	122	118	6	50
雾霾 denomai	容易 a	Facile	198	186	6	131
雾霾 denomai	京城 ns	Capitale	96	93	5	50
雾霾 denomai	伦敦 ns	Londres	126	120	5	54
雾霾 denomai	山西 province	Province du Shanxi	92	89	5	40
雾霾 denomai	鼻腔 Dns	Nasale	81	79	5	40
雾霾 denomai	柴静 nr	Chai Jing (journaliste chinoise)	243	224	5	58
雾霾 denomai	拍摄 v	Tourner	51	51	5	36
雾霾 denomai	人群 n	Foule	114	109	5	60
雾霾 denomai	新闻 n	Nouvelles	416	376	5	302
雾霾 denomai	中央气象台 nt	Station météo centrale	67	66	5	46
雾霾 denomai	督查 vn	Superviseur	54	54	5	23
雾霾 denomai	穹顶 n	Dôme	88	86	5	39
雾霾 denomai	尽量 d	Essayer	99	95	5	74
雾霾 denomai	沈阳市 city	La ville de Shenyang	65	64	5	19
雾霾 denomai	呼吸道 Dns	respiratoires	265	243	5	123
雾霾 denomai	维生素 nr	Vitamine	123	117	5	31
雾霾 denomai	其中 r	Où	276	254	5	189
雾霾 denomai	上午 t	Matin	107	103	5	70
雾霾 denomai	预计 vn	Estimation	181	169	5	94
雾霾 denomai	发布 v	Publier	478	433	5	231
雾霾 denomai	外出 v	Sortir	165	156	5	81
雾霾 denomai	冬季 t	Hiver	169	159	5	105

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	注意事项 n	les consignes à observer	48	48	5	40
雾霾 denomai	AQIeng	Air Quality Index	109	105	5	43
雾霾 denomai	统计表 n	Tableau statistique	57	57	5	57
雾霾 denomai	辽宁 province	Province du Liaoning	118	113	5	51
雾霾 denomai	元凶 n	Coupable	88	85	5	66
雾霾 denomai	百合 n	Lys	67	66	5	28

La fin du tableau

16 Liste de COOC de 雾霾 du WEIBO

Tab. 35 – COOC de 雾霾 (brouillard de pollution) du WEIBO

Le début du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
	戴 v	Porter	510	186	**	123
雾霾 denomai						
雾霾 denomai	口罩 n	Masque	1788	689	**	300
雾霾 denomai	天气 n	Météo	5864	1246	**	923
雾霾 denomai	防 v	Anti	572	245	**	157
雾霾 denomai	笼罩 v	Envelopper	217	106	**	105
雾霾 denomai	舆情 n	Sentiment public	61	57	**	56
雾霾 denomai	天 q	Jour	1686	592	**	477
雾霾 denomai	视频 n	Vidéo	2528	526	**	217
雾霾 denomai	严重 a	grave	1389	338	**	275
雾霾 denomai	追寻 v	Poursuivre	89	82	**	41
雾霾 denomai	防尘 n	Anti-poussière	233	102	45	83
雾霾 denomai	城市 n	Ville	1036	239	42	202
雾霾 denomai	我们 r	Nous	2820	481	41	284
雾霾 denomai	有害物质 n	Substance nocive	373	122	38	24
雾霾 denomai	松下 n	Panasonic	270	100	36	2
雾霾 denomai	梦想 n	Rêve	290	104	36	56
雾霾 denomai	让 v	Laisser	2978	484	36	356
雾霾 denomai	隔离病房 n	Salle des contagieux	33	33	35	33
雾霾 denomai	扇 v	Éventer	46	39	34	17
雾霾 denomai	低龄 n	Jeune âge	32	32	34	16
雾霾 denomai	纸质 n	Papier	33	32	33	31
雾霾 denomai	喷雾 n	Spray	286	99	33	55
雾霾 denomai	活性炭 n	Charbon actif	174	74	32	60
雾霾 denomai	抗击 vn	Lutter contre	32	31	32	31
雾霾 denomai	售卖 v	Vendre	35	32	31	31
雾霾 denomai	难民 n	Réfugié	29	29	31	15

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	包围 v	Entourer	68	45	31	45
雾霾 denomai	脚步 n	Pas	77	48	31	47
雾霾 denomai	音量 n	Volume	651	156	30	67
雾霾 denomai	防病毒 n	Antivirus	42	35	30	34
雾霾 denomai	管用 a	Utilie	43	35	30	33
雾霾 denomai	市面上 n	Sur le marché	58	40	29	36
雾霾 denomai	nanoeeng	Nano	115	57	29	1
雾霾 denomai	Timeeng	temps	670	156	28	67
雾霾 denomai	Niuniueng	Niuniu	26	26	28	26
雾霾 denomai	阻碍 v	Contraidre	71	43	27	43
雾霾 denomai	用处 n	Utilité	40	32	27	31
雾霾 denomai	进食 v	Manger	29	27	27	14
雾霾 denomai	渠道 n	Canal	32	28	26	14
雾霾 denomai	N95eng	N95eng	74	43	26	41
雾霾 denomai	厚厚的 u	Épais	71	42	26	42
雾霾 denomai	连续 a	Continu	569	136	26	113
雾霾 denomai	遭遇 n	Rencontre	141	58	25	53
雾霾 denomai	交换 v	Échanger	54	36	25	35
雾霾 denomai	追求 v	La poursuite	128	56	25	56
雾霾 denomai	光明 n	Brillant	79	43	25	43
雾霾 denomai	马路 n	Route	57	36	24	20
雾霾 denomai	烧结 v	Frittage	26	24	24	8
雾霾 denomai	哪里 r	Où	321	92	24	54
雾霾 denomai	中央气象台 nt	Station météo centrale	38	29	23	28
雾霾 denomai	抱 v	Tenir	120	52	23	33
雾霾 denomai	清肺 Dv	Humecter les poumons	41	30	23	22
雾霾 denomai	锡 n	Tin	21	21	23	3
雾霾 denomai	频繁 a	Souvent	71	39	23	39

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	胸闷 n	Oppression thoracique	52	33	22	32
雾霾 denomai	爆发 v	Épidémie	76	40	22	40
雾霾 denomai	中东部 nt	Centre Est	211	69	22	66
雾霾 denomai	心灵 n	Esprit	139	55	22	55
雾霾 denomai	哮喘 n	L'asthme	129	52	22	20
雾霾 denomai	想白 v	Pense blanc	60	35	22	35
雾霾 denomai	致使 v	Causant	52	32	21	30
雾霾 denomai	最火 a	La plupart tirent	40	28	21	28
雾霾 denomai	爆表 v	Table explosive	338	89	21	70
雾霾 denomai	食物 n	Nourriture	185	63	21	31
雾霾 denomai	医院 n	Hôpital	210	67	21	59
雾霾 denomai	拍 v	taper	1520	251	21	170
雾霾 denomai	幸福 a	Bonheur	529	120	21	76
雾霾 denomai	特有 b	spécial	83	40	21	3
雾霾 denomai	technologyeng	Technology	76	38	20	1
雾霾 denomai	我 r	Je	15451	1722	20	1018
雾霾 denomai	专家 n	Expert	459	106	20	86
雾霾 denomai	太阳 n	soleil	456	106	20	78
雾霾 denomai	不适 a	Inconfortable	85	40	20	39
雾霾 denomai	咳 v	Toux	37	26	20	25
雾霾 denomai	播放 v	émettre	581	125	20	72
雾霾 denomai	呼伦贝尔 city	Ville de Hulunbeier	26	22	20	1
雾霾 denomai	雨雪 n	Pluie et neige	116	47	20	29
雾霾 denomai	购物车 n	Panier	77	38	20	19
雾霾 denomai	防病 disease	Prévention des maladies	17	17	19	17
雾霾 denomai	滋润 n	Hydratant	104	43	19	42
雾霾 denomai	灾害 n	Catastrophe	36	25	19	15
雾霾 denomai	JJBOOMeng	JJBOOM	17	17	19	4

