Cahiers de praxématique

61 | 2013
Le discours radiophonique en pratiques

Quelques expériences de TAL sur le discours radiophonique : le cas de la revue de presse

Somme experiments in computational linguistics applied to radio press reviews

Agata Jackiewicz et Frédérik Bilhaut

Édition électronique
URL : http://journals.openedition.org/praxematique/1919
DOI : 10.4000/praxematique.1919
ISSN : 2111-5044

Éditeur
Presses universitaires de la Méditerranée

Édition imprimée
Date de publication : 12 décembre 2013
ISSN : 0765-4944

Référence électronique
Agata Jackiewicz et Frédérik Bilhaut, « Quelques expériences de TAL sur le discours radiophonique : le cas de la revue de presse », *Cahiers de praxématique* [En ligne], 61 | 2013, mis en ligne le 01 janvier 2016, consulté le 08 septembre 2020. URL : http://journals.openedition.org/praxematique/1919 ; DOI : https://doi.org/10.4000/praxematique.1919

Quelques expériences de TAL sur le discours radiophonique : le cas de la revue de presse

Somme experiments in computational linguistics applied to radio press reviews

Agata Jackiewicz et Frédérik Bilhaut

1. Introduction

1 Notre contribution a pour objet de présenter une série d'expériences informatiques réalisées sur les discours de revue de presse de France Inter. Il s'agit de montrer comment exploiter des savoirs linguistiques et des techniques du TAL, pour explorer automatiquement une conséquente archive de discours de revue de presse. Ce corpus est regardé avec les yeux d'informaticiens linguistes, sans connaissance des pratiques journalistiques ayant procédé à sa production et à sa diffusion. Les traitements appliqués emploient des techniques existantes, visant l'identification et l'extraction de contenus textuels sémantiquement caractérisés.

2 À notre connaissance, c'est une étude originale. Nos références apportent des éclairages sur la technique de la revue de presse (Claquin 1993), concernent des caractéristiques générales des discours médiatiques (Charaudeau 2010, 2011), ou s'intéressent à certains de leurs aspects, comme le rôle des émotions (Tetu 2004), (Wirth and Schramm 2005)[1].

3 Les propositions qui seront présentées sont purement exploratoires. A l'évidence, elles sont déconnectées des besoins réels.

4 Le corpus, composé de 716 textes (mai 2005 - juin 2011), provient d'une archive en libre accès sur le site de France Inter. Ces textes, rédigés pour être dits, ne sont pas une transcription du rendu final. Formellement simplifiés (abréviations, absence d'accents, de majuscules...), ils comportent de nombreux syntagmes averbaux, différentes traces de mise en valeur, des marques de pause. Fonctionnellement, il s'agit de discours

rapportés (ou en circulation), issus de la presse ou du Web, brièvement commentés, souvent articulés entre eux et parfois mis en scène de manière globale (grâce au recours à la métaphore filée notamment).

5 Le travail dont nous rendrons compte dans cette communication s'est déroulé en trois temps : (i) lecture et exploration manuelle d'un sous-ensemble de documents ; (ii) construction d'une grille d'analyse et création d'un ensemble de ressources linguistiques (lexiques et patrons) ; (iii) expérimentations informatiques et analyse des résultats.

6 La grille d'analyse proposée articule plusieurs dimensions. Elle distingue au premier niveau : (i) sources d'information et contenu informationnel à proprement parler ; (ii) contenus factuels et contenus subjectifs. Questionner les sources, c'est connaître le support de l'information, le média et sa catégorie, l'auteur... Interroger le contenu rapporté, c'est identifier les sujets traités, ce qui en est dit, de quelle manière et par qui (thèmes, questions, assertions, entités nommées...). Ciblant les sources, la subjectivité des chroniqueurs touche notamment à la valeur de l'information et à la qualité du travail des confrères. Face aux faits sélectionnés et relatés, les commentaires dévoilent une large gamme d'attitudes, notamment émotionnelles (enthousiasme, soulagement, inquiétude, étonnement...).

7 Les méthodes de TAL mises en œuvre visent à générer automatiquement des indices relatifs à cette grille d'analyse, puis à les représenter dans une base de connaissance que nous utilisons *in fine* pour formuler des requêtes d'observation et produire des analyses statistiques. Les modules d'analyse relèvent principalement de l'extraction d'entités nommées, de l'analyse de la tonalité, du discours rapporté, et de la projection de diverses ressources spécifiques à l'étude (principalement lexiques et grammaires contextuelles). Le processus de traitement est mis en œuvre au sein de la plate-forme SemioLabs (développée par la société Noopsis sur la base de (Widlöcher & Bilhaut 2008), et les informations sont représentées au sein d'un triple-store RDF.

