Mots. Les langages du politique

93 | 2010
Figures et filiations dans le discours politique latino-américain

Variété et distribution des sujets d'actualité sur Internet. Une analyse quantitative de l'information en ligne

Emmanuel Marty, Franck Rebillard, Nikos Smyrnaios et Annelise Touboul

revues.org

Éditeur
ENS Éditions

Édition électronique
URL : http://mots.revues.org/19832
DOI : 10.4000/mots.19832
ISSN : 1960-6001

Édition imprimée
Date de publication : 1 juillet 2010
Pagination : 107-126
ISBN : 978-2-84788-231-5
ISSN : 0243-6450

Référence électronique
Emmanuel Marty, Franck Rebillard, Nikos Smyrnaios et Annelise Touboul, « Variété et distribution des sujets d'actualité sur Internet. Une analyse quantitative de l'information en ligne », *Mots. Les langages du politique* [En ligne], 93 | 2010, mis en ligne le 01 octobre 2012, consulté le 30 septembre 2016. URL : http://mots.revues.org/19832 ; DOI : 10.4000/mots.19832

Ce document est un fac-similé de l'édition imprimée.

Emmanuel Marty, Franck Rebillard, Nikos Smyrnaios, Annelise Touboul

Variété et distribution des sujets d'actualité sur Internet. Une analyse quantitative de l'information en ligne

La question du pluralisme de l'information est essentielle dans les sociétés démocratiques dès lors que l'on considère les médias comme des moyens d'expression et d'échange entre diverses opinions, autant que comme des outils d'appréhension de la vie sociale. Afin d'assurer ce pluralisme, les pouvoirs publics assurent une régulation du secteur de la presse écrite (exemple : aides de l'État pour les quotidiens nationaux d'information politique et générale à faibles ressources publicitaires) et de l'audiovisuel (exemple : contrôle du temps de parole des différentes formations politiques par le Conseil supérieur de l'audiovisuel). Pour Internet en revanche, aucune mesure d'envergure n'a été prise à ce jour en France : le pluralisme est supposé découler de la multiplicité des sources d'information disponibles, comme le laissent penser les récents rapports publics consacrés à ces questions (Lancelot, 2005 ; Tessier, 2007). Puisque l'offre de contenus d'actualité sur la toile émane d'une diversité d'acteurs – appartenant au secteur professionnel du journalisme ou non (blogueurs, rédacteurs « citoyens »), intégrés à l'équipe rédactionnelle d'un média existant par ailleurs (journal, radio, télévision) ou exclusivement numérique (*Internet pure player*) –, il en est automatiquement déduit une substantielle pluralité des contenus offerts à l'internaute. Or, une étude plus approfondie fait apparaitre une piste d'investigation inverse : la multiplicité des espaces de publication sur Internet pourrait favoriser le pluralisme autant qu'une certaine redondance des informations en circulation. En effet, nombre d'articles publiés sur des sites de presse en ligne, fondés dans bien des cas sur des dépêches d'agence, se voient par exemple compilés par des agrégateurs ou commentés sur des blogs (Rebillard, 2006).

Université Toulouse 3, Lerass
emmanuel.marty@iut-tlse3.fr
Université de Lyon, Université Lumière Lyon 2, Elico
Franck.Rebillard@univ-lyon2.fr
Université Toulouse 3, Lerass
smyrnaios@gmail.com
Université de Lyon, Université Lumière Lyon 2, Elico
Annelise.Touboul@univ-lyon2.fr

Nous rendons compte ici de certains résultats obtenus dans le cadre d'un projet interdisciplinaire nommé « Internet, pluralisme et redondance de l'information »[1], qui visait à valider ou infirmer empiriquement ces hypothèses en prenant appui sur des méthodes quantitatives, ce qui était jusqu'ici inédit.

Le pluralisme de l'information peut être abordé de plusieurs manières : par exemple du point de vue de la concentration dans les médias (Miège, 2005), des lignes éditoriales mises en œuvre (Czepek *et al.*, 2009) ou encore des opinions relayées par les journalistes (Anderson *et al.*, 2005). Dans le cadre de cette recherche, nous nous sommes concentrés sur deux paramètres : d'une part, la diversité des sujets traités par les médias et d'autre part, la diversité lexicale de traitement de ces sujets. Il va de soi qu'il ne s'agit là que d'une mesure partielle du pluralisme de l'information. En effet, un même sujet peut être traité de façon pluraliste, et des sujets d'une grande diversité peuvent être traités selon un point de vue identique (par exemple un positionnement politique fort). Cependant, les paramètres concernés par notre étude renseignent sur le degré de puissance de l'effet d'agenda dans le domaine de l'information en ligne. Comme cela a été montré à de multiples reprises, les médias tendent à privilégier les mêmes sujets d'actualité au même moment et en parlent souvent en des termes similaires (Dearing, Rogers, 1992). Cette caractéristique forte du système médiatique aboutit à ce que des questions sociales, politiques, économiques ou culturelles soient particulièrement mises en avant dans la sphère publique, ou qu'elles soient à l'inverse occultées. Notre étude s'efforce d'évaluer ce phénomène sur Internet. De ce point de vue, il s'agit bien d'une mesure, partielle, du pluralisme de l'information.

Sur la base d'un corpus d'une soixantaine de sites représentatifs des différentes catégories d'espaces de publication en ligne, une analyse de contenu semi-automatisée de leurs titres[2] est d'abord déployée. Elle vise à saisir la quantité de sujets abordés au cours d'une journée, et surtout l'importance relative de chacun d'entre eux au sein de l'agenda médiatique numérique.

Après avoir établi ce panorama général de la production d'actualités en ligne, l'analyse lexicométrique des titres sera affinée dans un second temps afin de fournir des indicateurs linguistiques tangibles du positionnement éditorial propre à chaque source, via l'identification d'une production originale de

1. Programme de recherche IPRI (Internet, pluralisme et redondance de l'information) soutenu en 2008 par la Maison des sciences de l'homme Paris-Nord et réunissant des chercheurs en information-communication et informatique des laboratoires CRAPE (Université Rennes 1), ELICO (Université de Lyon), GRICIS (UQAM Montréal), LERASS (Université Toulouse 3), LIRIS (INSA Lyon).
2. Pour lever toute ambigüité terminologique, précisons que l'emploi du mot *titre* fera ici toujours référence à la titraille de premier niveau, qui constitue notre matériau d'analyse, et non aux supports ni aux acteurs de cette information, qui seront désignés par les mots *sites* et *sources*. Une confusion est en effet possible avec la notion de titre de presse, communément employée, désignant la source du message.

contenu, ou au contraire de reprises de formules stéréotypées. Toutes ces opérations ont été réalisées sur la base d'échantillons de données dont la méthode de constitution mérite d'être présentée au préalable.