Continuation du tableau

Pôle	Cooccurrent	Traduction	FQ(cooc)	NB_co-freq	Indice	NB_contextes
雾霾 denomai	侵害 v	Violation	135	50	19	47
雾霾 denomai	气候 n	Climat	109	44	19	21
雾霾 denomai	小妖 n	Petit démon	19	18	19	5
雾霾 denomai	起来 v	Lève-toi	605	125	19	99
雾霾 denomai	防御 v	Défense	61	33	19	27
雾霾 denomai	变异性 n	Variabilité	16	16	18	16
雾霾 denomai	一片 m	un morceau	332	83	18	74
雾霾 denomai	每个 r	Chaque	261	71	18	62
雾霾 denomai	送 v	Envoyer	546	115	18	80
雾霾 denomai	聚惠 vn	Juhui	85	38	18	1
雾霾 denomai	能 v	Peut	2468	351	18	236
雾霾 denomai	车灯 n	Phares	50	29	18	17
雾霾 denomai	应对 v	Faire face	186	58	18	46
雾霾 denomai	空调 n	Climatisation	858	158	18	25
雾霾 denomai	国人 n	Chinois	44	27	18	26
雾霾 denomai	推荐 v	Recommandé	501	108	18	94
雾霾 denomai	检测仪 n	Détecteur	262	69	17	52

La fin du tableau

17 Liste des segments répétés de 雾霾 (brouillard de pollution) et 天气 (temps) de l'Ins

Tab. 36 – Liste des segments répété de 雾霾 (brouillard de pollution) et 天气 (temps) de l'Ins

Le début du tableau			
Fréquence	Segment	Longueur	Traduction
	雾霾 denomai 天气 n	2	le temps *wumai*
1829			
234	x 雾霾 denomai 天气 n	3	le temps *wumai*
199	的 u 雾霾 denomai 天气 n	3	le temps *wumai*
153	雾霾 denomai 天气 n 的 u	3	le temps *wumai*
114	雾霾 denomai 天气 n 。 x	3	le temps *wumai*
102	。 x 雾霾 denomai 天气 n	3	le temps *wumai*
88	雾霾 denomai 天气 n 将 d	3	le temps *wumai* va
82	出现 v 雾霾 denomai 天气 n	3	le temps *wumai* apparaît
71	应对 v 雾霾 denomai 天气 n	3	affronter le temps *wumai*
62	雾霾 denomai 天气 n 过程 n	3	la duration du temps *wumai*
52	持续 vd 雾霾 denomai 天气 n	3	le temps *wumai* va durer
51	大 a 范围 n 雾霾 denomai 天气 n	4	le temps *wumai* s'étend à grande échelle
42	地区 n 雾霾 denomai 天气 n	3	le temps *wumai* dans la région de
42	雾霾 denomai 天气 n 对 p	3	le temps *wumai* va influencer
42	严重 a 雾霾 denomai 天气 n	3	le temps *wumai* est grave
42	在 p 雾霾 denomai 天气 n	3	dans le temps *wumai*
41	雾霾 denomai 天气 n 影响 vn	3	le temps *wumai* va influencer
38	有 v 雾霾 denomai 天气 n	3	il y a le temps *wumai*
36	雾霾 denomai 天气 n 持续 vd	3	le temps *wumai* va durer
34	遭遇 n 雾霾 denomai 天气 n	3	subir le temps *wumai*
28	雾霾 denomai 天气 n 将 d 自 p	4	le temps *wumai* va du xxx au xxx
27	是 v 雾霾 denomai 天气 n	3	est le temps *wumai*
25	雾霾 denomai 天气 n 将 d 自 p 北向南 nr	5	le temps *wumai* va du nord au sud
24	对 p 雾霾 denomai 天气 n	3	contre le temps *wumai*

17 Liste des segments répétés de 雾霾 (brouillard de pollution) et 天气 (temps) de l'Ins

Continuation du tableau			
Fréquence	Segment	Longueur	Traduction
23	重度 n 雾霾 denomai 天气 n	3	le temps *wumai* atteint le pic de pollution
22	上述 b 地区 n 雾霾 denomai 天气 n	4	le temps *wumai* de ces régions mentionnées
22	雾霾 denomai 天气 n 时 n	3	lorsqu'il fait le temps *wumai*
21	x 上述 b 地区 n 雾霾 denomai 天气 n	5	le temps *wumai* de ces régions mentionnées
20	出现 v 雾霾 denomai 天气 n 。x	4	le temps *wumai* apparaît
20	雾霾 denomai 天气 n 中 f	3	dans le temps *wumai*
19	的 u 雾霾 denomai 天气 n 将 d	4	le temps *wumai* va
19	将 d 有 v 雾霾 denomai 天气 n	4	il y aura le temps *wumai*
19	地区 n 雾霾 denomai 天气 n 将 d	4	le temps *wumai* de ces régions va
19	x 雾霾 denomai 天气 n 将 d	4	le temps *wumai* va
17	雾霾 denomai 天气 n 将 d 自 p 北向南 nr 逐渐 d	6	le temps *wumai* va du nord au sud devenir progressivement
17	的 u 雾霾 denomai 天气 n 。x	4	le temps *wumai* de xxx
17	雾霾 denomai 天气 n 频发 d	3	le temps *wumai* est fréquent
17	雾霾 denomai 天气 n 多发 m	3	le temps *wumai* est fréquent
16	雾霾 denomai 天气 n 将 d 自 p 北向南 nr 逐渐 d 减弱 v	7	le temps *wumai* va du nord au sud s'affaiblir progressivement
16	受 v 雾霾 denomai 天气 n	3	sous l'influence du temps *wumai*
16	此次 r 雾霾 denomai 天气 n	3	le temps *wumai* de cette fois-ci
16	雾霾 denomai 天气 n 形成 v	3	le temps *wumai* se forme
15	受 v 雾霾 denomai 天气 n 影响 vn	4	sous l'influence du temps *wumai*
15	中东部 nt 地区 n 雾霾 denomai 天气 n	4	le temps *wumai* des régions du centre et de l'est
15	多日 m 的 u 雾霾 denomai 天气 n	4	le temps *wumai* de ces derniers jours
15	雾霾 denomai 天气 n 的 u 形成 v	4	le temps *wumai* résulte de

Continuation du tableau			
Fréquence	Segment	Longueur	Traduction
15	雾霾 denomai 天气 n 里 f	3	dans le temps *wumai*
14	雾霾 denomai 天气 n 应急 vn	3	les nécessités urgentes face au temps *wumai*
13	出现 v 重度 n 雾霾 denomai 天气 n	4	a subi le grave temps *wumai*
13	持续性 n 雾霾 denomai 天气 n	3	le continuum du temps *wumai*
12	雾霾 denomai 天气 n 北京 city	3	le temps *wumai* de Beijing
12	造成 v 雾霾 denomai 天气 n	3	a causé le temps *wumai* va
12	雾霾 denomai 天气 n 导致 v	3	le temps *wumai* va conduire à
11	雾霾 denomai 天气 n 。x 中央气象台 nt	4	le temps *wumai*. le Centre météorologique de Chine
11	造成 v 雾霾 denomai 天气 n 的 u	4	a causé le temps *wumai*
11	持续 vd 的 u 雾霾 denomai 天气 n	4	le temps *wumai* va durer
10	中东部 nt 持续 vd 雾霾 denomai 天气 n	4	le temps *wumai* va durer encore quelque jours dans les régions du centre et de l'est
10	雾霾 denomai 天气 n 少 a	3	peu du temps *wumai*
La fin du tableau			

Glossaire

1 Glossaire

AFC : analyse factorielle des correspondances. Famille de méthodes statistiques d'analyse multidimensionnelles 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.
Algorithme : ensemble des règles opératoires propres à un calcul.
Annotation : processus permettant de fournir des informations linguistiques à certains segments selon différents niveaux, annotation phonétique, annotation grammaticale, annotation sémantique, etc.
APN : Assemblée populaire nationale.
AQISTUDY : la plateforme en ligne de détection et d'analyse de l'IQA (voir IQA) de Chine.
CCPPC : Conférence consultative politique chinoise et du Congrès national du peuple **Charbon à gaz** : une politique prise par le gouvernement chinois à la fin de l'année 2017, qui interdit l'utilisation du charbon dans 28 villes et encourage l'utilisation du gaz. **Classe sémantique** : domaine, taxème, dimension.
Caractère : signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.
Clé chinois : élément servant à classer les caractères chinois et à les retrouver dans un dictionnaire.
Concordance : ensemble de lignes de contexte se rapportant à une même forme-pôle.
Cooccurrence/Cooccurrent : 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 formes données.
Corpus : ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.