2. Apports du TAL à l'analyse des discours radiophoniques

8 L'introduction au présent volume (Fauré, XXXX) fait état d'un important champ d'études, à la fois vaste et spécialisé, autour des discours radiophoniques. L'expertise accumulée fait entrevoir divers axes d'approche, tout en montrant l'étendue de qui reste à explorer, étant données l'évolution des pratiques et la densification des liens entre les différents types de médias.

9 Pour notre part, n'étant ni analystes des médias ni journalistes, nous n'avons pas l'ambition de questionner dans ce travail les pratiques médiatiques ni les productions qui en émanent. Notre objectif est seulement de montrer qu'il est possible d'explorer un corpus de discours de revue de presse en exploitant le savoir-faire propre à nos disciplines d'origine qui sont la linguistique et le traitement automatique des langues. Nous constatons en effet que les professionnels des médias ne sont généralement pas au fait des possibilités offertes par l'ingénierie linguistique, et n'ont donc pas d'attentes précises vis-à-vis de ce domaine. Au contraire, certains d'entre eux s'estiment entièrement satisfaits des systèmes de recherche dits « plein texte », fonctionnant par mots-clés, car retrouver des contenus qui parlent d'une personne (Christiane Taubira),

d'un événement (élection présidentielle, États-Unis, 2012) ou d'une série d'événements (élection présidentielle, États-Unis) paraît réalisable de façon assez satisfaisante via un choix de mots-clefs appropriés.

10 Pourtant ces outils sont largement perfectibles, et surtout, au-delà des besoins de recherche d'information, d'autres finalités peuvent être envisagées afin de mener des explorations plus ouvertes et observer des tendances globales. Par exemple : identifier automatiquement la teneur émotionnelle dominante d'un ensemble de discours, ou encore répondre à des questionnements croisés qui impliquent à la fois propriétés formelles des discours et éléments de leur contenu. Ceci ne peut se faire sans disposer de moyens linguistiques et logiciels adéquats, car aucune requête à base de mots clefs ne permettra d'aboutir à ce résultat.

11 Les approches de type TAL peuvent en revanche répondre à de tels besoins en déployant de l'ingénierie et des ressources linguistiques associées à une expertise sur la langue et le discours, et beaucoup pourraient apporter des connaissances et des techniques applicables aux revues de presse. On pense en premier lieux aux travaux sur la construction automatique de résumé de textes, basés sur l'exploitation de différents types d'énoncés fonctionnels : annonces thématiques, énoncés conclusifs, reformulations, soulignements d'importance (Minel et al. 2001). On peut également citer les études sur l'identification et la catégorisation automatiques des discours rapportés et en circulation (Quintin et al. 2011). Et, plus récemment, les recherches dans le domaine de la fouille d'opinions et d'analyse de sentiments, qui ont ouvert des perspectives innovantes pour aborder la dimension subjective des discours (Pang and Lee, 2008 ; El-Bèze et al. 2011).

12 C'est dans cette optique que s'inscrit le travail ici décrit : il s'agit d'appliquer des méthodes de TAL aux revues de presse avec l'objectif de proposer une palette d'outils simples qu'un professionnel des médias pourrait s'approprier facilement et déployer selon ses besoins propres pour mener des analyses fines sur des corpus de type journalistique. A mi-chemin entre une plate-forme informatique ouverte (solution inadaptée pour des non spécialistes) et un logiciel spécialisé (solution trop restreinte et rigide), c'est une solution à la fois souple et puissante qui est recherchée.

3. Corpus

13 Notre corpus provient d'une archive en libre accès disponible sur le site de France Inter. Il compte 716 textes, couvrant la période entre mai 2005 et juin 2011. Il s'agit d'un ensemble de discours relativement homogène, produit dans le cadre d'une pratique professionnelle que nous n'avons pas étudiée pour elle-même. Nous notons une diversité des matériaux exploités et une finalité d'analyse qui ne s'interdit pas une approche subjective.

14 « Chaque jour, un regard sur la presse, à la fois subjectif et le plus large possible : presse papier, presse étrangère, sites d'informations, blogs...Dans les colonnes des journaux, sur le papier des magazines, sur les home-pages des sites et les posts des blogs, il y a chaque jour des pépites, des polémiques, des histoires, des bons mots, des reportages, des images. Essayer de recueillir ces pépites, analyser ce qu'elles disent de nos sociétés : c'est le défi, tous les matins, de cette revue de presse. Aux micros, aux manettes... » (http://www.franceinter.fr/emission-revue-de-presse)

15 Nos analyses sont guidées entièrement et uniquement par le contenu des textes archivés. Autrement dit, nous ne savons que ce que ces discours disent explicitement d'eux-mêmes et à propos des contenus relatés. Il est ainsi possible de relever les noms des chroniqueurs, savoir quels jours la revue de presse se fait à deux voix (1), etc. Les chroniqueurs prennent soin de signaler le caractère exceptionnel de certaines revues de presse, focalisées sur un événement majeur de l'actualité (la mort de Ben Laden, par exemple). De même, des revues fortement thématisées (révolutions arabes, départs en vacances...) peuvent être identifiées grâce à des annonces spécifiques (2).

