La subjectivité à travers les médias : étude comparée de les médias participatifs et de la presse traditionnelle

Lydia-Mai Ho-Dac, Anne Küppers

La subjectivité à travers les médias : étude comparée des médias participatifs et de la presse traditionnelle[1]

Subjectivity in mass media: comparing participatory and traditional journalese

Lydia-Mai HO-DAC et Anne KÜPPERS
Université de Toulouse-le Mirail et Université Catholique de Louvain / Louvain-la-Neuve

Résumé : Cette étude propose une analyse des différences et similitudes linguistiques dans la presse écrite traditionnelle et les médias participatifs en ligne afin d'évaluer dans quelle mesure la production et la diffusion en ligne peuvent modifier nos usages linguistiques. Les analyses effectuées se basent sur un large corpus (8 millions de mots) qui représente des modes d'expression et des degrés de subjectivité a priori différents.

Abstract: This paper investigates linguistic differences and similarities in a traditional newspaper and participatory on-line media in order to examine to what extent the on-line production has an impact on the language use. Our method of investigation is based on corpus analysis. We analyze a large-scale corpus (8,000,000 words) composed of datasets representing different steps on a graduate scale from traditional printed newspaper to online citizen press.

Mots clés : linguistiques de corpus, nouveaux médias, subjectivité, organisation du discours
Key words: corpus linguistics, on-line mass media, subjectivity, discourse organisation

[1] Avec le soutien accordé, dans le cadre des Partenariats Hubert Curien, par les Ministères des Affaires étrangères et européennes (MAEE) et de l'Enseignement supérieur et de la Recherche (MESR), Wallonie-Bruxelles International (WBI) et le Fonds de la Recherche Scientifique (FRS).

1. Introduction

Cette étude vise à observer les différences et similitudes linguistiques entre les articles issus de la presse écrite traditionnelle et de médias participatifs en ligne afin d'évaluer dans quelle mesure la production et la diffusion en ligne des textes peuvent modifier les usages linguistiques. Les analyses effectuées se basent sur un large corpus (8 millions de mots) contrastant des processus de mise en ligne différents (voir la section suivante). Le corpus a été préparé à l'aide de divers modules de traitement automatique (POS-tagger, analyseur syntaxique, extraction de constructions syntaxiques et éléments discursifs) afin de permettre des analyses quantitatives concernant à la fois la distribution d'éléments potentiellement subjectifs (Wiebe et al. 2004) et l'organisation discursive des articles afin de valider leur différence en termes de type de texte.

La constitution d'un corpus contrastant *a priori* trois modalités d'expression de la subjectivité est envisagée dans l'optique d'une analyse plus qualitative visant à répondre aux deux questions suivantes :

1. Dans quelle mesure l'implication de l'auteur diffère-t-elle dans la presse écrite traditionnelle et les médias participatifs en ligne ? S'agit-il d'une question de fréquence et/ou de mode d'expression de cette implication en termes de choix lexicaux, syntaxiques et discursifs ?

2. Dans quelle mesure l'organisation du discours diffère-t-elle dans les médias en question ? Les stratégies de mise en texte dans les nouveaux médias en ligne sont-elles différentes comparées aux médias plus traditionnels ?

Dans cet article, nous présentons essentiellement l'aspect méthodologique de notre étude en détaillant tout d'abord la constitution et la préparation du corpus d'étude qui en est la base. Nous présenterons ensuite les hypothèses de travail et leur évaluation via un repérage automatique d'indices de surface et donnant lieu à des analyses pour l'instant uniquement quantitatives. Nous finirons par la présentation des tendances générales qui se dégagent de ces analyses et nous permettent de

valider notre choix de corpus et de dresser une première typologie des trois médias en question.

2. Un corpus de presse pour l'étude de la subjectivité

Nous avons constitué un corpus de 8 000 000 de mots qui permet de contraster trois médias présentant *a priori* des caractéristiques différentes au niveau de l'implication de l'auteur et de l'expression de la subjectivité.