Délimiteurs de séquence : 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).
Dialectique : la dialectique étudie la succession des intervalles dans le temps textuel, comme les états qui y prennent place et les processus qui s'y déroulent (Rastier 2005).
Dialogique : la dialogique fonde la typologie des énonciateurs représentés, et rend compte des modalités énonciatives et évaluatives. La dialogique introduit des différences au niveau des acteurs et de leurs positionnements énonciatifs.
Dimension : classe souvent binaire exprimant une opposition générale (féminin/ vs /masculin//animé/ vs /inanimé//mélioratif/ vs /dépréciatif/etc.).
Discours : ensemble d'usages linguistiques codifiés attachés à un type de pratique sociale (Rastier 2001).
Sous-corpus institutionnel (Ins) : textes produit par un « énonciateur singulier ou collectif qui occupe une position juridiquement inscrite dans l'appareil d'état » (Oger et Ollivier-Yaniv 2003). Les articles publiés sur le site du gouvernement chinois par les éditeurs officiels peuvent être tous définis comme des textes institutionnels.
Sous-corpus informel-médiatique (InfM) : textes produit dans un cadre socio-économique et sémiologique en dehors du cadre institutionnel, par les rédacteurs employés professionnels pour « maintenir leur position compétitive et de capter le plus grand nombre possible d'audiences ».
Sous-corpus institutionnel-médiatique (InsM) : textes produits par des journalistiques, des rédacteurs d'institutions pour un groupe d'« audiences cibles » à l'aide d'une mise en forme médiatique et un mode de transmission sémiotique, l'InsM comprend l'ensemble des textes que l'on peut considérer à des degrés divers comme des discours autorisés dans un milieu donné (en ou hors contexte officiel), sans référence nécessaire à l'état (production des syndicats, des états - majors des partis politiques, etc.).
Sous-corpus Profane (Profane) : discours vernaculaire, discours du réseaux sociaux, c'est les weibo produits directement dans les environnements du web, sur les réseaux sociaux numériques, sous forme textuelle (voire agrémentée d'éléments audiovisuels) et assortie de méta- données, donc plurisémiotique. Il s'agit

d'énoncés natifs du web. Étant donné qu'ils sont produits dans le cadre de polylogues asynchrones médiés par ordinateurs, ils partagent certains traits avec le langage parlé et sont plus ou moins fortement la transcription d'une oralité.
Document numérique : document poly-sémiotique dématérialisé destiné à être lu sur un écran et non sur papier. Ex. Un blog est un document numérique.
Domaine : ensemble de taxèmes correspondant à une pratique déterminée.
Émoticône : signe qui imite une émotion, afin de la rendre perceptible au cours de l'énonciation d'un contenu, dans le cadre d'une communication médiée par un ordinateur (Halté, 2013). Les émoticônes participent donc au sens des productions discursives natives du web. Celles-ci prennent la forme soit d'une suite de caractères de ponctuation et alphabétiques, soit d'une icône, le plus souvent de la hauteur d'un interligne, insérée dans le texte. Il est souvent fait référence aux émoticônes par leur équivalent anglais smileys.
Fond sémantique : ensemble des isotopies d'un texte.
Forme (ou forme graphique) : archétype correspondant aux occurrences identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence.
Forme caractéristique : (d'une partie) synonyme de spécificité positive.
Forme sémantique : regroupement syntagmatique de sèmes stabilisé en corpus.
Fréquence : (d'une unité textuelle) le nombre de ses occurrences dans le corpus.
Fréquence maximale : fréquence de la forme la plus fréquente du corpus.
Fréquence relative : la fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties rapportée à la taille du corpus (resp. de cette partie).
Genre : programme de prescriptions qui règlent la production et l'interprétation d'un texte (Rastier 2001).
IQA : abréviation du terme « Indice de la qualité de l'air ». **Hapax** : forme dont la fréquence est égale à un dans le corpus (hapax du corpus) ou dans une de ses parties (hapax de la partie).
HEPA : High Efficiency Particulate Air.
Index alphabétique : index dans lequel les formes-pôles sont classées selon l'ordre alphabétique.
Index hiérarchique : index dans lequel les formes-pôles sont classées selon l'ordre de la fréquence des formes généralement par ordre de fréquence décroissante.

Isotopie : récurrence d'un sème donné et d'empan variés (de la collocation à l'intertexte). **Lemmatisation** : regroupement sous une forme canonique (lemme) des occurrences du texte.

Lemme : forme canonique du mot à partir de laquelle sont dérivées les formes fléchies.

Lexicométrie : exploration statistique de données lexicales. La discipline de l'analyse des données textuelles tarde à fixer sa terminologie (Mayaffre, 2004), mais, que l'approche des données se fasse par le prisme du lexique, du texte ou du discours, les calculs tendent tous vers le même objectif d'exploration statistiques de productions langagières sous forme de données numériques. Il s'agit d'un 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. Synonymes textométrie & logométrie.

Longueur : (d'un corpus d'une partie de ce corpus, d'un fragment de texte, d'un segment, etc.) le nombre des occurrences contenues dans ce corpus.

Macrosémantique : la macrosémantique vise à étudier les caractéristiques du genre textuel.

Mésosémantique : la mésosémantique rend compte du palier intermédiaire entre la lexie et le texte. Elle traite donc de la phrase, ou plus précisément de l'espace qui s'étend du syntagme pourvu d'une fonction syntaxique jusqu'à la phrase complexe et à ses connexions immédiates.

Microsémantique : la microsémantique vise à étudier la sémantique du palier inférieur du texte sur les unités minimales d'un signe, allant des morphèmes (le signe linguistique minimal s'appelle un morphème) jusqu'à la lexie (unité fonctionnelle, qui regroupe plusieurs morphèmes, parfois une lexie ne correspond qu'à une seule position).

miniblog : Les messages publiés sur le réseau social WEIBO est appelés weibo, que l'on traduit en français en miniblog.

N95 (respirateur) : Le respirateur peut bloquer au moins 95% des particules de diamètre supérieur inférieur ou égal à 0,3 µm.

Occurrence : suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs de forme.