(1) <« vendredi 28 janvier 2011 »> Patrick Cohen : La Revue de Presse à deux voix du vendredi, Guyonne de Montjou, Bruno Duvic. A la Une, ce matin évidemment : la contagion révolutionnaire... Bruno Duvic : De la Mauritanie au Golfe persique, Le Figaro publie une grande carte de l'Afrique du Nord et du Proche Orient. L'onde de choc de la révolution tunisienne.

(2) <« vendredi 24 juillet 2009 »> « UN » Mot Omniprésent dans la presse ce matin....le Mot : « Vacances » ! ...c'est « de saison » me direz-vous, de Fait' !

16 Ces textes sont un support pour l'oral. Ils ne correspondent pas à la transcription du rendu final. Nous ne vérifions pas la correspondance avec ce qui a effectivement été prononcé en antenne. Nous ne revenons pas aux sources dans la presse citée pour recouper l'information. Nos données correspondent à une « réalité » réduite, amputée de l'enveloppe vocale et comportementale.

17 Formellement simplifiés (abréviations, absence d'accents, de majuscules, de marques de pluriel), ils comportent de nombreux syntagmes averbaux, différentes traces de mise en valeur, des marques de pause. Fonctionnellement, il s'agit de discours rapportés (ou en circulation), issus de la presse ou du Web, brièvement commentés, souvent articulés entre eux et parfois mis en scène de manière globale, grâce au recours à la métaphore filée notamment (3,4). Loin d'être un catalogue de citations, les revues de presse articulent et fusionnent des discours. Tetu (1993) parle de subjectivités orchestrées.

(3) <chronique date = « mercredi 25 mai 2011 »> Patrick Cohen : Dans la presse aujourd'hui : l'heure du bricolage... (...) Bricolage aussi à Pôle-Emploi...

(4) <chronique date = « mardi 21 juin 2011 »> Patrick Cohen : Quel est le maître-mot, ce matin ? Denis Astagneau : « Domino », un mot qui fait de l'effet... ça n'est pas le jeu proprement-dit, c'est plutôt son dérivé : un domino qui fait tomber l'autre, qui fait tomber le suivant. (...) Aujourd'hui, la Grèce est le domino faible. (...) « Si Athènes tombe, Rome vacillera très vite et Paris tremblera aussitôt ».

18 Il est à noter que notre corpus renferme un important nombre de citations directes (plus de 3700), dont on peut extraire celles qui constituent des titres (5). A l'évidence, cet objet spécifique peut intéresser des analystes de discours qui travaillent sur la presse écrite.

(5) ...« Chômage : qui dit vrai ? »... L'interrogation est en Une du Télégramme...

4. Observations sur les contenus

19 En schématisant, une revue de presse radiophonique fait alterner citations et commentaires, qui renvoient des discours sélectionnés dans la presse papier et du web, lesquels, à leur tour, prennent pour objets des événements, des situations ou des discours. Ce résultat est obtenu grâce à un travail de sélection, de montage et d'élaboration opéré à plusieurs niveaux. Nous pensons que les marques formelles de ces différentes opérations peuvent guider utilement des explorations outillées.

20 Les spécificités du matériau discursif créé et manipulé dans des revues de presse radiophonique sont nombreuses. De manière heuristique, nous avons élaboré une grille d'analyse permettant d'en esquisser les grandes dimensions, d'identifier et de catégoriser des éléments potentiellement significatifs.

21 La première distinction opère entre les sources d'information et ce qui touche au contenu de l'information, à proprement parler. Le deuxième axe de lecture possible est celui qui trace une limite, pas toujours très nette d'ailleurs, entre des éléments factuels et des contenus plus subjectifs. La subjectivité peut toucher tant les sources que le contenu des informations. Questionner les sources, c'est connaître le support de l'information, la catégorie du média... l'auteur, son statut... Interroger le contenu, c'est identifier (i) les sujets traités (thèmes, problématiques...) (ii) ce qui en est dit, de quelle manière et par qui. L'attitude subjective des chroniqueurs peut porter sur l'information elle-même, sa valeur et la façon de l'élaborer (la méthode, la qualité du travail des confrères...). Face aux faits sélectionnés et relatés, les commentaires dévoilent une large gamme d'attitudes : de nature évaluative (sur l'importance ou l'intérêt...) ou émotionnelle (enthousiasme, soulagement, inquiétude, étonnement...).