D'un côté, le quotidien francophone belge *Le Soir*, existant depuis 1887, propose depuis 1996[2] une version papier et une version en ligne. Ce quotidien, l'un des principaux quotidiens francophones de Belgique, représente, dans notre corpus, la presse dite traditionnelle. Les journalistes qui publient pour la version papier de ce média sont confrontés aux "classiques du métier" de journaliste : *deadlines* pour l'impression, normes de rédaction (longueur, ton, style, etc.), objectivité de l'information (en dehors des rubriques consacrées à l'expression d'opinion *e.g.* éditorial, commentaires, critiques, etc.). Pour la version en ligne, le standard est autre : par rapport aux "classiques du métier" (*e.g.* évaluation et validation de l'information, temps de rédaction jusqu'au moment du bouclage, etc.), le journaliste web subit davantage l'immédiateté de la mise à disposition de l'information, cherchant à mettre la nouvelle en ligne le plus vite possible, à être le premier à publier l'information pour avoir l'exclusivité, ce qui peut l'amener à simplement copier-coller et adapter à la ligne éditoriale une dépêche d'agence de presse[3]. La distinction entre publications en ligne et publications papier est difficile à cerner dans ce média : jusqu'en mai/juin 2009, les deux rédactions étaient séparées mais l'archivage des articles tel qu'il est

[2]D'après l'article de Nicolas Vuille, URL : http://archives.lesoir.be/-soir-en-ligne-plus-clair-et-aere_t-20060608-005KMW.html

[3]Cette étude appartient au projet ARC 08/13-11 (Action de Recherche Concertée), URL : http://www.onlinejournalism.be/. Cette ARC pluridisciplinaire se compose de six recherches particulières, menées chacune par un doctorant ou post-doctorant affiliés à différents départements de l'Université Catholique de Louvain et des Facultés Universitaires Notre-Dame de la Paix. Parmi celles-ci, une thèse concerne expressément l'étude des pratiques journalistiques depuis l'avènement d'Internet (Degand 2009).

accessible sur le site du journal ne garde aucune trace de cette distinction de rédaction. Par conséquent, les articles de notre corpus issus de ce média correspondent indistinctement à des parutions en ligne et à des parutions papier.

De l'autre côté, *Rue89* et *AgoraVox*, sites d'informations qui n'existent qu'en ligne[4], se définissent comme des médias participatifs alternatifs. *Rue89*, site d'actualité en réseau avec comité de rédaction "professionnel", publie depuis 2007 des articles rédigés par des journalistes ou des experts. Les internautes participent au contenu uniquement en proposant des thématiques et en publiant leur(s) commentaire(s) (Benhamou 2009). À l'inverse, *AgoraVox* est un média citoyen qui suit une ligne éditoriale davantage "démocratique" : les internautes proposent des articles et acquièrent le statut de modérateur, appellation des membres du comité de rédaction, à partir du moment où au moins quatre de leurs articles ont été acceptés (voir Dugué 2008 pour une description comparative de ces deux médias participatifs).

En résumé, nous pouvons distinguer ces trois médias (*Le Soir*, *Rue89* et *AgoraVox*) selon deux caractéristiques extra-linguistiques interdépendantes qui sont :

1. le **degré de professionnalisme** des auteurs *: Le Soir* fait appel des journalistes professionnels, *Rue89* est un lieu de journalisme en réseau (journalistes ou spécialistes de domaine) et *AgoraVox* est un lieu d'expression de citoyens tout court.

2. la **philosophie du média** relayée par l'intention de l'auteur de publier via tel ou tel média : d'un côté *Le Soir* recherche le scoop et l'information objective (hors genre "opinion" comme les éditoriaux, les commentaires, les critiques) ; de l'autre, *Rue89* et *AgoraVox* affichent une volonté de "faire de l'information" différemment, sans ligne éditoriale particulière et avec une certaine liberté quant au ton, au temps et au format de publication.

Les articles extraits dans les trois médias appartiennent aux rubriques présentées dans le tableau 1. Ces rubriques

---

[4] Depuis juin 2010, *Rue89* existe également sous la forme d'un mensuel papier qui sélectionne les meilleurs articles parus en ligne au cours du mois précédent.

couvrent toutes celles présentes dans Rue89, le plus jeune des média, et leurs équivalents dans les deux autres médias.