Partie : (d'un corpus de textes) fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

Partition : (d'un corpus de textes) division d'un corpus en parties constituées par des fragments de texte consécutifs n'ayant pas d'intersection commune et dont la réunion est égale au corpus.
Passage : association stabilisée d'un fond et d'une forme sémantique caractéristique d'un usage lexical d'un texte ou d'un corpus.
PM10 : particule ultrafine dont le diamètre est inférieur à 10 micromètres. Elles sont désignées sous le terme de PM 10 (d'après la terminologie anglaise particulate matter).
PM2,5 : particule ultrafine dont le diamètre est inférieur à 2,5 micromètres. Elles sont désignées sous le terme de PM2,5 (d'après la terminologie anglaise particulate matter).
Profil : ensemble des informations d'un membre permettant de l'identifier pseudonyme, avatar, statut, nombre d'interventions, etc.
Quatre composantes sémantiques : Dialectique, dialogique, tactique et thématique. Cf. Dialectique, Dialogique, Tactique et Thématique.
Répartition : (des occurrences d'une forme dans les parties du corpus) nombre des parties du corpus dans lesquelles cette forme est attestée.
Réseau sémique : fond ou forme sémantique (i.e. les objectivations sémantiques à l'exception du signifié).
Réseaux sociaux numériques : ensemble des applications web servant à constituer un réseau social virtuel, en permettant aux internautes d'élaborer une identité sociale en ligne et d'interagir entre eux.
Rubrique : niveau de structuration le plus haut et le plus figé des forums étudiés. Les rubriques sont créées par les administrateurs du forum et non par n'importe quel membre du forum. Elles sont déterminées de manière verticale et constituent le cadre dans lequel les internautes interviennent.
SC : Système circulaire d'« exportation-importation-exportation » de sous-corpus Profane.
sc : sème /cause/.
SE : Système d'« exportation » de sous-corpus Ins.
SM : Système mixe d'« exportation » et d'« importation » additionnel optionnel des sous-corpus InfM et InsM.
sm : sème /maladie/.
sp : sème /mesures préventives/.

Segment : toute suite d'occurrences consécutives dans le corpus et non séparées par un Séparateur de séquence est un segment du texte.
Segment répété : suite de forme dont la fréquence est supérieure ou égale à 2 dans le corpus.
Segmentation : opération qui consiste à délimiter des unités minimales dans un texte.
Segmentation automatique : ensemble d'opérations réalisées au moyen de procédures informatisées qui aboutissent à découper, selon des règles prédéfinies, un texte stocké sur un support lisible par un ordinateur en unités distinctes que l'on appelle des unités minimales.
Sème : propriété sémantique d'ordre métalinguistique résultant d'une validation par le linguiste. Les regroupements paradigmatiques de sèmes constituent des fonds et des formes sémantiques. Les regroupements syntagmatiques des sémèmes ou des signifiés. Ex. Dans le signifié de 'chaise' /pour s'asseoir/ est un sème.
Séquence : suite d'occurrences du texte non séparées par un délimiteur de séquence.
Seuil : 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.).
Site traditionnel : Les sites d'Internet médiatiques ou institutionnels.
Signifiant : La partie graphique du signe.
Signifié : Le contenu sémantique d'un signe exprimé en collection de sèmes (sémème et sémie sont deux manières de qualifier le signifié).
Spécificité négative : 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 au seuil fixé au départ.
Spécificité positive : 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 (ou forme caractéristique de cette partie), 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.
Stable : dont la structure sémique varie peu (notamment statistiquement) dans un corpus donné -dans un genre, un discours ou un domaine.
Tactique : la tactique rend compte de la linéarité du signifié et de la disposition des unités textuelles (Rastier 2005).
Taille : (d'un corpus) sa longueur mesurée en occurrences (de formes simples).
TAL : traitement automatique des langues.
Taxème : petite classe sémantique correspondant à une situation pratique précise. La cohésion de la classe est assurée par les sèmes génériques.
Textométrie : exploration statistique de données lexicales. La discipline de l'analyse des données textuelles tarde à fixer sa terminologie (Mayaffre, 2004), mais, que l'approche des données se fasse par le prisme du lexique, du texte ou du discours, les calculs tendent tous vers le même objectif d'exploration statistiques de productions langagières sous forme de données numériques. Il s'agit d'un 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. Synonymes lexicométrie & logométrie.
Thématique : la thématique rend compte des thèmes, décrits comme des formes sémantiques (Rastier 2005).
Thème : une structure stable de traits sémantiques, c'est à dire les sèmes, récurrents dans un corpus, et susceptibles de lexicalisations diverses.
Topographie textuelle : la représentation graphique des phénomènes langagiers mis en évidence par l'étude statistique afin d'apprécier leurs positions dans le texte.
Trois systèmes : Système unidirectionnel d'« exportation » de l'Ins ; Système d'« exportation » et d'« importation » additionnel à option des discours InsM et InfM ; Système circulaire d'« exportation-importation-exportation » des discours profane.
Trois genres traditionnels : Ins, InsM, InfM et Profane.

Ventilation : (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.

Trois paliers sémantiques : le palier macrosémantique, le palier mésosémantique et le palier micro-sémantique.Cf. Macrosémantique, Mésosémantique et Microsémantique.

Vocabulaire : ensemble des formes attestées dans un corpus de textes.

weibo : microblogs produits par les internautes dans le site de Sina WEIBO.

wumai : transcription en pinyin de mots chinois 雾霾. Nous utilisons la transcription pour faire référence à ce mot chinois.

2 Tags proposés par JIEBA

a : adjectifs.
ad ou d : adverbes.
an : adjectifs nominaux.
b/nn : mots distinctifs.
c/cc : conjonctions.
denomai : mots qui contiennent le caractère 霾 (dont la transcription en pinyin est le mai).
denopm : terminologies de pm2,5 et pm10.
denopollu : deux mots 空气污染 (pollution de l'air) et 大气污染 (pollution atmosphérique).
denowu : mots qui contiennent le caractère 雾 (dont la transcription en pinyin est le wu).
disease : termes de maladie.
Dns : terminologies médicales.
Dv : les locutions verbales relatives, par exemple : 看病 Dv (voir le médecin).
e : interjections.
f/lc : mots de localité.
i : expressions proverbiales.
j : abréviations ou des sigles.
l : idiomes.
n : substantifs.
nr : noms de personnes.
nt : établissements.
nw : néologismes.
nz : noms propres.
o : onomatopées.
p : prépositions.
pba : prépositions de 把.
pbei : prépositions de 被.
q/m : quantifieurs.
r : pronoms.
s/lc : noms de lieux.
symptom : symptômes de maladies.

t : termes de temps.
u : auxiliaires.
ule : auxiliaires 了.
uzhe : auxiliaires 着.
v : verbes.
vd : expressions adverbiales.
vn : verbes normalisés.
w/pu : ponctuations.
x : non-séquentiels morphèmes.
y/sp : particules modales.
z : mots qui indiquent un état.

Bibliographie

Adam, Jean-Michel. «Genres, textes, discours : pour une reconception linguistique du concept de genre». In : *Revue belge de philologie et d'histoire* 75.3 (1997), p. 665–681. doi : 10.3406/rbph.1997.4188.

Airhihenbuwa Collins, Obregon Rafael. «A Critical Assessment of Theories/Models Used in Health Communication for HIV/AIDS». In : *Journal of health communication* 5 Suppl (fév. 2000), p. 5–15. doi : 10.1080/10810730050019528.

André Salem, Ludovic Lebart. *Statistique textuelle (French Edition)*. Dunod, 1994. isbn : 2100022393.

Arbex, Marcos Abdo et al. «Air pollution and the respiratory system». In : *Jornal Brasileiro de Pneumologia* 38.5 (oct. 2012), p. 643–655. issn : 1806-3713. doi : 10.1590/S1806-37132012000500015. url : http://www.scielo.br/scielo.php?script=sci_abstract&pid=S1806-37132012000500015&lng=en&nrm=iso&tlng=en (visité le 26/09/2018).

Bai, Zhipeng, Jinbao Han et Merched Azzi. «Insights into measurements of ambient air PM2.5 in China». In : *Trends in Environmental Analytical Chemistry* 13 (2017), p. 1–9. issn : 2214-1588. doi : https://doi.org/10.1016/j.teac.2017.01.001. url : http://www.sciencedirect.com/science/article/pii/S2214158816300654.

Ballabriga, M Michel. «LA SEMANTIQUE TEXTUELLE 1». fr. In : *Texto !* (2005), p. 9.

Bamman, David, Brendan O'Connor et Noah Smith. «Censorship and deletion practices in Chinese social media». In : *First Monday* 17.3 (2012). doi : 10.5210/fm.v17i3.3943.

Beaudouin, Valérie. *Mètre et rythmes du vers classique : Corneille et Racine*. Honoré Champion, 2002. isbn : 2745305093.