22 Comme nous l'avons dit plus haut, ces différents éléments sont exprimés par des marques spécifiques. Outre des lexiques référentiels regroupant des noms propres des journaux, des personnes, des lieux... il y a également des termes techniques de la presse (types d'articles, de publications...), ainsi que des constructions récurrentes permettant la mise en discours des contenus relatés (introducteurs de citations, commentaires, transitions...). Des schémas d'expression préférentiels qui articulent ces différents éléments ont été mis au jour, avec l'élaboration de patterns et de grammaires. Plus concrètement, ces schémas captent des relations entre différents constituants des discours de revues de presse (référentiels, modaux...), par exemple <NP_journal>+ <Emotion_Inquiétude> + <Citation>, voir (a, 1er et 2e items).

23 Ce travail a été opéré à plusieurs niveaux de grain. Les trois paradigmes qui suivent exemplifient des séquences prototypiques correspondant respectivement à (a) des structures citationnelles composées, (b) une grammaire particulière des titres rapportés, (c) des commentaires d'introduction ou de transition employés par des chroniqueurs. Toutes ces attestations concernent l'expression de l'inquiétude (l'« Inquiétude » est l'une des catégories d'émotion de la ressource EMOTAIX, voir la section 5.1).

(a) ... Le Figaro s'inquiète : « Nucléaire : l'Iran accélère »...
Et LES ECHOS s'inquiète de la « baisse surprise des ventes de téléphones mobiles en Europe »...
L'inquiétude de « La Croix » : « La tentation du clonage thérapeutique »...
« Alerte rouge sur la monnaie européenne », confirme La Tribune.
Selon « Le Parisien », l'inquiétude monte...
Plus que troublante, la Une de L'Humanité...
(b) Le cri d'alarme de X
Alerte (rouge...) sur/contre Y
X craint/redoute... Y
Les X s'inquiètent
Les X inquiets pour leurs Y
Ces Y qui inquiètent X
X en danger
Les X doivent-ils s'inquiéter ?
La crainte de Y resurgit/...
(c) Pour terminer... cette nouvelle, qui va rassurer les plus inquiets...

(...) disponibilité des traitements... Inquiétude donc.
On parlait d'inquiétudes... Ce n'est pas le prix du pétrole qui va les calmer...
Parole à l'inquiétude, et à une certaine nostalgie...
Et ça peut effectivement inquiéter...

24 L'analyse de la phraséologie employée dans les textes de notre corpus a permis d'identifier des formes d'expression privilégiées. Les fréquences élevées d'apparition de certains termes (*problème, mal...*)[2] ont conduit à l'élaboration des concordances qui révèlent des constructions récurrentes, par exemple le tour « Le problème, c'est que... », (d) et (6).

(d) Le problème, c'est que le soleil est en option.
Le problème, c'est que la course à l'Élysée, en brouillant...
Le problème, c'est que le bachotage politique, ça existe aussi...
Le problème c'est que cet homme est en fait réputé pour sa...
Le problème, c'est qu'ils ont l'impression que leur situation ne change pas... (voir l'extrait 6).
(6) ... Plus largement, ce sondage... eh bien, « il est dérangeant », commente Éric Hacquemand... « Il y a malaise, écrit le journaliste... 56 % des Noirs de France se disent victimes de discrimination... et d'abord dans les espaces publics et les transports en commun... Et **le problème, c'est qu**'ils ont l'impression que leur situation ne change pas... Pour 37 % d'entre eux, elle s'aggraverait même » [...] En fait... **le problème** est peut-être ailleurs... C'est Paul Burel, dans Ouest-France, qui le pointe du doigt... **Le problème, c'est** la qualité de l'emploi... sa précarisation...
(<chronique date = « mercredi 31 janvier 2007 »>)

25 De telles fonctionnalités d'analyse permettent non seulement un accès plus ciblé et raffiné au contenu des discours, mais également un moyen d'observer leurs formes d'expression. Les discours archivés offrent une image des pratiques journalistiques, y compris sur le plan langagier. Leur analyse peut intéresser aussi bien les différents observateurs des médias que les journalistes eux-mêmes (en exercice, en qualité de formateurs).

26 Les questionnements peuvent être divers. Détecter des routines d'expressions, observer le spectre d'emplois de certaines termes (*danger, crise, scandale...*), travailler sur l'équilibre (thématique, émotionnel...) de son discours, garantir une présence équitable aux différentes sources, surveiller la visibilité des personnes ou des catégories de personnes... Ou encore, en cours de composition de la revue, surveiller la concentration des termes à polarité négative et leur équilibre dans les séquences citationnelles et les commentaires (7,8), apprécier finement leur intensité (9, 10)... Enfin, porter une attention accrue aux présupposés (10), ainsi qu'aux degrés de prise en charge ou de distanciation par rapport aux états des choses exprimés.