Tableau 1: Rubriques retenues dans les trois corpus

Le Soir	*Rue89*	*AgoraVox*
Actualité	Monde	International
		Europe
	Politique	Politique
	Société	Société
Culture	Culture	Culture
	Médias	Médias
		Religion
		Étonnant
		Mode et tendances
		People

2.1. L'expression de la subjectivité dans ces médias

Sans s'avancer sur une possible définition de genres différents entre ces trois médias, ou du moins entre presse traditionnelle et médias participatifs, nous posons l'hypothèse que les modalités de production et le degré de professionnalisme associé à la philosophie du média auquel un auteur soumet son texte ont une influence sur son implication et sur l'expression de la subjectivité. Concrètement, nous voulons vérifier les hypothèses suivantes :

1. *AgoraVox* et *Rue89* sont *'plus subjectifs' que Le Soir*, eu égard à la contrainte d'immédiateté de la mise à disposition de l'information dans ce dernier média qui suppose un style d'écriture moins individuel, moins subjectif, moins diversifié ;

2. Selon le paramètre de professionnalisme, *AgoraVox* est encore plus subjectif que *Rue89* ;

> Il est bien difficile de trancher entre les deux sites d'information si bien que c'est au niveau de la présentation que tout semble se jouer, donnant à Rue89 [qui devrait satisfaire des lecteurs exigeants envers une information qu'ils veulent complète et objective] un aspect plus professionnel avec les articles rédigés par des journalistes formés et/ou d'expérience. [...] AgoraVox donne une fausse impression d'amateurisme et, surtout, on sent un esprit

différent, une sorte de proximité, de regard plus subjectif sur le monde, les événements. (Dugué 2008)

3. L'expression de la subjectivité est plus variée dans *AgoraVox* et *Rue89* comparés à *Le Soir*, selon l'hypothèse de Wiebe et al. (2004) concernant le lien entre opinion et créativité linguistique : « people are creative when they are being opinionated » (*ibid* : 286).

Selon ces hypothèses, nous aurions l'échelle : ***Le Soir* < *Rue89* < *AgoraVox***, du moins au plus subjectif. Notre méthodologie tire parti de ces trois types ou degrés de subjectivité afin de réaliser une analyse comparée des caractéristiques linguistiques de ces trois médias.

2.2 Constitution et préparation du corpus

La procédure de constitution du corpus suit trois étapes :

1. l'extraction des archives listant les articles à extraire ;
2. l'extraction des articles et leur normalisation selon la TEI-P5 ;
3. le nettoyage des corpus générés.

L'extraction des articles se fait au moyen de scripts perl développés spécialement pour ces trois médias. Ces scripts permettent à la fois une extraction des articles, leur encodage selon la norme éditoriale TEI-P5 (avec renseignement du header) et leur concaténation dans un <teiCorpus> pour chaque source retenue.

De façon générale, le script d'extraction des articles repère les URL des articles mentionnés avec, si possible, mention de leur titre, de leur date de publication, de leur auteur, de leur(s) rubriques de classement. Au cours de cette phase, les brèves (très nombreuses dans *Le Soir*) sont en partie éliminées automatiquement. Les brèves restantes, les entretiens/interviews et les poèmes ou chansons (ainsi que les parties de texte correspondant à ce type de texte) sont supprimés lors du nettoyage manuel des corpus. Lors de cette extraction, les articles sont mis automatiquement au format XML selon la norme TEI-P5, ce formatage fera ensuite l'objet d'une vérification manuelle. Les articles sont alors rassemblés en sous-corpus. Un script de nettoyage gère ensuite certaines

erreurs redondantes, enlève les articles en doublons et, par la même occasion, liste les articles et leur nombre de mots.