Beauvisage, Thomas. «Exploiter des données morphosyntaxiques pour l'étude statistique des genres - Application au roman policier». In : *Texto !* (2001).

Belghanem, Ali. «La sémantique interprétative Du mot au corpus et du sème aux formes sémantiques». fr. In : *Texto !* (2014), p. 15.

Benzécri, Jean-Paul. *L'analyse des données*. Sous la dir. de Dunod. 1973.

– *L'analyse des données, tome 1 : La taxinomie*. Sous la dir. de Dunod. 1984. isbn : 2040156097.

– *L'ANALYSE DES DONNEES. Tome 2*. Sous la dir. de Bordas Editions. 1993. isbn : 2040155155.

Biber, Douglas. «Conversation text types : A multi-dimensional analysis». en. In : *DIMENSIONAL ANALYSIS* (2004), p. 20.

– «The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings». In : *Computers and the Humanities* 26.5/6 (1992), p. 331–345. issn : 0010-4817. url : https://www.jstor.org/stable/30204629 (visité le 13/11/2018).

Bonhomme, Marc. «La problématique des genres de discours dans la communication sur Internet». fr. In : *Travaux neuchâtelois de linguistique* 63 (2015), p. 17.

Bourion, Évelyne et Denise Malrieu. «Concepts, systèmes signifiants et organisation d'un domaine». fr. In : *Classiques Garnier* (2012). doi : 10.15122/isbn.978-2-8124-4316-9.p.0085.

Bourion-Jacquemin, Evelyne. «L'aide à l'interprétation des textes électroniques». français. Thèse doctorat. France : Université de Nancy II, 2001.

Boutet, Josiane, Bernard Gardin et Michèle Lacoste. «Discours en situation de travail». In : *Langages* 29.117 (1995), p. 12–31. doi : 10.3406/lgge.1995.1703.

Boutet, Josiane et Dominique Maingueneau. «Sociolinguistique et analyse de discours : façons de dire, façons de faire». In : *Langage et société* 114.4 (2005), p. 15. doi : 10.3917/ls.114.0015.

Bronckart, Jean-Paul. «GENRES DE TEXTES, TYPES DE DISCOURS ET « DEGRÉS » DE LANGUE». fr. In : *Texte inédit prononcé au Deuxième congrès international d' interactionnisme socio-discursif (ISD2), Lisbonne, 10-13 octobre 2007* vol. XIII.nº 1 (2008), p. 96.

Calabrese, Laura. «Rectifier le discours d' information médiatique. Quelle légitimité pour le discours profane dans la presse d' information en ligne ?» fr. In : *Les Carnets du Cediscor. Publication du Centre de recherches sur la didac-*

ticité des discours ordinaires 12 (fév. 2014), p. 21–34. ISSN : 1242-8345. URL : http://journals.openedition.org/cediscor/916 (visité le 30/08/2019).

CHARAUDEAU, Patrick. «Le discours d'information médiatique. La construction du miroir social». fr. In : *Mots. Les langages du politique* 72 (juil. 2003), p. 181–182. ISSN : 0243-6450. URL : http://journals.openedition.org/mots/6763 (visité le 30/08/2019).

CHEN, R. et al. «Association of Particulate Air Pollution With Daily Mortality: The China Air Pollution and Health Effects Study». en. In : *American Journal of Epidemiology* 175.11 (juin 2012), p. 1173–1181. ISSN : 0002-9262, 1476-6256. DOI : 10.1093/aje/kwr425. URL : https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwr425 (visité le 26/09/2018).

CHEN, Yuyu et al. «Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy». en. In : *Proceedings of the National Academy of Sciences* 110.32 (août 2013), p. 12936–12941. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.1300018110. URL : http://www.pnas.org/content/110/32/12936 (visité le 26/09/2018).

CHEN, Yuyu et al. «Gaming in Air Pollution Data? Lessons from China». In : *The B.E. Journal of Economic Analysis & Policy* 12.3 (2012). ISSN : 1935-1682. DOI : 10.1515/1935-1682.3227. URL : https://www.degruyter.com/view/j/bejeap.2012.12.issue-3/1935-1682.3227/1935-1682.3227.xml (visité le 20/09/2018).

CHETOUANI, Lamria. «Les mots de la controverse sur le changement climatique». fr. In : *Le Telemaque* n° 31.1 (2007), p. 81–104. ISSN : 1263-588X. URL : https://www.cairn.info/revue-le-telemaque-2007-1-page-81.htm (visité le 09/08/2019).

CLAS, André. «TOURNIER, Jean (1991) : Structure lexicales de l'anglais. Guide alphabétique, Paris, Nathan, Collection Nathan-Université, 190 p.» In : *Meta : Journal des traducteurs* 38.2 (1993), p. 349. DOI : 10.7202/002259ar.

DANG, qinran, Nicolas TURENNE et Mathieu VALETTE. «Using smog-related data of Chinese Sina Weibo to explore correlation between health issues and relevant regions». In : *Natural Language Processing and Cognitive Science* (2018).

DENG, Qi et al. «中文文本体裁分类中特征选择的研究 (Research on Feature Selection in Chinese Text Genre Classification)». In : *Computer Engineering* (2008).

DENISE, Malrieu et Rastier FRANÇOIS. «Genres et variations morphosyntaxiques». In : *Texto !* (2001). URL : http://www.revue-texto.net/Inedits/Malrieu_Rastier/Malrieu-Rastier_Genres1.html (visité le 09/10/2018).

EENSOO, Egle et Mathieu VALETTE. «Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité». fr. In : *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015), Caen (France)* (2015), p. 13.

FIALA, Pierre. «L'interprétation en lexicométrie. Une approche quantitative des données lexicales». In : *Langue française* 103.1 (1994), p. 113–122. DOI : 10.3406/lfr.1994.5731.

FLØTTUM, Kjersti et al. «Representations of the future in English language blogs on climate change». In : *Global Environmental Change* 29 (nov. 2014), p. 213–222. DOI : 10.1016/j.gloenvcha.2014.10.005.

FOREST, Dominic et Hélène BROUSSEAU. «L' environnement vu par ses documents : utilisation de techniques de fouille de textes dans un contexte de description linguistique». fr. In : *Université de Montréal* (2016), p. 12.

FRANCESIAZ, Théo, Raphaël GRAILLE et Brahim METAHRI. «Introduction aux modèles probabilistes utilisés en Fouille de Données». fr. In : (2015), p. 27.

FREED, Alice F. «Institutional Discourse». en. In : *The International Encyclopedia of Language and Social Interaction*. American Cancer Society, 2015, p. 1–18. ISBN : 978-1-118-61146-3. DOI : 10.1002/9781118611463.wbielsi151. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118611463.wbielsi151 (visité le 12/06/2019).

GENETTE, Gérard. *Seuils (French Edition)*. Editions Du Seuil, 2002. ISBN : 978-2-02-052641-8.

GJERSTAD, Øyvind. «Web 2.0 et genres discursifs : l'exemple de blogs sur le changement du climat». fr. In : *Synergies Pays Scandinaves* 9 (2014), p. 13.

GLEDHILL, Christopher, Stéphane PATIN et Maria ZIMINA. «Lexico-grammaire et textométrie : identification et visualisation de schémas lexico-grammaticaux caractéristiques dans deux corpus juridiques comparables en français». fr. In : *Corpus* 17 (jan. 2017). ISSN : 1638-9808. URL : http://journals.openedition.org/corpus/2868 (visité le 18/09/2018).

Gold, D. R. et J. M. Samet. «Air Pollution, Climate, and Heart Disease». en. In : *Circulation* 128.21 (nov. 2013), e411–e414. issn : 0009-7322, 1524-4539. doi : 10.1161/CIRCULATIONAHA.113.003988. url : http://circ.ahajournals.org/cgi/doi/10.1161/CIRCULATIONAHA.113.003988 (visité le 26/09/2018).