La méthode de constitution des données

S'intéresser au pluralisme de l'information sur Internet implique d'intégrer les particularités de la publication de contenus sur le web, en comparaison des supports antérieurs comme l'imprimé ou l'audiovisuel. Car le processus de publication sur le web, ainsi que l'a montré un programme de recherche précédent dédié de façon plus large aux contenus informationnels et culturels (Chartron, Rebillard, 2007), ne se limite pas au modèle classique de diffusion médiatique : il comprend aussi le registre de l'autopublication (ex : blogs), de la publication distribuée (ex : *peer to peer*), et le niveau méta-éditorial (ex : agrégateurs). Pour les contenus journalistiques, cela revient à considérer que la logique de l'*editorial content* voisine avec celle de la *public connectivity*, comme l'a exprimé Mark Deuze dans un article de référence sur le sujet (2003) où il distinguait les catégories de *mainstream news sites*, *index and category sites*, *meta and comment sites*, *share and discussion sites*.

Cette typologie est reprise dans de nombreux travaux anglophones qui se penchent sur le journalisme en ligne mais sans jamais vraiment l'analyser globalement. Y compris dans les travaux les plus récents qui, bien qu'ils doivent prendre acte de cette existence plurielle du journalisme en ligne, se limitent au mieux à analyser les interactions entre deux catégories d'espaces de publication d'informations d'actualité : par exemple entre les sites de médias traditionnels et leurs pourvoyeurs d'audience transnationaux que sont les agrégateurs (agrégateurs « manuels » comme le *Drudge Report* ou « automatisés » comme *Google News* ; voir Thurman, 2007) ; ou entre blogs amateurs et sites professionnels, malgré l'annonce d'une analyse de la « *global news arena* » (Reese *et al.*, 2008).

Délimitation du corpus

Dans le cadre de cette recherche, nous nous sommes efforcés de poursuivre le travail de défrichage initié par Deuze (2003) jusqu'à aboutir à une typologie complète et actualisée des différents espaces web de publication d'information d'actualité :

- sites d'organes de presse ;
- sites d'agences de presse ;
- publications individuelles exclusivement en ligne (blogs) ;
- publications collectives exclusivement en ligne (webzines) ;

– publications « collaboratives » exclusivement en ligne (sites de journalisme participatif) ;
– composantes informationnelles de plateformes multiservices (rubriques *Actualités* des portails) ;
– regroupements automatisés d'informations d'actualité (agrégateurs de nouvelles).

En accord avec la problématique de cette recherche, nous avons ainsi fait en sorte que le corpus retenu contienne des sites représentant tout l'éventail des catégories énumérées ci-dessus. Une liste d'une centaine de sites d'information d'actualité – sites français et francophones afin de tenir compte de la dimension internationale d'Internet et avec au moins cinq sites par catégorie – a été établie à partir des répertoires spécialisés suivants : bases de Google Actualités, de Wikio, de Rezo.net et du site professionnel IPLJ (« Internet pour les journalistes »).

Extraction du corpus

De cette première liste d'une centaine de sites, on n'a pu finalement exploiter qu'une soixantaine, ceux qui présentaient un flux RSS « À la Une » ou « Actualités », afin de satisfaire au procédé informatique d'extraction des titres en continu. Cette application informatique, développée spécifiquement pour les besoins de la présente recherche, repose en effet sur une « aspiration » (*crawling*), en temps réel, des titres publiés sur les fils RSS des différents sites du corpus.

Elle nous a permis de collecter des données d'une ampleur rarement atteinte dans les analyses de contenu des médias : en fin de compte, c'est la production quotidienne d'environ soixante sources d'information qui a été rendue disponible pour l'observation, à comparer avec la taille standard des corpus ordinairement mobilisés dans les études sur les seuls grands quotidiens nationaux ou sur les seules chaines nationales hertziennes, dépassant rarement la dizaine de sources.

Ce procédé aura toutefois également montré une limite, par rapport aux deux catégories de sites n'ayant pas (ou peu) de flux RSS : les agences de presse et les agrégateurs. Pour la première catégorie, cet obstacle a pu être anticipé et contourné en relevant « à la main », pour chaque journée d'observation, les titres des dépêches publiées dans les catégories « À la Une » et « Actualités » des sites de l'AFP, Reuters, Xinhuan et La Presse canadienne. Pour la dernière catégorie, celles des agrégateurs, un problème technique d'enregistrement des flux a malheureusement révélé a posteriori la mise à l'écart involontaire d'un site majeur comme Google News, alors que son alter ego Wikio a, lui, vu ses flux correctement captés par notre application informatique.

Malgré ces quelques limites, nous estimons avoir pu travailler sur un corpus final satisfaisant par sa taille et sa globalité, puisqu'il rassemble plusieurs

milliers d'articles publiés quotidiennement par environ 60 sources d'informations, couvrant toutes les catégories de sites.

Structure des échantillons et production des sites

L'analyse a porté sur deux journées de novembre 2008, constituant autant d'échantillons de données, composés comme suit :

– échantillon du 6 novembre : 2 617 articles issus de 61 sources – échantillon du 10 novembre : 2 040 articles issus de 60 sources[3].

Le début du mois de novembre 2008 coïncidait avec un évènement politique – et médiatique – majeur : l'élection de Barack Obama à la présidence des États-Unis d'Amérique. Le scrutin a eu lieu le 4 novembre et ses résultats n'ont été connus qu'à partir du 5 novembre en France : le fait de s'attacher à deux échantillons, l'un proche temporellement de cet évènement et l'autre plus éloigné, visait à analyser autant une journée d'actualité « chaude » qu'une journée plus ordinaire sur le plan journalistique.

La production d'articles est très variable pour chacun des sites. Tout en haut de l'échelle, les agrégateurs et portails tels que Wikio, MSN et Canoë (portail québécois) diffusent quotidiennement jusqu'à plusieurs centaines d'articles. Tout en bas de l'échelle apparaissent les blogs, dont la production est quantitativement très faible. Entre ces deux extrémités figure tout un éventail de sites de presse en ligne, d'agences, de webzines et de sites participatifs dont la production quotidienne oscille entre 10 et 100 articles. Les écarts de rythme de production sont donc importants selon les catégories de site, et le poids d'un ou deux acteurs sur l'ensemble de l'échantillon s'avère extrêmement important (à eux deux, MSN et Wikio pèsent à peu près 20 % de l'échantillon).