L'étape suivante consiste à passer en revue chaque sous-corpus pour le nettoyer manuellement jusqu'à obtention d'un document XML bien formé (*i.e.* validé par l'éditeur XML Oxygen et visualisable sous FireFox). Lors de ce nettoyage manuel, il est très fréquent de devoir retourner à la version HTML en ligne pour comprendre la construction de certains articles, notamment pour vérifier les titres de section, les listes, la complétude du texte ou pour supprimer des parties de textes parasites (cadres promotionnels, questionnaires, figures remplaçant les puces, etc.) Nous avons également décidé de supprimer dans les articles certaines parties de discours susceptibles de biaiser notre analyse, notamment car elles relevaient d'un autre discours que celui de l'article même : entretiens et interviews ; poèmes et paroles chansons (certains articles proposaient des extraits de poèmes ou chansons) ; légendes de photos et vidéos ; notes de fin d'article, annexes, liste des “lire aussi” ou “à voir aussi”.

Pour les entretiens/interviews et poèmes/chansons, leur mention est gardée par une balise de division auto-fermante (,). Les parties de fin d'article (notes, annexes, bibliographie, etc.) sont supprimées. La suppression des notes est plus problématique puisqu'il faut au maximum enlever également la mention de ces notes dans les corps de texte. Au final, nous obtenons trois corpus dont les caractéristiques sont détaillées dans le tableau 2.

Tableau 2: Caractéristiques générales des trois corpus

	Mots	Articles	Mots/Articles
Rue89	2 187 333	3 879	564
AgoraVox	**3 281 208**	4 368	**751**
Le Soir	2 744 270	**5 873**	467
	8 212 811	14 120	582

Les caractéristiques visuelles des textes dans nos trois médias montrent déjà certaines différences, avec d'un côté *AgoraVox* qui contient les articles les plus longs composés des paragraphes les plus longs et présentant le plus de titres de

section et de l'autre, *Le Soir* qui contient le plus d'articles mais de taille beaucoup plus courte. *Rue89* se situe dans une position intermédiaire.

Ces trois corpus font ensuite l'objet d'un étiquetage morpho-syntaxique (TreeTagger) et une analyse syntaxique (SYNTEX) afin de permettre le marquage et l'extraction automatique des indices.

3. Méthodologie

Notre analyse comparée des caractéristiques linguistiques des trois médias se décline en deux analyses descriptives complémentaires. Premièrement, nous réalisons une caractérisation des stratégies textuelles principalement utilisées dans les trois médias, en les comparant à des résultats antérieurs permettant de classer les sous-corpus selon leur type textuel : à tendance descriptive/narrative ou argumentative. Une seconde analyse repose sur l'étude distributionnelle d'éléments potentiellement subjectifs (EPS).

3.1 Typologie textuelle via la position initiale

Les éléments situés en position initiale de phrase (désormais « éléments en position initiale ») constituent une porte d'entrée dans l'organisation du discours du fait même de leur position : à l'articulation de deux segments. La position initiale fonctionne comme un pivot servant à la fois d'ancrage et de point de départ au nouveau segment. Notre conception s'accorde entièrement à celle exposée par (Virtanen 1992, 2004) qui définit la position initiale par les trois fonctionnements suivants : lier le discours antérieur au discours à venir ; orienter l'interprétation des segments à venir ; conférer aux éléments initiaux une certaine saillance dans la construction de la représentation mentale, ces éléments étant alors associés à l'information « cruciale ».

Cette idée d' « information cruciale en premier » permet d'appréhender le fonctionnement général de la position initiale en discours : ce qui est jugé comme « le plus important » par l'auteur varie selon la situation. L'auteur juge parfois plus important d'insister sur le lien à établir entre deux segments en

mettant, par exemple, l'information donnée en premier. Par ailleurs, il peut parfois trouver plus important de préciser que le cadre dans lequel les informations sont à interpréter a changé.