(7) Alerte noire titre Libération. C'est un drame planétaire renchérit France soir.
Les États-Unis ont déclaré l'état de catastrophe nationale.
(8) savoir si depuis nous avons tiré les leçons du drame ? C'est là qu'est l'angoisse, c'est qu'on redoute un éventuel nouveau manquement.
(9) Des scores inattendus... incroyablement élevés... qui donnent la mesure du malaise de la société française...
(10) Retour à la violence de la société française. Elle n'émane pas que des Bleus.

5. Outils et ressources mis en œuvre

27 Le système mis en place pour mener ces travaux s'appuie sur la plate-forme « Sense Miner » développée par la société Noopsis, et en particulier le moteur « SemioLabs »

qui est conçu pour opérationnaliser des modèles linguistiques complexes de façon robuste et efficace. La plate-forme incorpore également les formalismes et outils du « Semantic Web » qui sont ici utilisés pour représenter, stocker, et interroger les données extraites à partir des modèles linguistiques.

5.1. Ressources

28 Les ressources linguistiques intégrées dans le système ont essentiellement trois origines : (i) ressources issues de nos travaux passés ; (ii) ressources et modules « standards » de la plate-forme Noopsis, notamment Entity Miner qui permet l'extraction automatique d'entités nommées et de termes thématiques ; (iii) ressources constituées par des pairs, en particulier la ressource EMOTAIX issue du domaine de la psychologie.

29 Les ressources issues de travaux précédents sont principalement composées de lexiques structurées et de différents types de grammaires formelles. Elles concernent notamment :

- Des entités nommées « spécifiques » correspondant aux chroniqueurs, et aux publications susceptibles d'être citées : ressources lexicales complémentaires pour le module Entity Miner (environ 150 entrées).
- Des éléments de typologie des objets journalistiques : types d'articles (chronique, édito, interview, etc.), fréquences de publications (mensuel, hebdomadaire, etc.), types de publications (journal, magazine, blog, etc.) : ressources lexicales catégorisées (environ 50 entrées).
- Les introducteurs thématiques (ex. « En ce qui concerne X ») : ressources lexicales (environ 100 entrées) associées à une grammaire syntagmatique.
- Le discours rapporté direct : ressources lexicales (ex. verbes de parole, environ 700 entrées) associées à une grammaire phrastique.
- Marqueurs liés à l'expression de la subjectivité : intensificateurs, champs lexicaux de l'accord et du désaccord, émotions, etc. : ressources lexicales catégorisées (environ 350 entrées).

30 Le module Entity Miner procède à l'extraction et à la désambiguïsation d'entités nommées sur la base de vastes ressources a priori combinées à un large ensemble de grammaires contextuelles permettant l'extraction d'entités non connues a priori, la désambiguïsation des entités le cas échéant, et l'extraction de « concepts » thématiquement saillants dans le document.

31 La ressource EMOTAIX (Ginouves 2008 ; Piolat et Bannour 2009) a été originellement développée pour identifier le lexique émotionnel et affectif au sein du logiciel « Tropes ». Il est composé de plus de 4000 termes du lexique émotionnel et affectif de la langue française (substantifs, verbes, adjectifs, adverbes, et locutions diverses). Les entrées de la ressource couvrent les émotions d'arrière-plan (énergie, malaise, excitation, etc.), les émotions primaires (peur, colère, dégoût, tristesse, etc.), les émotions sociales (sympathie, embarras, honte, etc.) et les sentiments (amour, haine, etc.).

5.2. Annotation des documents et stockage des données extraites

32 La phase de collecte, prise en charge par les modules appropriés de la plate-forme Sense Miner, consiste à crawler le site de France Inter sur la base de règles définies de façon spécifique pour la récupération des contenus proprement dits ainsi que des quelques métadonnées accessibles (date, chroniqueur, etc.). Les contenus sont stockés localement de façon à faciliter et accélérer la mise en œuvre des différentes expérimentations.