Classification et indexation des articles

La classification des articles en sujets est une étape déterminante vis-à-vis des résultats obtenus. Pour répondre à l'objectif de mesure quantitative du pluralisme de l'information sur Internet, les articles collectés ont en effet été regroupés par sujets, sur la base de leurs titres respectifs. La méthode employée pour parvenir à ces classifications est semi-automatisée et inductive. Fondée au départ sur une analyse lexicométrique (automatisée) des titres qui permet d'opérer un premier défrichage, cette méthode repose conjointement sur une identification par les chercheurs des sujets constituant l'actualité du jour au sein de notre corpus.

Par *sujet*, nous désignons une expérience factualisée, telle que définie

3. Le fait que la journée du 10 novembre compte 60 sources et non plus 61 est lié à la méthode d'aspiration utilisée, ne retenant que les sources qui avaient publié de manière effective dans les 24 heures concernées.

par Jean-Pierre Esquenazi (2002) : la « factualisation de l'expérience » revient « dans le temps et l'espace, à délimiter, circonscrire le fait en le présentant comme une discontinuité ». En prenant appui, avec Esquenazi, sur la notion de « cadre » d'Erving Goffman (1991), on considère le sujet d'actualité comme une « expérience cadrée », résultant de l'application du cadre primaire de la perception. La rédaction de l'article de presse, en tant que production discursive médiatique, consiste à appliquer à cette réalité perçue et cadrée un cadre médiatique second, un « recadrage » (Esquenazi, *ibid*.). Notre conception du sujet n'est pas relative à ce cadre médiatique second mais au cadre primaire du sujet d'actualité en amont de son traitement médiatique, c'est-à-dire avant la situation de rédaction. En effet, comme le soulignent Roselyne Ringoot et Yvon Rochard (2005), le « rapport au terrain va déterminer la pratique professionnelle du journaliste en articulant la perception du réel en situation de recherche d'informations et le processus d'écriture en situation de rédaction ». Pour résumer notre démarche de regroupement des articles en sujets, nous nous sommes attachés à l'identification du sujet d'actualité avant qu'il ne devienne un sujet journalistique, auquel on a apposé « angles, format, genres, positionnement dans le journal » (*ibid*.). Par exemple, pour la journée du 6 novembre, l'élection de Barack Obama à la présidence des États-Unis a constitué pour nous un sujet d'actualité donnant lieu à plusieurs sujets journalistiques, tels que les déclarations d'hommes d'État sur l'élection, la portée symbolique de l'accession d'un candidat noir à la fonction suprême ou encore la future formation gouvernementale. C'est sur la base de cette définition, tentant d'objectiver les différents sujets d'actualité et d'aboutir à un niveau de classification homogène pour chacun d'entre eux, que ces derniers ont été progressivement inventoriés, à l'issue de plusieurs étapes.

L'étape automatisée, tout d'abord, a été réalisée à l'aide du logiciel d'analyse des données textuelles Lexico3 (Lebart, Salem, 1994)[4]. Une des fonctionnalités de ce logiciel est le comptage des segments répétés. Il s'agit de repérer les groupes de mots très fréquents dans un corpus. À cette étape de l'analyse, cet outil nous permet de comptabiliser les titres identiques (ou présentant une forte proximité lexicale) les plus présents au sein de chaque échantillon. Ce premier tri automatique ne permet pas simplement de regrouper les articles dont les titres sont similaires, mais aussi de faire émerger un certain nombre de sujets particulièrement récurrents.

Dans un second temps, sont pris en considération les titres plus originaux sur le plan lexical et donc non repérés par l'outil de lexicométrie. Ceux-ci sont analysés un à un par les chercheurs : certains sont rattachés aux sujets déjà mis en lumière dans l'étape précédente ; les autres titres conduisent à l'identification de nouveaux sujets, au sein desquels ils sont regroupés.

4. Le logiciel effectue tout d'abord un comptage des occurrences de formes ou groupes de formes similaires, puis construit un tableau des formes récurrentes avec leur fréquence d'apparition dans le corpus.

Dans un dernier temps, après constitution définitive des sujets, des codes sujets sont affectés à chacun d'eux selon une numérotation classique allant de 1 à n. Après traitement, chacune des deux bases (6 novembre et 10 novembre) se présente donc sous la forme d'un tableau qui comporte autant de lignes que d'items récoltés automatiquement pendant la journée en question, et trois colonnes : titre de l'article, code sujet, nom de la source. Ces bases de titres indexés sont alors utilisées pour les calculs de variété et d'équilibre des sujets.

Quantifier le pluralisme : variété et répartition des sujets abordés

Pour réaliser une mesure du pluralisme des sujets traités au sein de nos échantillons de données, nous avons appliqué au cas de l'information journalistique des formules déjà expérimentées pour calculer le degré de diversité d'autres types de contenus culturels et médiatiques, comme cela a été fait par exemple dans le secteur du livre (Benhamou, Peltier, 2006). Notre étude reprend donc des axes conceptuels déjà éprouvés, à travers les notions de variété et d'équilibre de l'information. La variété correspond au nombre de sujets au sein de l'échantillon considéré, l'équilibre est pour sa part déterminé par la répartition des sujets entre les différents articles[5].

Échantillon du 6 novembre 2008 : sur un total de 2 617 articles, 385 sujets différents ont été identifiés. On observe ainsi une moyenne de 7 articles par sujet, avec d'immenses écarts à la moyenne et une médiane proche de 1 article par sujet. Dans le détail, alors que les 5 sujets les plus fréquents regroupent chacun plus de 100 articles (voir tableau 1), 301 des 385 sujets identifiés regroupent moins de 5 articles chacun.

Sujet	Nombre d'articles	Pourcentage de l'échantillon
L'élection américaine de Barack Obama	435	16,62
Le congrès du PS et son actualité	187	7,15
Les institutions internationales face à la crise financière	178	6,80
L'actualité boursière	121	4,62
Grève à la SNCF	120	4,59

Tableau 1. Sujets les plus fréquemment abordés lors de la journée du 6 novembre 2008

5. Un troisième critère de la diversité est la disparité, correspondant pour notre problématique à la différence de traitement d'un même sujet entre plusieurs articles, que nous n'avons pas eu le temps d'aborder de façon suffisante dans le cadre de la présente recherche. Nous revenons sur cette perspective en conclusion, restant convaincus de la nécessaire complémentarité entre la démarche lexicométrique présentée ici et une approche plus qualitative d'analyse du discours journalistique.

Échantillon du 10 novembre 2008 : sur un total de 2 040 articles, 309 sujets différents ont été identifiés. On observe là encore une moyenne de 7 articles par sujet, et les écarts à la moyenne restent colossaux, bien qu'un peu plus lissés que dans l'échantillon du 6 novembre, en particulier parce qu'il n'y a plus de sujet extrêmement dominant comme pouvait l'être l'élection de Barack Obama. La médiane est ici aussi proche de 1 : en fait, 169 sujets sur 309 correspondent à un seul article, inclus dans un total de 236 sujets qui comptent moins de 5 articles chacun. Les 5 sujets les plus fréquents totalisent chacun plus de 80 articles. Ce nombre, bien qu'inférieur à celui du 6 novembre, reste néanmoins élevé.