Gomez-Mejia, Gustavo. «La métropole parisienne au prisme du réseau : réalités discursives et marqueurs symboliques». fr. In : *Quaderni. Communication, technologies, pouvoir* 73 (oct. 2010), p. 53–64. issn : 2105-2956. doi : 10.4000/quaderni.447. url : http://journals.openedition.org/quaderni/447 (visité le 30/08/2019).

Gonçalves, Matilde. «Similitudes et différences textuelles dans les genres numériques : blog et site web.» FR. In : *Studii de Lingvistică* 2014, Vol. 4 (2014), p75–91.

Greimas, A.-J. *Structural Semantics : An Attempt at a Method*. University of Nebraska Press, 1984. isbn : 978-0803221123.

Guetzkow, Harold. «Unitizing and categorizing problems in coding qualitative data». en. In : *Journal of Clinical Psychology* 6.1 (jan. 1950), p. 47–58. issn : 00219762, 10974679. doi : 10.1002/1097-4679(195001)6:1<47::AID-JCLP2270060111>3.0.CO;2-I. url : http://doi.wiley.com/10.1002/1097-4679%28195001%296%3A1%3C47%3A%3AAID-JCLP2270060111%3E3.0.CO%3B2-I (visité le 26/09/2018).

Gunson, A. J. et Marcello M. Veiga. «Mercury and Artisanal Mining in China». In : *Environmental Practice* 6.2 (2004), p. 109–120. doi : 10.1017/s1466046604000225.

Halté, Pierre. «Émoticône et modalisation : ancrage énonciatif du locuteur dans un corpus de chatPierre Halté». In : *Les émotions à travers les corpus*. Poitiers, France, sept. 2014. url : https://hal.archives-ouvertes.fr/hal-01618894 (visité le 26/08/2019).

Hébert, Louis. «La sémantique interprétative en résumé». In : *Texto !* (2002).

Hồ-Đình, Océane. «Caractérisation différentielle de forums de discussion sur le VIH en vietnamien et en français : Éléments pour la fouille comportementale du web social». Thèse de doct. INALCO, déc. 2017.

Hồ-Đình, Océane et Mathieu Valette. «Analyse différentielle des discours de prévention du VIH : textes institutionnels et textes informels en français et

en vietnamien». In : *JADT 2014 : 12es Journées internationales d' Analyse statistique des Données Textuelles* (2014).

Hopke, Philip K. «Contemporary threats and air pollution». In : *Atmospheric Environment*. Atmospheric Environment - Fifty Years of Endeavour 43.1 (jan. 2009), p. 87–93. ISSN : 1352-2310. DOI : 10.1016/j.atmosenv.2008.09.053. URL : http://www.sciencedirect.com/science/article/pii/S1352231008009151 (visité le 26/09/2018).

Huchet, Jean-Françaois. *La crise environnementale en Chine : Evolution et limites des politiques publiques*. Les Presses de Sciences Po, 2016. ISBN : 9782724619508.

Jackiewicz, Agata. «Matérialité linguistique des controverses sociétales. Rapports intersubjectifs et interdiscursifs dans des tweets polémiques». In : *SHS Web of Conferences* 27 (2016). Sous la dir. de F. Neveu et al., p. 02008. DOI : 10.1051/shsconf/20162702008.

Kan, Haidong, Renjie Chen et Shilu Tong. «Ambient air pollution, climate change, and population health in China». In : *Environment International*. Emerging Environmental Health Issues in Modern China 42 (juil. 2012), p. 10–19. ISSN : 0160-4120. DOI : 10.1016/j.envint.2011.03.003. URL : http://www.sciencedirect.com/science/article/pii/S0160412011000535 (visité le 26/09/2018).

Karlgren, Jussi. «The Wheres and Whyfores for Studying Textual Genre Computationally». In : *SICS - Swedish Institute of Computer Science* (avr. 2012).

Karlgren, Jussi et Douglass Cutting. «Recognizing Text Genres with Simple Metrics Using Discriminant Analysis». In : *arXiv :cmp-lg/9410008* (oct. 1994). arXiv : cmp-lg/9410008. URL : http://arxiv.org/abs/cmp-lg/9410008 (visité le 13/11/2018).

Kessler, Brett, Geoffrey Nunberg et Hinrich Schutze. «Automatic Detection of Text Genre». In : *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. Madrid, Spain : Association for Computational Linguistics, juil. 1997, p. 32–38. DOI : 10.3115/976909.979622. URL : http://www.aclweb.org/anthology/P97-1005 (visité le 13/11/2018).

Krieg-Planque, Alice. «Thomas Fromentin, Stéphanie Wojcik éd., Le profane en politique. Compétences et engagements du citoyen». fr. In : *Mots. Les langages du politique* 92 (mai 2010), p. 121–129. ISSN : 0243-6450. URL : http://journals.openedition.org/mots/19599 (visité le 30/08/2019).

Labbé, Cyril. «Lexicométrie : quels outils pour les sciences humaines et sociales ?» In : *Usages de la lexicométrie en sociologie*. Guyancourt, France, 2013. URL : https://hal.archives-ouvertes.fr/hal-00834039 (visité le 20/09/2019).

Lafon, Pierre. «Sur la variabilité de la fréquence des formes dans un corpus». In : *Mots* 1.1 (1980), p. 127–165. DOI : 10.3406/mots.1980.1008.

Larssen, Thorjorn et al. «Acid Rain in China». In : *Environmental Science & Technology* 40.2 (2006), p. 418–425. DOI : 10.1021/es0626133.

Lavorel, Pierre. «Éléments pour un calcul du sens». fre. Thèse de doct. Paris Saint-Sulpice-de-Favières : Dunod Association Jean-Favard pour le développement de la linguistique quantitative, 1975.

Li, Yuan et al. «What are Chinese talking about in hot weibos ?» In : *Physica A : Statistical Mechanics and its Applications* 419 (fév. 2015), p. 546–557. DOI : 10.1016/j.physa.2014.10.043.

Liu, Jianguo et Jared Diamond. «China's environment in a globalizing world». In : *Nature* 435.7046 (2005), p. 1179–1186. DOI : 10.1038/4351179a.

Liu, Jianguo et al. «Effects of household dynamics on resource consumption and biodiversity». In : *Nature* 421.6922 (2003), p. 530–533. DOI : 10.1038/nature01359.

Loiseau, Sylvain. «Sémantique du discours philosophique : du corpus aux normes : autour de G. Deleuze et des années 60». français. Thèse de doctorat. France : Université Paris Nanterre, 2006.

Loiseau, Sylvain, Céline Poudat et Driss Ablali. «Exploration contrastive de trois corpus de sciences humaines». fr. In : *8ème Journées d'analyse des données textuelles (JADT 2006), 2006, Besançon, France.* (2006), p. 12.

Maingueneau, Dominique. «L'analyse du discours, introduction aux lectures de l'archive». fre. In : *Mots, n°29, décembre 1991. Politique et sport. Retours de Chine, sous la direction de Simone Bonnafous* . (1991). ISSN : 0243-6450. URL : https://www.persee.fr/doc/mots_0243-6450_1991_num_29_1_1658.

Malrieu, Denise. «Genre textuel, surlignages et marques linguistiques d'importance». In : *Linx* 31.2 (1994), p. 123–140. DOI : 10.3406/linx.1994.1329.

– «Linguistique de corpus, genres textuels, temps et personnes». In : *Langages* 38.153 (2004), p. 73–85. DOI : 10.3406/lgge.2004.935.

MATUS, Kira et al. «Health damages from air pollution in China». In : *Global Environmental Change* 22.1 (fév. 2012), p. 55–66. DOI : 10.1016/j.gloenvcha.2011.08.006.

MAYAFFRE, Damon. «Quand “travail ”, “famille ”, “patrie ” co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur la co-occurrence». In : *JADT 2008 - 9e Journées d'Analyse statistique des Données Textuelles*. Sous la dir. de Serge HEIDEN (ED.) T. Volume 2. Lyon, France : Presses Universitaires de Lyon, mar. 2008, p. 811–822. URL : https://hal.archives-ouvertes.fr/hal-00551300 (visité le 26/11/2018).