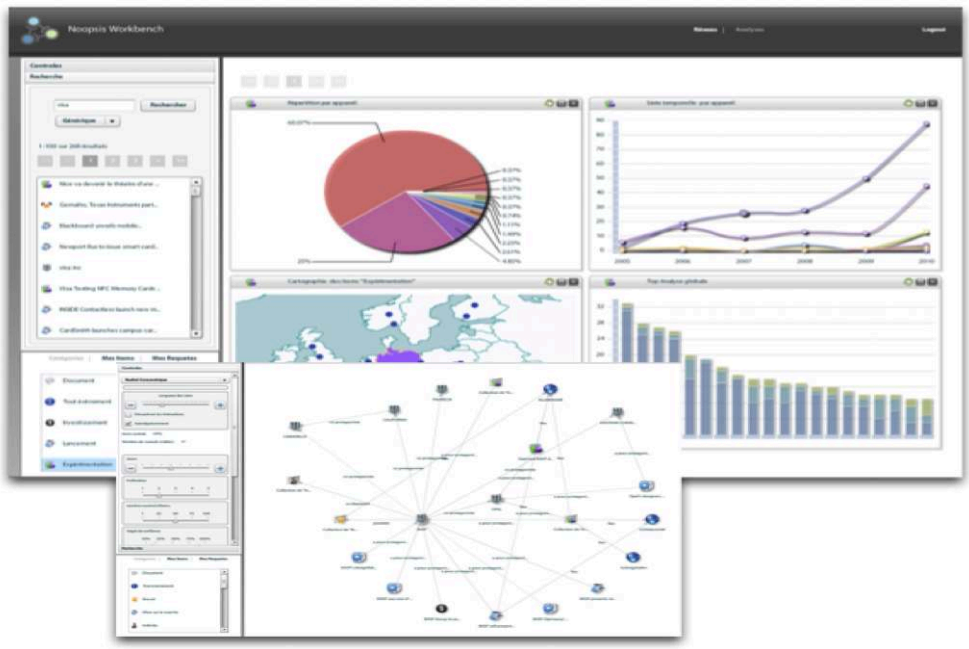
33 La phase d'analyse consiste à projeter l'ensemble des ressources précédemment décrites sur chacun des textes collectés. Le moteur SemioLabs permettant de gérer simultanément un nombre quelconque d'annotations d'ordres et de niveaux divers, toutes les occurrences repérées à l'aide des différentes ressources peuvent donc coexister au sein d'un même texte, y compris en se chevauchant si nécessaire.

34 À l'issue de l'analyse proprement dite, l'ensemble des annotations (y compris les métadonnées recueillies au moment de la collecte) sont exportées sous la forme de triplets RDF (Resource Description Format), c'est à dire sous la forme d'un graphe conceptuel conforme aux standards définis par le World Wide Web Consortium (W3C) dans le cadre de son initiative « Semantic Web ». L'intérêt de cette approche est multiple :

- elle s'adapte automatiquement à la typologie des annotations effectivement repérées dans le texte, le processus restant inchangé en cas d'ajout de nouveaux marqueurs ;
- le modèle est cumulatif est permet d'accueillir à tout moment de nouvelles informations sans remettre en cause les informations déjà existantes ;
- les annotations sont représentées dans un format parfaitement standard et sont donc facilement réutilisables ;
- elles sont en outre potentiellement « connectables » à toute autre ressource déjà existante au sein du Web Sémantique (ex. les chroniqueurs sont décrits dans la base dbPedia, les lieux dans GeoNames, etc.), ce qui permet d'élargir considérablement le champ des les analyses possibles.

35 L'ensemble des graphes RDF produits par l'analyse de chaque revue de presse est stocké dans un « triple-store », c'est à dire une base de données spécialisée dans le stockage de connaissances formelles sous forme de triplets RDF, qui permet également de procéder à des inférences à base de règles, et d'interroger la base à l'aide du langage de requêtes « SPARQL » (SPARQL Protocol and RDF Query Language).

36 In fine, l'application Workbench développée par Noopsis permet de manipuler visuellement la base de connaissances pour réaliser des analyses, naviguer dans le graphe, retourner aux extraits, etc.



6. Exemples d'expérimentations

37 L'objectif du présent travail est de mettre en place une plate-forme préfigurant un outil qui pourrait être utile aux acteurs et aux observateurs des pratiques médiatiques pour réaliser des expériences sur corpus, et non pas de réaliser ces expériences à proprement parler. Nous espérons en revanche susciter, à travers ces premiers résultats, des collaborations qui permettraient de finaliser l'outil tout en aboutissant à des observations utiles.

38 À titre purement illustratif, nous avons toutefois mené trois séries d'expériences :

- la première consistait à mener de simples observations sur les différentes catégories d'informations pour en dégager quelques tendances marquantes ;
- la deuxième visait à observer des corrélations entre différents types d'informations, en croisant notamment des contenus factuels et des contenus subjectifs ;
- la troisième visait à comparer la teneur émotionnelle des citations et de leur discours environnant.

39 La première expérience a fait l'objet d'une démonstration lors de la journée d'étude : « *De l'autre côté du média : radiographie d'une matinale dans les coulisses de France Inter* » du 2 juillet 2012.