Sujet	Nombre d'articles	Pourcentage de l'échantillon
La question de la gouvernance du PS après le congrès	129	6,32
Obama rencontre Bush à la Maison Blanche	114	5,59
Nouvel acte de malveillance contre la SNCF	112	5,49
Le prix Goncourt décerné à l'auteur franco-afghan Atiq Rahimi	108	5,29
La situation en République démocratique du Congo	83	4,07

Tableau 2. Sujets les plus fréquemment abordés lors de la journée du 10 novembre 2008

Les deux échantillons présentent des caractéristiques très voisines : une grande variété de sujets abordés, mais très inégalement représentés. Quelques sujets agglomèrent des quantités importantes d'articles (plusieurs centaines), tandis que la majorité des sujets ne connait de traitement que dans un seul article. La production de l'information telle que nous l'avons mesurée sur la toile affiche donc une structure relativement constante, partagée entre une concentration extrême des articles sur un nombre réduit de sujets et une dispersion très forte des articles restants sur presque autant de sujets isolés.

Une concentration extrême sur quelques sujets dominants

Malgré le poids médiatique exceptionnel pris le 6 novembre par le sujet de l'élection de Barack Obama (environ 17 % des sujets), en raison de la proximité temporelle avec le scrutin présidentiel aux États-Unis, on observe une structure très semblable de la distribution des sujets au sein des deux échantillons.

Les articles se concentrent dans les deux premiers déciles de sujets de l'échantillon du 6 novembre 2008 : 20 % des sujets rassemblent 82 % des articles (en sachant que les seuls 10 % des sujets les plus traités rassemblent 71 % des articles).

La concentration est quasi identique dans l'échantillon du 10 novembre 2008 : 20 % des sujets rassemblent 81 % des articles (en sachant que les seuls 10 % des sujets les plus traités rassemblent 69 % des articles).

Une forte dispersion en de multiples sujets isolés

La concentration est ainsi très forte sur les sujets les plus mis en avant sur la scène médiatique, fût-elle numérique. Parallèlement, à l'autre extrémité de la courbe, apparait un phénomène inverse de très forte dispersion. Mais si les sujets originaux repérés sur la toile sont nombreux, ils sont loin d'entrainer un engouement médiatique, en comparaison des sujets qui s'imposent à la une de l'actualité. Ainsi, pour l'échantillon du 6 novembre, 50 % des sujets agrègent 93 % des articles tandis que l'autre moitié n'est présente que dans 7 % des articles de l'échantillon.

Les chiffres sont quasiment identiques pour l'échantillon du 10 novembre : 50 % des sujets agrègent 92 % des articles tandis que l'autre moitié n'est présente que dans 8 % des articles de l'échantillon.

Une structure identique aux niveaux national et international

La très forte dispersion constatée pourrait s'expliquer par la délimitation du corpus présidant à l'échantillonnage. En choisissant de faire figurer des sites étrangers francophones pour conserver la dimension internationale d'Internet dans notre analyse, nous aurions pu créer un émiettement artificiel des sujets, via la collecte d'articles concernant un sujet très spécifique à un pays (ex : compétitions de hockey sur glace dans les sites québécois).

Une réduction de l'échantillon aux seuls sites français a donc été effectuée pour tester l'existence d'un tel biais. Elle confirme la permanence de la structure déjà observée, celle d'une grande variété de sujets inégalement répartis entre concentration extrême et forte dispersion. En effet, sur chacune des deux journées, les cinq premiers sujets traités sont les mêmes et les ordres de grandeur sont très comparables : la médiane reste à 1 et 20 % des sujets continuent à rassembler environ 80 % des articles.

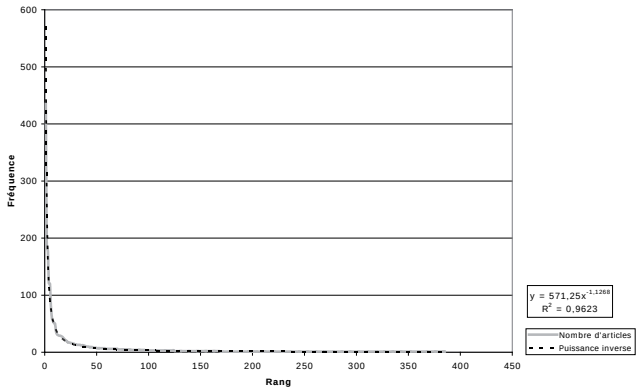
Cette première analyse quantitative a permis d'offrir, à propos d'échantillons de données nombreux (plus de 2 000 articles par jour collectés sur une soixantaine de sites), une représentation chiffrée de la production des informations d'actualité sur Internet. Celle-ci apparait donc à la fois comme variée (plus de 300 sujets) et comme très inégalement répartie : aux côtés de quelques sujets omniprésents, quantité de sujets n'émergent que par la grâce d'un article ou d'une poignée d'articles.

Un déséquilibre de l'information spécifique d'Internet ?

La distribution des sujets d'actualité sur Internet francophone, telle que nous avons pu l'observer lors de ces deux journées de novembre 2008, rappelle la règle de 20/80 dite de Pareto. La distribution *parétienne* caractérise plusieurs phénomènes sociaux et discursifs comme l'inégale répartition des ressources au sein d'une population (80 % des ressources sont alors détenues par environ 20 % des membres d'une population), ou encore l'inégale apparition des mots au sein d'un texte (80 % des occurrences sont dans ce cas rattachables à 20 % de l'ensemble des formes similaires présentes dans le texte).

Dans ce cadre, Internet ne présenterait alors pas une grande originalité en matière d'informations d'actualité : leur distribution présente une structure, entre concentration sur un nombre réduit de sujets et dispersion d'une multitude d'autres sujets, finalement assez habituelle. Afin de statuer plus précisément sur ce point, la distribution des sujets d'actualité observée sur Internet a été mise en parallèle avec des phénomènes plus anciens de distribution des contenus caractérisés, tout comme les distributions *parétiennes*, par des lois de puissance inverse : loi de Zipf en lexicométrie et loi de Lotka en scientométrie[6].