L'organisation du discours via les éléments en position initiale a été étudiée par Ho-Dac (2007, 2010) qui a mis en place une méthodologie en corpus permettant de mettre au jour des distributions significatives d'éléments en position initiale selon différentes positions textuelles (initiale de paragraphe, initiale de section, position intraparagraphique) et différents types de texte (descriptif, narratif et argumentatif). Cette étude a montré que, du point de vue de l'organisation discursive en position initiale, les textes les plus descriptifs présentent significativement plus de phrases mais surtout de paragraphes commençant par un circonstant antéposé à l'inverse des textes plus argumentatifs qui privilégient les séquenceurs et adverbiaux modalisateurs en position initiale. Parallèlement, les textes plus descriptifs/narratifs (dans le cas précis, des portraits encyclopédiques de célébrités) montrent significativement plus d'indices de continuité topicale (appositions et sujets pronominaux) et de connecteurs en position initiale, en milieu comme en début de paragraphe.

En appliquant cette méthodologie à notre corpus de presse et en contrastant les résultats issus de nos trois médias à ceux issus de Ho-Dac (2007, 2010), nous serons davantage renseignées sur la nature linguistique des textes qui composent notre corpus d'étude et des stratégies textuelles utilisées par les auteurs pour mettre en texte l'information.

3.2 La subjectivité à l'épreuve : l'expression d'un état personnel

Le phénomène linguistique de subjectivité fait actuellement l'objet de nombreuses études s'inscrivant dans diverses théories proposant différentes dénominations. En grammaire systémique fonctionnelle (SFG) on parle d'*appraisal* (*e.g.* Thompson & Hunston 2000, Thompson 2001, Hunston 2000, White 2001), Bednarek (2006) parle d'*évaluation* en se distanciant quelque peu du cadre théorique très stricte de la SFG, Biber *et al.* (1989, 1999, 2000) parlent de *stance*. C'est tout d'abord dans le cadre

de l'analyse du discours (Traugott 1995, Lyons 1982 *inter alia*) que l'on parle *subjectivité*[5].

Notre définition de la subjectivité reprend totalement la proposition de Thompson & Hunston (2000), bien que, dans leur définition, les auteurs parlent d'*évaluation* :

> The broad cover term for the expression of the speaker's or writer's attitude or stance towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about. That attitude may relate to certainty or obligation or desirability or any of a number of other sets of values.

Supposant que le terme d'évaluation implique toujours un jugement (positif/négatif, bien/mauvais, (peu) important/ (in)attendu/(im)possible/nécessaire/superflu etc. (Bednarek 2006 : 3)), nous préférons la dénomination *subjectivité*, qui nous semble moins restrictive et permettra d'autres catégorisations du phénomène.

La question de l'expression de la subjectivité a beaucoup été traitée au cours des deux décennies passées, avec des études focalisant surtout sur l'anglais et le discours académique. La majorité de ces études proposent des analyses en corpus guidées par l'étude de marqueurs spécifiques comme les marqueurs de discours (Hyland 1998, Harwood 2005a, 2005b) ou les pronoms déictiques de l'énonciateur (Lyons 1982, Hyland 2004, 2005, Kuo 1999 *inter alia*). Ce n'est que très récemment que des analyses plus inductives, basées sur de gros corpus, mêlant plusieurs indices potentiels de subjectivité, et étudiant d'autres langues que l'anglais sont menées (Conrad & Biber 2000, Wiebe *et al.* 2004). Notre recherche se place du côté innovateur des recherches sur l'expression de la subjectivité en proposant une analyse basée sur un large corpus francophone, contrastant trois sous-corpus associés *a priori* à différentes modalités d'expression de la subjectivité.

La première étape de l'étude que nous présentons ici consiste essentiellement à valider notre corpus en observant la distribution dans les différents sous-corpus de cinq catégories d'**éléments potentiellement subjectifs** (désormais **EPS)** sélectionnés à la fois pour leur fort potentiel subjectif et la

[5]D'autres travaux proposent les termes *hedging* (Hyland 1994), *private state* (Quirk *et al.* 1985) ou encore *commitment* (Stubbs 1986).

possibilité de les extraire automatiquement sans recours à une ressource particulière (ressource inexistante pour le français au début de ce projet) : pronoms personnels et possessifs de première personne, adverbiaux modalisateurs, constructions clivées, constructions impersonnelles et hapax legomena.