MELLET, Sylvie. «Corpus et recherches linguistiques. Introduction». fr. In : *Corpus* 1 (nov. 2002). ISSN : 1638-9808. URL : http://journals.openedition.org/corpus/7 (visité le 18/09/2018).

MOURIK, Maaike S M van et al. «Accuracy of administrative data for surveillance of healthcare-associated infections : a systematic review». en. In : *BMJ Open* 5.8 (août 2015), e008424. ISSN : 2044-6055, 2044-6055. DOI : 10.1136/bmjopen-2015-008424. URL : http://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2015-008424 (visité le 18/09/2018).

MULLER, Charles. «La statistique lexicale». In : *Langue française* 2.1 (1969), p. 30–43. DOI : 10.3406/lfr.1969.5419.

NURMINEN, Tiia. «Le substantif épithète dans les désignations de produits cosmétiques». Mém.de mast. Université de Tampere, 2010.

OGER, Claire et Caroline OLLIVIER-YANIV. «Analyse du discours institutionnel et sociologie compréhensive : vers une anthropologie des discours institutionnels». In : *Mots* 71 (mar. 2003), p. 125–145. DOI : 10.4000/mots.8423.

PAVEAU, M.-A. «”Tweet”, Technologies discursives, [Carnet de recherche]». In : *https ://www.liens-socio.org/Le-profane-en-politique* (2012). https ://technodiscours.hypotheses.org/385, p. 385.

PINCEMIN, Bénédicte. «Sémantique interprétative et analyses automatiques des textes : qui devient les sèmes ?» In : *Texto !* (1999).

– «Sémantique interprétative et textométrie». In : *Texto !* XVII.3 (2012).

– «Sémantique interprétative et textométrie - Version abrégée». fr. In : *Corpus* 10 (nov. 2011), p. 259–269. ISSN : 1638-9808. URL : http://journals.openedition.org/corpus/2121 (visité le 27/10/2018).

PINCEMIN, Bénédicte et Serge HEIDEN. «Qu'est-ce que la textométrie ? Présentation». Site du projet Textométrie , http ://textometrie.ens-lyon.fr/spip.php ?rubrique80. 2008. URL : http://textometrie.ens-lyon.fr/spip.php?rubrique80.

PLEAU, Joannie. «Le texte à l'ère du numérique : Analyse du concept de genre numérique». In : *Canadian Journal for New Scholars in Education/ Revue canadienne des jeunes chercheur(e)s en éducation* 8 (jan. 2017), p. 144–149.

PORTILLO Serrano, Verónica. «LA NOTION DE GENRE EN SCIENCES DU LANGAGE». fr. Thèse de doct. Université de Franche-Comté, 2016, p. 137.

POTTIER, Bernard. «Un mal-aimé de la sémiotique». In : *Exigences et perspectives de la sémiotique*. John Benjamins Publishing Company, 1985, p. 499–503. DOI : 10.1075/z.23.42pot.

POUDAT, Céline. «Étude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres». Thèse de doct. Université d'Orléans, 2006. URL : http://www.revue-texto.net/Corpus/Publications/Poudat/Etude.html (visité le 11/11/2018).

RASTIER, François. «Eléments de théorie des genres». In : *Texto !* (2001). URL : http://www.revue-texto.net/Inedits/Rastier/Rastier_Elements.html (visité le 11/11/2018).

– «Enjeux épistémologiques de la linguistique de corpus». In : *Texto !* (2004). URL : https://www.researchgate.net/publication/280816977_Enjeux_epistemologiques_de_la_linguistique_de_corpus (visité le 22/03/2019).
– «Formes sémantiques et textualité». In : *Langages* 40.163 (2006), p. 99–114. DOI : 10.3406/lgge.2006.2686.
– «La macrosémantique (1)». fr. In : *Texto !* (2002), p. 16.
– *La mesure et le grain Sémantique de corpus*. HONORE CHAMPION, 2011. ISBN : 9782745322302.
– «La sémantique des textes : concepts et applications». In : *Texto !* (1996).
– «La sémantique des thèmes - ou le voyage sentimental.» In : *Texo !* (1995). URL : http://www.revue-texto.net/Inedits/Rastier/Rastier_Themes.html.
– «L'isotopie sémantique, du mot au texte». In : *L'Information Grammaticale* 27.1 (1985), p. 33–36. DOI : 10.3406/igram.1985.2168.
– «Mésosémantique et syntaxe». In : *Texto !* (2005).
– «Pour une sémantique des textes théoriques». In : *Texto !* 17 (2005), p. 151–180.

RASTIER, François. «Principes et conditions de la sémantique componentielle». In : *Exigences et perspectives de la sémiotique*. John Benjamins Publishing Company, 1985, p. 505–527. DOI : 10.1075/z.23.43ras.

– «Sémantique du web vs. Semantic Web ?» In : *Syntaxe et sémantique* 9.1 (2008), p. 15. DOI : 10.3917/ss.009.0015.

– *Sémantique Interprétative*. Paris, Presse universitaire de France, 1987.

– *Sémantique interprétative*. FR. T. 3e éd. Formes sémiotiques. Paris cedex 14 : Presses Universitaires de France, 2009. ISBN : 978-2-13-057495-8. URL : https://www.cairn.info/semantique-interpretative--9782130574958.htm.

– *Sens et textualité*. Hachette Paris, 1989.

– «Vers une linguistique des styles». In : *L'information Grammaticale* 89.1 (2001), p. 3–6. DOI : 10.3406/igram.2001.2707.

RASTIER, François, Marc CAVAZZA et Jacques ABEILLE. *SEMANTIQUE POUR L'ANALYSE. De la linguistique à l'informatique*. Dunod, 1994. ISBN : 978-2225845376.

RENAUD, Clément. «Les mèmes internet : dynamiques d'énonciations sur le réseau social chinois Sina Weibo». In : *Travaux de linguistique* 73.2 (2016), p. 27. DOI : 10.3917/tl.073.0027.

SALEM, André. «Segments répétés et analyse statistique des données textuelles». In : *Histoire & Mesure* 1.2 (1986), p. 5–28. DOI : 10.3406/hism.1986.1518.

SANTÉ, Organisation Mondiale de la. *Lignes directrices OMS relatives à la qualité de l'air : particules, ozone, dioxyde d'azote et dioxyde de soufre*. Organisation Mondiale de la Santé. Organisation Mondiale de la Santé. 2005.

SCHMIDT, Vivien A. «Discursive Institutionalism : The Explanatory Power of Ideas and Discourse». In : *Annual Review of Political Science* 11.1 (2008), p. 303–326. DOI : 10.1146/annurev.polisci.11.060606.135342.

SCOTTO, Lionel Apollonia, Luxardo GIANCARLO et Piet GRÉGORY. «Approche lexicométrique des controverses climatiques». In : *JADT 2014 : 12es Journées internationales d' Analyse statistique des Données Textuelles*. Juin 2014.

SIGNORINI, Alessio, Alberto Maria SEGRE et Philip M. POLGREEN. «The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic». In : *PLoS ONE* 6.5 (mai 2011). Sous la dir. d'Alison P. GALVANI, e19467. DOI : 10.1371/journal.pone.0019467.

Sullivan, Jonathan. «China's Weibo : Is faster different ?» In : *New Media & Society* 16.1 (fév. 2013), p. 24–37. doi : 10.1177/1461444812472966.

Therriault, Guy. «Le concept d'isotopie : un instrument sémantique pour l'analyse du discours criminologique». In : *Déviance et société* 7.2 (1983), p. 115–130. doi : 10.3406/ds.1983.1366.

Tournier, Maurice. «D'où viennent les fréquences de vocabulaire ? La lexicométrie et ses modèles». In : *Mots* 1.1 (1980), p. 189–209. doi : 10.3406/mots.1980.1010.