Une première comparaison a ainsi été effectuée avec la loi de Zipf. Celle-ci, établie depuis le milieu du 20e siècle, montre que la distribution des occurrences verbales au sein d'un texte est constante. Le schéma suivant représente la mise en correspondance de la distribution des sujets dans notre échan-



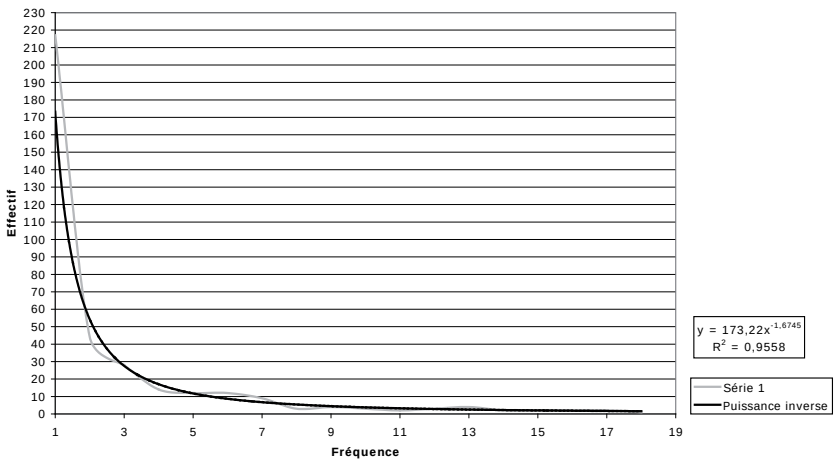
Graphique 1. Distribution des sujets par fréquence décroissante (6 novembre 2008). Comparaison avec la loi de Zipf

6. Nous remercions très vivement nos collègues spécialistes d'infométrie, Thierry Lafouge et Stéphanie Pouchot (Université Lyon 1, Elico), pour leur participation sur ce point précis à la recherche « Internet, pluralisme et redondance de l'information ».

tillon du 6 novembre 2008[7] avec la courbe théorique de Zipf. On peut remarquer une quasi-superposition entre les deux courbes.

Une deuxième comparaison a été menée avec la loi de Lotka, portant, elle, sur la distribution non pas d'items mais de classes d'items. Élaborée au début du 20e siècle, la loi de Lotka a mis en évidence, au sein d'une discipline scientifique donnée, de forts écarts entre un faible nombre d'auteurs de très nombreux articles et, à l'inverse, une grande majorité de chercheurs publiant en quantité limitée.

Cela nous amène, afin d'opérer une comparaison avec la production d'informations d'actualité sur Internet, à regrouper les sujets d'actualité selon leur niveau de fréquence d'apparition. Pour l'échantillon du 6 novembre, de tels regroupements ont ainsi été opérés, depuis les 217 sujets traités isolément (dans un seul article) jusqu'aux sujets donnant lieu à plusieurs articles (en s'arrêtant au seuil de 18 articles, à partir duquel les niveaux de fréquence deviennent trop discontinus pour une représentation graphique satisfaisante). Dans le schéma qui suit, la courbe hyperbolique tirée de cette classification est mise en parallèle avec la courbe théorique de puissance inverse de Lotka.



Graphique 2. Distribution des sujets par effectifs de même fréquence (6 novembre 2008). Comparaison avec la loi de Lotka

Là encore ressort de cette comparaison graphique une relative homologie entre les deux courbes. On remarquera toutefois que, pour le niveau de fréquence le plus bas (sujets qui ne sont traités que dans un seul article), la courbe de

7. D'autres comparaisons ont été effectuées sur la base de l'autre échantillon (10 novembre 2008), tant avec la loi de Zipf qu'avec la loi de Lotka. Le format de cet article ne nous permet pas de présenter l'ensemble de ces comparaisons, mais elles ont donné des résultats très voisins pour les deux échantillons.

distribution des sujets dépasse assez nettement la courbe théorique de Lotka. Ce dernier constat laisse penser que les sujets « monotraités » sont particulièrement nombreux sur le web sans pour autant que les sujets faiblement traités dans leur ensemble n'échappent, eux, à une distribution relativement classique.

Les enseignements à tirer de ces deux comparaisons ne sont cependant pas à surinterpréter. En effet, les lois infométriques mobilisées pour ces comparaisons sont issues de terrains d'observation (la diversité lexicale d'un texte et la production de documents scientifiques) assez éloignés de la diffusion d'informations sur Internet. Ces comparaisons ont néanmoins le mérite de rappeler que la répartition inégalitaire des sujets d'actualité observée dans le cas présent recoupe des lois de distribution plus générales, et ne doit pas être considérée comme une spécificité d'Internet.

De ce point de vue, la prise de recul offerte par de telles comparaisons permet notamment de revenir sur des hypothèses en vogue, telles que celle de « longue traîne » (Anderson, 2006) qui présente Internet comme un dispositif de communication favorisant la diversité culturelle. Une telle hypothèse a commencé à être infirmée par des recherches portant sur les contenus cinématographiques ou littéraires sur le web (Benghozi, Benhamou, 2008). Notre propre recherche apporte une contribution supplémentaire à ce sujet, à propos des contenus journalistiques. Elle montre que l'originalité des informations déployées sur certains sites est contrebalancée par un fort mimétisme d'ensemble dans le choix des sujets. Cette dualité de l'espace médiatique numérique, entre blogs et portails par exemple, a déjà été relevée (pour une synthèse des études en la matière, voir Flichy, 2008). Nous proposons de l'analyser dans le détail à partir de notre propre observation.

Retour aux sources du pluralisme et de la redondance

En nous appuyant sur les notions de variété (nombre de sujets au sein de l'échantillon considéré) et d'équilibre (distribution des articles en sujets), nous avons obtenu une vision panoramique de l'offre d'informations d'actualité sur le web francophone. Nous manque cependant une analyse permettant d'intégrer un facteur essentiel : les particularités des différentes sources d'information. Après avoir mis en évidence et quantifié la variété et l'équilibre – ou plus exactement le déséquilibre – de l'information, il nous appartient à présent d'en distinguer les acteurs. L'enjeu est d'être en mesure de situer les différentes sources sur un continuum allant des acteurs de la redondance aux acteurs du pluralisme. Pour entrer dans le détail de la production propre à chaque source et voir dans quelle mesure chacune pèse sur l'ensemble de l'échantillon analysé, l'outil Lexico3 a vu ses fonctionnalités mobilisées au-delà du tri initial des différents titres en présence.

Sur chacune des deux journées, deux analyses ont été effectuées. La première concerne l'identification des segments extrêmement stéréotypés : le seuil était alors fixé à 10 mots consécutifs et plus avec une fréquence de 15 occurrences et plus. Cette étape nous a permis d'identifier les sujets intitulés de manière redondante. Dans un second temps, l'analyse de ventilation des formes nous a permis de trier les segments en fonction de leur source. Il s'agit alors de segments plus courts, mais encore relativement caractéristiques (7 mots et plus), avec une fréquence de 2 occurrences et plus, c'est-à-dire répétés au moins une fois.