Les exemples (1) et (2)[6] présentent des occurrences de pronoms personnels référant à l'auteur du texte. A côté des pronoms de la première personne singulier et pluriel, nous incluons également le pronom o*n* utilisé dans la signification de *nous* comme dans l'exemple (2).

(1) **Je** ne reviendrai pas sur les questions rhétoriques toujours aussi efficaces. [S]

(2) **On** assiste ainsi à une soirée organisée en l'honneur des Amis américains de Versailles dans la Galerie des Glaces du fameux château. [S]

Dans cette même catégorie, nous prenons en compte les pronoms possessifs de première personne qui servent fréquemment à exprimer une auto-référence dans les textes journalistiques, *e.g. de **notre** envoyé(e) spécial(e)*, *de **notre** correspondant(e)* ou *dans **nos** colonnes* comme dans l'exemple (3).

(3) Vendredi dernier dans **nos** colonnes, les recteurs de l'Université libre de Bruxelles (ULB), Pierre de Maret, et de la Vrije Universiteit Brussel (VUB), Ben Van Camp, signaient une Carte blanche. [S]

Autre catégorie d'EPS, les adverbiaux modalisateurs peuvent être employés pour exprimer la subjectivité. À l'origine définis comme marqueurs d'atténuation (Lakoff 1972), les adverbiaux modalisateurs sont désormais considérés comme des indices de subjectivité exprimant le degré de certitude ou d'attente concernant un certain sujet, pour des raisons liées au contenu (*e.g.* le manque d'information sur un certain sujet) ou pour des fins interpersonnelles (*e.g.* le fait de vouloir porter un jugement sans imposer son opinion) (Hyland 1996, 1998). La définition s'élargit en incluant l'expression d'attitudes, d'émotions et d'opinions outre la simple atténuation de

[6]Le média d'où sont tirés les exemples est indiqué entre crochet : [A] = *AgoraVox*, [R] = *Rue89*, [S] = *Le Soir*.

l'énonciation, comme *Bien sûr* dans l'exemple (4) ou encore *évidemment* dans l'exemple (5).

(4) **Bien sûr**, en France, on estimait à une très large majorité (plus de 90 %!) qu'il vaudrait mieux un Démocrate qu'un Républicain à la Maison Blanche, mais on abordait ce sentiment très différemment. [A]

(5) Car il serait **évidemment** bien dangereux de se replier frileusement sur les égoïsmes nationaux d'antan. [R]

L'exemple (5) illustre également notre quatrième type d'EPS : les constructions impersonnelles. Ces constructions dites « spéciales »[7] permettent à l'auteur de porter un jugement indirect sur le contenu propositionnel du message (Lambrecht 2001).

Une autre construction spéciale qui est prise en compte dans notre analyse de la subjectivité sont les clivées, qui permettent entre autres de mettre en avant un élément précis du message : topique, circonstance ou encore modalisation, comme illustré dans les exemples (6) et (7):

(6) **C'est l'amer constat que** l'on peut faire soit qu'on y habite ou qu'on y arrive pour la première fois dans cette ville qui jadis, présentait fière allure. [A]

(7) **C'est d'abord parce qu**'ils gardent au début l'espoir insensé d'un miracle, et qu'ensuite il est trop tard. [A]

Comme le souligne Charaudeau (2006) ces deux constructions (impersonnelles et clivées) sont relativement fréquentes dans les textes journalistiques, avec pour distinction pragmatique le fait que la fonction principale des clivées est de mettre en emphase un élément extrait et isolé du reste de la phrase, alors que celle des impersonnelles est de communiquer un jugement tout en laissant l'auteur en arrière-plan.

Dernier indice inclus dans notre étude, les hapax (*i.e.* les mots à occurrence unique) on été étudiés pour leur potentiel subjectif par Baayen (1996), Baayen *et al.* (2001), Weeber *et al.* (2000) et Wiebe *et al.* (2004) *inter alia*. À l'instar de Wiebe *et al.* (2004), nous supposons que la créativité lexicale constitue

[7]Les constructions « spéciales » ou encore « thétiques » sont caractérisées par un sujet grammatical vide, ce qui en fait, du point de vue de la grammaire systémique fonctionnelle, des constructions à thème vide aux fonctions discursives particulières (cf. Halliday 1985).

un indice d'implication du locuteur et donc un indice potentiel de subjectivité, comme l'illustre les exemples (8) et (9).