Turenne, Nicolas. «Apprentissage d'un ensemble pré-structuré de concepts d'un domaine : l'outil GALEX». en. In : *Mathématiques et Sciences humaines* 148 (1999), p. 41–71. url : http://www.numdam.org/item/MSH_1999__148__41_0/ (visité le 21/10/2019).

Tvardik, Nastassia et al. «Accuracy of using natural language processing methods for identifying healthcare-associated infections». In : *International Journal of Medical Informatics* 117 (sept. 2018), p. 96–102. issn : 1386-5056. doi : 10.1016/j.ijmedinf.2018.06.002. url : http://www.sciencedirect.com/science/article/pii/S1386505618304362 (visité le 18/09/2018).

Valette, Mathieu. «Approche textuelle du lexique». In : *HDR INALCO* (2009).

– «SÉMANTIQUE INTERPRÉTATIVE APPLIQUÉE À LA DÉTECTION AUTOMATIQUE DE DOCUMENTS RACISTES ET XÉNOPHOBES SUR INTERNET». fr. In : *Le poids des mots, Actes des 7èmes Journées internationales d' Analyse statistique des Données Textuelles (JADT), 10-12 mars 2004, Louvain-la-Neuve (Belgique), G. Purnelle, C. Fairon, A. Dister, éds., UCL-Presses Universitaires de Louvain, 2004, p. 1106-1116.* Louvain-la-Neuve (Belgique), G. Purnelle, C. Fairon, A. Dister, éds., UCL-Presses Universitaires de Louvain, 2004, p. 1106-1116. (2004).

Wang, Shiliang, Michael J. Paul et Mark Dredze. «Exploring Health Topics in Chinese Social Media : An Analysis of Sina Weibo». In : *AAAI 2014*. 2014.

Wong, Coby S.C. et al. «Sources and trends of environmental mercury emissions in Asia». In : *Science of The Total Environment* 368.2-3 (2006), p. 649–662. doi : 10.1016/j.scitotenv.2005.11.024.

WU, Li-Chi. «APPROCHE TEXTOMÉTRIQUE DE L'ANALTYSE D'OPINIONS (L'exemple de la crise entre la Chine et Google 2010)». Thèse de doct. UNIVERSITÉ SORBONNE NOUVELLE - Paris 3, oct. 2016.

YANG, F. et al. «Characteristics of PM2.5 speciation in representative megacities and across China». en. In : *Atmospheric Chemistry and Physics* 11.11 (juin 2011), p. 5207–5219. ISSN : 1680-7324. DOI : 10.5194/acp-11-5207-2011. URL : https://www.atmos-chem-phys.net/11/5207/2011/ (visité le 18/09/2018).

YANG, Qinghua (Candy), Fan YANG et Chun ZHOU. «What health-related information flows through you every day ? A content analysis of microblog messages on air pollution». In : *Health Education* 115.5 (juil. 2015), p. 438–454. ISSN : 0965-4283. DOI : 10.1108/HE-05-2014-0066. URL : https://www.emeraldinsight.com/doi/abs/10.1108/HE-05-2014-0066 (visité le 26/09/2018).

YU, Louis Lei, Sitaram ASUR et Bernardo A. HUBERMAN. «Trend Dynamics and Attention in Chinese Social Media». In : *American Behavioral Scientist* 59.9 (avr. 2015), p. 1142–1156. DOI : 10.1177/0002764215580619.

ZHANG, Shuqing et al. «微博文本和传统文本体裁特征对比 (Comparison of Genre Features Between Micro§blog Text and Traditional Texte)». In : *Journal of University of South Chine(Science and Technology)* (2015).

ZHEFEI FANG Hongfei Lin, Zhihao Yang. «中文文本体裁的自动分类机制 (Automatic Classification of Chinese texte Genre)». In : 中文信息学报 20.2, 26 (2006), p. 26. URL : http://jcip.cipsc.org.cn/CN/abstract/article_1977.shtml.

Table des figures

Liste des tableaux

Table des matières

Qinran DANG

Brouillard de pollution en Chine. Analyse sémantique différentielle de corpus institutionnels, médiatiques et de microblogues

Résumé

Au fur et à mesure de la dégradation de la qualité de l'air en Chine, de plus en plus d'articles journalistiques et de microblogues (*weibo* en chinois, équivalent de *tweet*), provenant de sites web gouvernementaux, médiatiques, de réseaux sociaux, de forums ou de blogs, traitent le problème du « 雾霾 » (*wumai* en chinois, pour désigner le brouillard de pollution) en Chine sous plusieurs angles : politique, écologique, économique, sociologique, sanitaire, etc. La sémantique des thèmes abordés dans ces textes diffère sensiblement en fonction de leur genre textuel. Dans cette thèse, nous avons pour objectif d'une part, de relever les différents thèmes d'un corpus numérique traitant du *wumai* et spécifiquement construit à cette fin, et d'autre part, d'interpréter de façon différentielle la sémantique de ces thèmes. Dans un premier temps, nous collectons les données textuelles en langue chinoise relatives au *wumai*. Ces textes provenant de trois sites web chinois traditionnels et du réseau social sont divisés en quatre genres textuels. Après une série de traitements préparatoires : nettoyage, segmentation, normalisation, annotation, balisage et organisation, nous étudions les caractéristiques des quatre genres textuels du corpus à partir d'une série de variables discriminantes - hyperstructurelles, lexicales, sémiotiques, rhétoriques, modales et syntaxiques - réparties au niveau infratextuel et intratextuel. Ensuite, en nous basant sur les caractéristiques de chaque genre textuel, nous relevons les thèmes principaux exposés dans chaque genre de sous-corpus, et analysons de manière contrastive la sémantique de ces thèmes récupérés. Les résultats d'étude sont interprétés de manière quantitative et qualitative. Les analyses quantitatives s'effectuent à l'aide d'outils textométriques, les interprétations sémantiques s'inscrivent dans le cadre théorique de la sémantique interprétative (SI) proposée par Rastier (1987).

Mots-clés : Humanité numérique, fouille de texte, textométrie, analyses sémantique, genre textuel, analyses des réseaux sociaux, analyses du discours institutionnel, traitement automatique du chinois, corpus écologique

Abstract

Air pollution has increasingly become a serious problem in China, more and more journalistic articles and miniblogs (*weibo* in Chinese, equivalent to *tweet*), comming from governmental or media websites, social networks, blogs and forums, etc., discuss the issue of «雾 霾» (*wumai* in Chinese, means *smog*) in China through several angles : political, ecological, economic, sociological, health, etc. The semantics of the themes adressed in these texts differ significantly from each other according to their textual genre. In the framework of our research, our objectif is double-fold : on the one hand, to identify different themes of a digital propose-bulit corpus relating to *wumai* ; and on the other hand, to interpret differentially the semantics of these themes. Firstly, we collect the textual data written in chinese and related to *wumai*. These journalistic articles and *weibo* deriving from three traditional chinese and the social network are divided into four genres of sub-corpus. Secondly, we constitute our corpus through a series of data processing : data cleaning, word segmentation, normalization, POS tagging, benchmarking and data organization. We study the characteristics of the four genres of sub-corpus through a series of discriminating variables - hyperstructural, lexical, semiotic, rhetorical, modal and syntactic - distributed at the infratextual and intratextual level. After that, based on the characteristics of each textual genre, we identify the main themes exposed in each genre of sub-corpus, and analyze the semantics of these identified themes in a contrastive way. Our analysis results are interpreted from two angles : quantitative and qualitative. All statistical analysis are assisted by textometric tools ; and the semantic interpretations are implemented on several fundamental concepts of SI (Sémantique interprétative) proposed by Rastier (1987).

Keywords : Digital humanity, text mining, textometrie, semantic analysis of corpus, textual genre, social network analysis, institutionnal discours analysis, chinese language processing, ecological corpus