L'analyse des segments répétés

Les segments répétés longs et à forte fréquence constituent un marqueur tangible de redondance de l'information, correspondant soit à la répétition volontaire d'un titre de la part d'une source, soit à une reprise (au moins partielle) par une deuxième source de la formulation du titre de la première. Ces segments apportent un complément d'analyse aux calculs de variété et d'équilibre des sujets, en appréhendant la problématique du pluralisme et de la redondance sous l'angle de la production originale de titres (ou au contraire de la reprise/répétition de leur formulation originelle) et non plus uniquement du point de vue du sujet traité. En d'autres termes, nous ne nous situons plus dans une approche relative aux questions d'agenda des médias en ligne, mais dans une démarche liée à la manipulation du langage à travers la question des formulations.

Dans nos résultats (voir tableaux 3 et 4), il n'est pas étonnant de retrouver dans chaque journée les sujets les plus traités. Mais les seuils choisis, 10 mots d'une part et 15 occurrences d'autre part, ont d'abord vocation à identifier, sur un plan purement lexical, les groupes de formes textuelles employées de manière récurrente, témoignant d'une production discursive routinisée.

Au vu de la multiplicité des sources de notre échantillon, une telle récurrence de segments caractéristiques (jusqu'à 46 occurrences le 10 novembre) pourrait surprendre. De tels résultats appellent une distinction des sources et une caractérisation de leurs productions textuelles, permettant d'affiner notre approche lexicométrique des titres de l'actualité en ligne. L'enjeu est d'identifier les acteurs du pluralisme et les acteurs de la redondance, toujours du point de vue lexical.

Répartition des segments répétés et des hapax dans les différentes sources

Le calcul des spécificités lexicales est une fonctionnalité du logiciel Lexico3. Il repose sur le modèle hypergéométrique de Lafon (1984), analogue au calcul

Segments	Nombre d'occurrences du segment
Japon des tissus cérébraux créés à partir de cellules souches	28
Obama le premier président noir élu face à des défis colossaux	27
perpétuité pour le beau père 30 ans pour la mère	23
Les marchés replongent malgré une baisse des taux en Europe	21
Enfant mort de coups perpétuité pour le beau père 30 ans	19
Obama met en place l'équipe chargée de préparer la transition	18
La BCE abaisse ses taux et garde des cartouches pour décembre	18
Ligue des champions Lyon en tête de son groupe la Juventus	17
Le FMI annonce pour 2009 la première récession dans les pays	15
Discrimination la Halde recommande de lutter via les programmes et manuels	15

Tableau 3. Segments de 10 mots minimum apparaissant au moins 15 fois (6 novembre 2008)

Segments	Nombre d'occurrences du segment
attentat à Bagdad fait 28 morts et des dizaines de blessés	46
Le Goncourt à l'Afghan Atiq Rahimi le Renaudot au Guinéen	33
SNCF nouvel acte de malveillance le parquet anti terroriste saisi	23
États-Unis Obama doit rencontrer Bush à la Maison Blanche	23
l'UE lance une opération navale historique contre les pirates	21
Gilles Simon bat Roger Federer son deuxième succès face au Suisse	19
les deux camps se font face dans l'Est sans affrontement	19
Le plan de relance chinois applaudi par des marchés en nette	18
Partenariat avec la Russie l'UE prête à reprendre les négociations	16
PS Ségolène Royal réunit ses représentants lundi après midi au Sénat	15
Le plan de relance chinois et le G20 offrent un bref	15

Tableau 4. Segments de 10 mots minimum apparaissant au moins 15 fois (10 novembre 2008)

du KHI² mais permettant d'améliorer ses approximations. Les spécificités sont dites « positives », « négatives » ou « nulles » : elles correspondent à la fréquence d'apparition d'une ou plusieurs formes lexicales dans une source, au regard à la fois de la taille de cette dernière dans le corpus et de la fréquence moyenne d'apparition de la ou des formes en question dans les différentes sources. L'indice de spécificité correspond alors à l'exposant de probabilité d'apparition des termes dans la partie du corpus. C'est pourquoi la fréquence spécifique de formes dans une source s'exprime en spécificité positive ou négative : cela signifie qu'une source peut, de manière statistiquement significative, suremployer ou sous-employer cette forme par rapport à un emploi également réparti. Lorsque le nombre de formes spécifiques (segments répétés ou hapax par exemple) employées par une source se situe dans la moyenne, la spécificité est alors nulle car non significative en termes de probabilité.

Les résultats sur lesquels nous nous appuyons[8] sont, d'une part, la répartition sur chacune des sources des segments répétés au moins une fois de longueur minimum de 7 mots, longueur qui nous a paru suffisante pour constituer une unité de sens relativement caractéristique et pour réduire le « bruit » inhérent à cette méthode (ex : « Les États baltes ne comprennent plus Moscou ») ; d'autre part, la répartition des hapax, mots qui n'apparaissent qu'une seule fois dans l'ensemble du corpus. Ils constituent donc un indice de richesse lexicale, à considérer toutefois avec prudence, étant entendu que la présence d'hapax est corrélée à la taille des parties[9]. La surreprésentation d'hapax apporte donc un complément d'information à mettre en regard avec la sous-représentation de segments, et inversement entre la sous-représentation d'hapax et la surreprésentation de segments.

Sur l'ensemble des deux journées, les sources apparaissant comme les plus redondantes en termes à la fois de choix de sujets et de reprise de formulations stéréotypées sont les trois portails MSN Actualités, Yahoo Actualités et Orange Actualités, ainsi que l'AFP. Ce résultat vient confirmer l'importance de l'espace occupé par ces sources et illustre leur politique de flux continu d'information, qui privilégie la réactivité plutôt que la créativité. Viennent ensuite de manière encore très significative[10] les sites de chaines françaises de télévision,

8. En termes statistiques, les spécificités sont considérées comme significatives à 0,05, qui indique une probabilité de 95 % pour que cette répartition ne soit pas due au hasard. Dans nos résultats, les niveaux de significativité vont de 2 à 51, traduisant une probabilité comprise entre 10^{2} et 10^{51}, soit bien au-delà du seuil de significativité minimum de 0,05.
9. Le calcul des spécificités sur Lexico3, en s'appuyant sur le modèle de Lafon (1984), doit théoriquement modérer les résultats en fonction de cette donnée, mais des disparités dues à de grands écarts de taille des parties peuvent toutefois subsister.
10. Les sources sont ici présentées par ordre décroissant de significativité, au regard de la répartition des segments répétés et des hapax sur l'ensemble des deux journées. Par exemple, le 6 novembre, MSN Actu présente un suremploi de segments répétés à un taux de spécificité de 10^{51} et une sous-représentation d'hapax à un taux identique. À l'inverse, Agoravox affiche une sous-représentation de segments répétés à un taux de 10^{3} et une surreprésentation d'hapax à 10^{20}.

particulièrement France 2 et France 3, très proches des modèles proposés par l'AFP (TF1 et France 24 apparaissant comme moins redondantes en comparaison) ainsi que la station de radio RTL.