40 La troisième expérience n'a pas permis de faire apparaître de résultat tangible à ce stade, aucune différence réellement significative n'ayant pu être observée pour le moment entre les propriétés « subjectives » des citations et de leur discours citant. Il s'agit cependant d'un travail en cours qui fera peut-être apparaître des résultats ultérieurement.

41 La seconde expérience fait quant à elle apparaître un certain nombre de résultats que semblerait confirmer l'intuition profane, et mériterait peut-être d'être approfondie avec le concours d'un spécialiste des pratiques médiatiques. À titre d'exemple, et sans nous livrer à aucune interprétation, voici quelques unes des corrélations qui sont

apparues entres les marqueurs de subjectivité et des marqueurs factuels représentés par les entités nommées de type « publication » et « personne ».

42 La figure ci-dessous traduit des corrélations entre les catégories d'émotions EMOTAIX et les publications citées dans les revues de presse. Le signe « + » indique que la corrélation avec la catégorie est significativement surreprésentée, et le signe « - » indique qu'elle est significativement sous-représentée :

	Amour	Apaisement	Bouleversement	Colère	Délivrance	Désir	Douceur	Douleur	Estime	Gaîté	Humiliation	Humilité	Inquiétude	Insatisfaction	Irritation	Mépris	Plaisir	Rage	Santé mentale	Sérénité	SURPRISE	Tristesse
Canard Enchaîné	+			+			-			+				+			+	-			+	
Charente Libre				-										+								
Courrier International		-		-		-	+	+									-					
France Soir			-																			
L'Equipe															+							
L'Est Républicain																					+	+
L'Express		-																	+	+		-
L'Humanité				+	+																	
La Croix																						
La Dépêche du Midi											+		+									
La Tribune			+																			
Le Figaro																						
Le Midi Libre							-								+	-						
Le Monde																						
Le Parisien																						
Le Point										-												
Le Progrès		-					-			+		+		-								-
Les Echos																						
Libération																						
Nouvel Observateur	+	-								-												
République du Centre									+													

43 La figure ci-dessous traduit les mêmes corrélations avec les personnalités les plus fréquemment citées dans le corpus :

	Tranquillité	Terreur	Santé mentale	Rire	Ressentiment	Rage	Estime	Douleur	Désir	Déplaisir	Délivrance	Audace	Attirance	Apaisement	Amour
Barack Obama									+						
Dominique De Villepin					+								-		
Dominique Strauss-Khan						+		+			+		+	-	
François Bayrou											-			+	
François Hollande				+											+
Jacques Chirac	+														
Jean - François Copé		+					+			+	+	+			
Lionel Jospin															
Martine Aubry															
Nicolas Sarkozy															
Ségolène Royal			+												

6. Conclusion

44 Le travail que nous venons de présenter vise la création d'outils (quantitatifs et qualitatifs) permettant d'objectiver l'observation des discours médiatiques. Nous estimons que les techniques du traitement automatique des langues ont acquis aujourd'hui une maturité suffisante pour guider des explorations discursives fines et produire des résultats consistants. À titre d'illustration, nous avons livré quelques observations simples et intuitives. La réalisation d'expérimentations réelles ne pourra se faire sans la participation des spécialistes des médias, seuls aptes à définir les besoins et les conditions expérimentales pertinentes.

BIBLIOGRAPHIE

CHARAUDEAU P., (2011), *Les médias et l'information*. De Boeck, INA.

CHARAUDEAU P., (2010) « Une éthique du discours médiatique est-elle possible ? », Communication [En ligne], vol. 27/2 | 2010, mis en ligne le 31 mars 2010, consulté le 07 mai 2014. URL : http://communication.revues.org/3066 ; DOI : 10.4000/communication.3066.

CLAQUIN F., (1993), « La revue de presse : un art du montage, *Langage et société*, n° 84, *Les tailleurs de l'information*, pp. 43-71.

EL-BÈZE M., JACKIEWICZ A., HUNSTON S., (dir) (2011), « Opinions, sentiments et jugements d'évaluation », *Revue TAL*, n° 51 :3.

FAURÉ L. (2013/2016), « Analyser les pratiques discursives radiophoniques : nouveaux enjeux et perspectives », in L. FAURÉ (coord.) « Le discours radiophonique en pratiques », *Cahiers de praxématique* 61.

GINOUVES V. (2008), « Emotaix », *Aldébaran*, Méthode, [En ligne], mis en ligne le 27 août 2008 16h32. URL : http://aldebaran.revues.org/3463.