(8) Un pays gorgé de pétrole qui aurait pourtant pu, très facilement, développer et **bitumer** ses routes. [A]

(9) La résorption de ces conflits, ou l'absence de conflit doit permettre de créer et de révéler ce troisième espace sacré, celui de la relation elle-même qui doit être un espace de **ressourcement**, indépendant des deux espaces intenses de la personnalité de la mère et de la fille. [A]

Tous ces indices, parce qu'ils sont repérables automatiquement, de façon systématique et qu'ils représentent différentes modalités d'expression de la subjectivité, nous permettent de dresser une première caractérisation du degré de subjectivité des trois médias.

4 Analyses quantitatives

Les analyses quantitatives réalisées se basent sur le marquage automatique et systématique des EPS et des éléments en position initiale décrits précédemment. Pour mesurer les différences significatives entre les trois médias selon les indices extraits, nous utilisons le test du log-likelihood ratio (désormais LL) qui permet de comparer des fréquences deux à deux sur des corpus de grande taille (Rayson & Garside 2000). Ainsi, nous comparons pour chaque couple de médias (*AgoraVox* vs. *Rue89*, *AgoraVox* vs. *Le Soir* et *Rue89* vs. *Le Soir*) les fréquences de chacun des indices retenus, en tenant compte à chaque fois de leur position textuelle. Concernant la position textuelle elle-même, nous comparons également les fréquences pour chaque couple P1 (initiale de paragraphe)/P2 (position intraparagraphique) dans chaque média. Dans cette étude, nous fixons le seuil de significativité à $p < 0.0001$ ce qui correspond à un $LL > 15,13$.

4.1 Marquage et extraction automatique des indices

La caractérisation des EPS et des éléments en position initiale se base sur une analyse syntaxique réalisée par l'outil SYNTEX (Bourigault 2007). Le programme de caractérisation

automatique, réalisé dans le cadre de notre première étude sur l'organisation du discours vue à travers position initiale (Ho-Dac 2007, 2010), a été adapté à notre corpus. Il consiste à repérer les positions préverbales de chaque phrase[8] et à y distinguer trois types d'éléments : les connecteurs simples, les éléments détachés et les sujets grammaticaux. Les éléments détachés ainsi que les sujets grammaticaux passent au travers d'une série de patrons construits sur la base de règles morpho-syntaxiques associées à des lexiques spécifiques afin de caractériser les traits suivants :

- les **éléments antéposés** sont caractérisés selon leur fonction discursive (adverbial circonstanciel, adverbial modalisateur, séquenceur, apposition) et leur sémantique dans le cas des adverbiaux circonstanciels (temps, lieu, notion) ;

- les **sujets grammaticaux** sont caractérisés selon leur catégorie morpho-syntaxique et leur potentiel coréférentiel, en distinguant les pronoms, les SN possessifs, et les redénominations *i.e.* les SN dont la tête lexicale reprend un nom déjà mentionné dans la section en cours.

La détection des EPS vient se greffer sur ce programme de caractérisation automatique, ce qui permet pour chaque indice d'observer sa distribution relativement à sa position textuelle : P1 ou P2. Le tableau 3 montre la distribution des cinq catégories d'EPS dans nos trois sous-corpus, comptabilisée en nombre de phrase contenant au moins un élément appartenant à la catégorie concernée. La distribution des hapax correspond au nombre de token à occurrence unique par sous-corpus.

Tableau 3 : Distribution des EPS

Phrases contenant	*Rue89*	*AgoraVox*	*Le Soir*
Pronom 1re pers. + On	29 468	42 952	24 413
SN poss. 1re pers.	4 715	7 467	4 531
adv. modalisateurs	3 229	6 059	3 201
impersonnelles	1 219	2 814	1 103
clivées	4 567	6 256	4 762
	195 395	**298 636**	**249 830**
Nb de phrases total	181 192	122 630	153 974

[8]En cas de phrases complexes ou propositions coordonnées, etc. seule la première zone préverbale est prise en compte.