Du côté de la diversité, traduite par une sous-représentation de segments et une surreprésentation d'hapax, les sources les plus remarquables sur les deux journées sont situées à l'étranger : il s'agit d'Afrique en ligne, Canoë (Canada) et Ria Novosti (Russie). Ce résultat peut s'expliquer par le fait que l'agenda médiatique français était dominant dans notre corpus. Viennent ensuite, parmi les sources situées en France, quatre types de médias qui se distinguent par leur production originale de titres : il s'agit, dans l'ordre, du site participatif Agoravox, du webzine d'opinion Backchich, de la version internet du magazine d'opinion *Politis*, d'un autre site participatif Le Post et de certains blogs. Viennent seulement ensuite les déclinaisons numériques de la presse plus traditionnelle avec *Le Journal du Dimanche*, *Le Point*, *Les Échos* et, dans une moindre mesure, par ordre décroissant, *Libération*, *L'Humanité* et *Le Monde*, qui restent dans une singularité lexicale assez affirmée.

Enfin, certaines sources apparaissent au carrefour de la diversité et de ce que l'on appelle la « banalité lexicale », c'est-à-dire « un recours massif aux mots les plus fréquents » (Marchand, 2008). Cette position est flagrante pour *Le Nouvel Observateur*, mais c'est également le cas dans une moindre mesure pour *20 Minutes*, *Métro*, RFI et RMC. Ces sources ne présentent pas de surreprésentation des segments répétés, ce qui témoigne d'une réelle activité de production de titres, mais ils ne présentent pas pour autant une surreprésentation d'hapax (ces derniers sont même sous-représentés pour la journée du 10 novembre), indiquant l'emploi d'un langage sinon pauvre, tout au moins très commun et employé également par les autres sources. À la lumière d'études de nature socio-économique préalables qui montrent une tendance à la recherche de productivité dans les rédactions numériques des groupes de presse (Estienne, 2007 ; Rebillard *et al.*, 2007), on peut alors émettre l'hypothèse d'un tropisme de certains sites pour le journalisme « assis » (*desk*), via la réécriture de dépêches ou d'articles produits par des tiers, au détriment d'un travail d'élaboration de contenus originaux, passant, lui, plutôt par la voie du journalisme « debout ».

Les contenus et leur environnement : une interaction continue

Des liens peuvent alors se dessiner entre la disparité des conditions de production du contenu, l'enjeu, la nature et enfin les attentes liées à la consommation de ce contenu, qui diffèrent sensiblement selon les types de source. L'hypothèse de l'existence de ces liens entre en résonance avec la notion de « contrat de communication médiatique » avancée par Charaudeau (1997). Réactivité et

exhaustivité pour l'AFP, les agrégateurs et les sites de médias audiovisuels ; créativité et retraitement réflexif ou critique pour les *pure-players* et les indépendants (professionnels ou non), voilà les deux pôles du *continuum* redondance-pluralisme. En ce qui concerne les blogs, leurs spécificités langagières et communicationnelles ont déjà été explicitées et interprétées. Elles semblent liées au rôle, joué par ceux-ci, de remédiation et de retraitement subjectivé de l'information première (Serfaty, 2006). Les autres types de source se situent entre ces deux pôles dans une position intermédiaire : équilibre plus ou moins précaire entre alignement sur la concurrence et stratégies de distinction pour la presse traditionnelle, procédés de retraitement rapide et minimal de l'information pour les gratuits et certaines radios, telles peuvent être les valeurs implicitement partagées par les producteurs et les consommateurs de l'information en ligne[11], partage cristallisé dans cette disparité de contenus.

Ces quelques remarques montrent l'intérêt d'une recherche combinant analyse de contenu des sites web et prise en compte des conditions socio-économiques de production et de consommation de l'information. Une telle perspective a déjà été amorcée au cours de la présente recherche collective, dont certains résultats ont été livrés ici. Elle sera déployée à une plus large échelle dans les années qui viennent[12]. Ceci permettra de dépasser certaines des limites pointées au cours de cet article. Plus précisément, l'évaluation du pluralisme de l'information pourra être affinée à trois niveaux.

Premièrement, la dimension comparative du niveau de pluralisme entre Internet et d'autres médias sera plus pleinement intégrée à la recherche. Au-delà d'une mise en parallèle avec des phénomènes de distribution des contenus plus anciens, une comparaison intéressante sur le plan à la fois scientifique et sociétal pourrait s'établir entre Internet et télévision, média dominant jusqu'ici.

Deuxièmement, afin de cerner de façon plus précise le degré d'originalité des informations d'actualité circulant sur Internet, l'analyse de contenu dépassera le niveau lexical des titres. L'analyse sera approfondie à un niveau plus sémiologique, en considérant l'ensemble du texte d'un article. Par exemple, l'implicite de telle ou telle formulation, qui de fait échappe à une analyse lexicométrique, pourra être pris en compte. Plus généralement, le point de vue se dégageant d'un article pourra être saisi. Ceci permettra d'ajouter, à l'analyse de la variété et de la distribution des sujets d'actualité, une analyse de la disparité des traitements journalistiques d'un même sujet. Ainsi identifiés, les

11. Cette correspondance entre une offre et une demande d'information minimale, courte et factuelle, a déjà été démontrée dans le cas de la presse écrite pour les quotidiens gratuits (Augey *et al.*, 2005)
12. Pour la période 2009-2012, le programme « Internet, pluralisme et redondance de l'information » bénéficie d'une aide de l'Agence nationale de la recherche portant la référence ANR-09-JCJC-0125-01.

positionnements éditoriaux exprimés dans chacun des articles offriront une représentation plus satisfaisante des multiples dimensions que recouvre le pluralisme de l'information. En effet, si une similarité de titre apparait comme un indice tangible de redondance lexicale, la diversité éditoriale est plus difficilement déductible d'une originalité de titre, principalement du fait de la stratégie de certaines sources que nous venons d'évoquer (notamment les gratuits, mais pas seulement), dont l'activité se concentre précisément sur la reprise des dépêches d'agence (principalement l'AFP) et la reformulation (syntaxique et/ou lexicale) quasi exclusive des titres. On peut ainsi envisager à l'avenir des analyses plus poussées, portant sur le texte intégral des articles, dans lesquelles certaines hypothèses sur les spécificités des différents types de sources pourraient être testées de manière plus solide et approfondie.