MINEL, J.-L., DESCLÉS J.-P, CARTIER, E., CRISPINO, G., BEN HAZEZ, S., JACKIEWICZ A., « Résumé automatique par filtrage sémantique d'informations dans des textes», *Revue TSI*, Hermès, 2001, vol. 20, n° 3, pp. 369-396.

PANG B., LEE L., (2008), « Opinion Mining and Sentiment Analysis », *Fondations and Trends in Information Retrieval*, vol. 2, p. 1-135.

PIOLAT A. et BANNOUR R., (2009). « EMOTAIX : un scénario de Tropes pour l'identification automatisée du lexique émotionnel et affectif ». *L'Année psychologique*, 109, pp 655-698. doi : 10.4074/S0003503309004047.

QUINTIN E, JACKIEWICZ A., ROY T. (2011). « Analyse d'énoncés à fort impact et éléments pour leur détection automatique » J. BRES, A. NOWAKOWSKA, J.-M. SARALE, S. SARRAZIN (coord.) *Actes du colloque international Dialogisme : langue, discours* (8-10 septembre 2010, Montpellier), mis en ligne le 10 juillet 2011, consulté le 07/05/2014. URL : http://www.praxiling.fr/dialogisme-langue-discours.html

TETU J-F. (2004), « L'émotion dans les médias : dispositifs, formes et figures », *Mots. Les langages du politique* [En ligne], 75 | 2004, mis en ligne le 22 avril 2008, consulté le 31 janvier 2013. URL : http://mots.revues.org/2843.

WIRTH W. and SCHRAMM H., (2005), "Media and Emotions", in *COMMUNICATION RESEARCH TRENDS, vol.* 24 (2005) N° 3—43, accessible en ligne sur http://cscc.scu.edu/trends/v24/v24_3.pdf, consulté le 7/05/2014.

WIDLOCHER, A. et BILHAUT, F. (2008). « Articulation des traitements en TAL - Principes méthodologiques et mise en œuvre dans la plate-forme LinguaStream », *Revue Traitement Automatique des Langues (TAL),* 49(2), p. 73-101.

NOTES

1. Sur cette question, on peut consulter également « Les médias et la peur », *2ème colloque de l'Institut de journalisme et communication*, « Les médias créent-ils ou reflètent-ils les peurs collectives ? » (2003), Recueil de textes préparatoires, accessible sur http://www.thierryherman.ch/wp-content/uploads/colloque_peur_recueil.pdf

2. Dans l'ensemble des extraits annotés avec la catégorie « Bouleversement » (la plus représentée dans le corpus, avec 12 % des attestations « émotionnées »), le terme *problème* apparaît 328 fois, dont 50 fois au sein de la construction « *le problème, c'est que…* ». Au total, 56 catégories d'émotion définies dans la ressource EMOTAIX sont attestées (avec un taux moyen de 1,75 %). La part de la catégorie « Inquiétude » s'élève, quant à elle, à 6 %.

RÉSUMÉS

Nous présentons une série d'expériences linguistico-informatiques appliquées aux revues de presse de France Inter (716 textes, de mai 2005 - juin 2011). Le corpus a fait l'objet d'une annotation sémantique automatique sur différents axes : sources et relais d'informations (type de publication, périodicité, chroniqueurs, etc.), contenus factuels (entités, faits, marqueurs thématiques, etc.), discours rapporté, et marques de subjectivité traduisant différentes attitudes, notamment émotionnelles (enthousiasme, inquiétude, etc.) ou axiologiques (accord, validité, etc.). L'étude se décompose en trois volets : (i) analyse de corpus et construction d'une grille d'analyse ; (ii) constitution de ressources linguistiques opérationnelles ; (iii) mise en œuvre informatique et analyse des résultats.

We present an experiment in computational linguistics applied to press reviews issued by France Inter (716 texts, May 2005 - June 2011). The corpus has been automatically annotated following various formal and semantic criteria: sources and information channels (kind of publication, periodicity, columnists, etc.), factual contents (named entities, facts, topic markers, etc.), quotes, and subjective aspects related to various attitudes such as emotional ones (enthusiasm, anxiety, etc.) or axiological ones (agreement, validity, etc.). The study is divided into three parts: (i) corpus analysis and building of the analytical framework; (ii) establishment of operational language resources; (iii) implementation and analysis of results.

INDEX

Keywords : experimental studies on corpora, France Inter, natural language processing (NLP), radio press review, SemioLabs

Mots-clés : études expérimentales sur corpus, France Inter, revue de presse radiophonique, SemioLabs, traitement automatique des langues (TAL)

AUTEURS

AGATA JACKIEWICZ

STIH Paris-Sorbonne
Agata.Jackiewicz@paris-sorbonne.fr

FRÉDÉRIK BILHAUT

Noopsis, Caen
frederik.bilhaut@noopsis.fr