Troisièmement, notre mesure quantitative actuelle porte en elle une limite structurelle en s'appliquant à des échantillons restreints chacun à une journée. En effet, les résultats des calculs que nous effectuons produisent la représentation artificiellement figée d'une production de l'information dont les sujets seraient circonscrits à la journée observée. Or ces sujets sont bien évidemment l'objet d'une continuité temporelle. Cet artefact est la contrepartie de l'obtention d'une photographie du pluralisme de l'information à un jour J. Nos conclusions devront donc être éprouvées sur d'autres échantillons avec, notamment, une profondeur chronologique plus importante.

Mais au-delà de ses limites, traçant les perspectives des travaux à venir, la présente recherche est déjà parvenue à une première description de l'existence conjointe du pluralisme et de la redondance de l'information sur Internet, deux phénomènes qui ne peuvent plus être considérés comme exclusifs. Elle a également esquissé une première cartographie des types de sites qui peuplent l'espace de l'actualité en ligne. Cette exploration constitue un travail à poursuivre, dont l'enjeu est scientifique (identifier des constantes linguistiques liées aux différents supports) mais aussi éminemment politique (mettre à l'épreuve l'idéal démocratique d'Internet).

Références

Anderson Alison, Petersen Alan, David Matthew, 2005, « Communication or spin ? Source-media relations in science journalism », *Journalism. Critical Issues*, S. Allan éd., Buckingham, Open University Press, p. 188-198.

Anderson Chris, 2006, *The Long Tail. Why the Future of Business Is Selling Less of More*, New-York, Hyperion.

Augey Dominique, Lipani-Vaissade Marie-Christine, Ruellan Denis, Utard Jean-Michel, 2005, « Dis à qui tu te donnes… La presse quotidienne gratuite ou le *marketing* du don », *Le journalisme en invention. Nouvelles pratiques, nouveaux acteurs*, R. Ringoot, J.-M. Utard éd., Rennes, Presses universitaires de Rennes, p. 89-123.

Benghozi Pierre-Jean, Benhamou Françoise, 2008, « Longue traîne : levier numérique de la diversité culturelle ? », *Culture prospective*, n° 1, en ligne [URL : http://www2.culture.gouv.fr/deps/fr/traine.pdf], consulté le 30 mars 2010.

Benhamou Françoise, Peltier Stéphanie, 2006, « Une méthode multicritère d'évaluation de la diversité culturelle. Application à l'édition de livres en France », *Création et diversité au miroir des industries culturelles. Actes des journées d'économie de la culture*, X. Greffe éd., Paris, La Documentation française, p. 313-344.

Charaudeau Patrick, 1997, *Le discours d'information médiatique. La construction du miroir social*, Paris, Nathan / Institut national de l'audiovisuel (Médias-Recherches).

Chartron Ghislaine, Rebillard Franck, 2007, « La publication sur le web, entre filiations et innovations éditoriales », *La redocumentarisation du monde*, R.-T. Pedauque éd., Toulouse, Cépaduès, p. 201-213.

Czepek Andrea, Hellwig Melanie, Nowak Eva éd., 2009, *Press Freedom and Pluralism in Europe. Concepts and Conditions*, Bristol, Intellect Books.

Dearing James W., Rogers Everett M., 1992, *Communication Concepts 6. Agenda-setting*, Thousand Oaks, Sage.

Deuze Mark, 2003, « The web and its journalisms. Considering the consequences of different types of newsmedia online », *New Media and Society*, vol. v, n° 2, p. 203-230.

Esquenazi Jean-Pierre, 2002, *L'écriture de l'actualité. Pour une sociologie du discours médiatique*, Grenoble, Presses universitaires de Grenoble, p. 15-47.

Estienne Yannick, 2007, *Le journalisme après Internet*, Paris, L'Harmattan.

Flichy Patrice, 2008, « Internet et le débat démocratique », *Réseaux*, vol. xxvii, n° 150, p. 159-185.

Goffman Erving, 1991, *Les cadres de l'expérience*, Paris, Minuit.

Lafon Pierre, 1984, *Dépouillements et statistiques en lexicométrie*, Genève, Slatkine / Paris, Champion.

Lancelot Alain, 2005, *Les problèmes de concentration dans le domaine des médias*, Rapport pour le Premier ministre, Paris, La Documentation française, en ligne [URL : http://lesrapports.ladocumentationfrancaise.fr/BRP/064000035/0000.pdf], consulté le 30 mars 2010.

Lebart Ludovic, Salem André, 1994, *Statistique textuelle*, Paris, Dunod.

Marchand Pascal, 2008, « La déclaration de politique générale a-t-elle vécu ? », *Parlement(s). Revue d'histoire politique*, n° 10, p. 152-164.

Miège Bernard, 2005, « La concentration dans les industries de contenu (présentation de dossier) », *Réseaux*, n° 131, p. 9-14.

Rebillard Franck, 2006, « Du traitement de l'information à son retraitement. La publication de l'information journalistique sur Internet », *Réseaux*, n° 137, p. 29-68.

Rebillard Franck, Cabedoche Bertrand, Damian Béatrice, Smyrnaios Nikos, 2007, *Les mutations de la filière Presse et information. Note sectorielle pour le ministère de la Culture et de la Communication*, Programme « Diversité culturelle et mutations des industries de la culture, de l'information et de la communication », MSH Paris-Nord.

Reese Stephen D., Rutigliano Lou, Hyun Kideuk, Jeong Jaekwan, 2007, « Mapping the blogosphere. Professional and citizen-based media in the global news arena », *Journalism*, vol. viii, n° 3, p. 235-261.

Ringoot Roselyne, Rochard Yvon, 2005, « Proximité éditoriale. Normes et usages des genres journalistiques », *Mots. Les langages du politique*, n° 77, *Proximité*, p. 73-90.

Serfaty Viviane, 2006, « Les blogs et leurs usages politiques lors de la campagne présidentielle de 2004 aux États-Unis », *Mots. Les langages du politique*, n° 80, *La politique mise au net*, p. 25-35.

Tessier Marc, 2007, *La presse au défi du numérique*, Rapport pour le ministre de la Culture et de la Communication, en ligne [URL : http://www.culture.gouv.fr/culture/actualites/rapports/tessier/rapport-fev2007.pdf], consulté le 30 mars 2010.

Thurman Neil, 2007, « The globalization of journalism online. A transatlantic study of news websites and their international readers », *Journalism*, vol. viii, n° 3, p. 285-307.