Nb d'hapax	**26 849**	**38 006**	**29 777**
Nb de token	3 131 675	5 092 708	3 836 117

5. Premiers résultats

Nos premiers résultats montrent des différences statistiquement significatives entre les trois médias, à la fois au niveau de l'organisation discursive et au niveau de l'expression de la subjectivité.

Du point de vue de l'organisation textuelle et du type de texte, les médias plus "professionnels" se distinguent d'*AgoraVox*. En effet, les indices organisationnels en général sont significativement plus fréquents dans *Le Soir* et *Rue89* (le LL de chaque couple est indiqué dans la première ligne du tableau 4). Si l'on regarde ces différences dans le détail, on s'aperçoit que dans les articles de *Rue89* et *Le Soir*, les indices organisationnels significativement plus fréquents sont de nature circonstancielle, en P1 et en P2[9]. A noter cependant que pour les initiales de paragraphe dans *Le Soir*, seuls les circonstants temporels sont concernés par cette caractéristique, les circonstants spatiaux et notionnels apparaissant plutôt en position intraparagraphique. En comparant la fréquence de ces mêmes circonstants dans *Rue89* et *Le Soir*, c'est le quotidien belge qui se montre le plus descriptif, avec un LL de 39 en sa faveur, mais uniquement concernant les circonstants en P2. A l'opposé de ces deux médias, nous trouvons *AgoraVox* qui montre significativement plus d'indices d'organisation davantage argumentative : connecteurs en P1 et séquenceurs P1/P2.

Tableau 4 : LL concernant l'organisation discursive

	R vs. A[10]	**A vs. S**	**R vs. S**
circonstants	316 (R)	427 (S)	P2 : 39 (S)
séquenceurs	P1[11] : 22 (A)	P1 : 33 (A)	

[9]Par manque de place, nous ne pouvons pas indiquer le détail des mesures effectuées.

[10] (*R,A,S*) indique le corpus présentant un écart significativement positif, R = *Rue89*, A = *AgoraVox*, S = *Le Soir*.

[11]*P1* indique que la différence significative n'a été observée qu'en *P1*, idem pour *P2*. Lorsqu'aucune position textuelle n'est précisée, la différence

connecteurs	P1 : 16 (A)	P1 : 38 (A)	
Continuité topicale		647 (S)	499 (S)
appositions	P2 : 72 (R)	1 716 (S)	832 (S)
redénominations	P2 : 91 (A)	P1 : 35 (S)	94 (S)
pronoms et SN poss.		P1 : 151 (A)	P1 : 16 (R)
Organisation discursive	**117 (R)**	**880 (S)**	**263 (S)**

Pour ce qui est de la continuité topicale, c'est *Le Soir* qui se distingue avec une très forte fréquence d'appositions et de redénominations, ce qui corrobore les LL mesurés pour les circonstants en initiale et associe davantage le journal traditionnel au type descriptif/narratif. Cependant, en regardant maintenant les indices de continuité topicale dans le détail, il semble que le saut de paragraphe remplit des fonctions différentes selon les trois médias. Dans *Le Soir*, le saut de paragraphe semble aller de pair avec l'apparition d'une apposition ou d'une redénomination en position sujet, ce qui indique davantage une continuité topicale à renforcer qu'une continuité topicale forte, comme dans *AgoraVox* ou *Rue89* où l'on trouve significativement plus de pronoms et SN possessifs en P1. Cette différence de la fonction paragraphique est sans doute à relier à la taille des paragraphes, plus longs et plus nombreux dans les deux médias participatifs. Ces premières observations sont évidemment à valider par une étude (ultérieure) plus fine des différences entre P1 et P2 et un retour au texte.

Parallèlement, la distribution des EPS oppose le média traditionnel *Le Soir* aux deux médias participatifs. En effet, tous les EPS, à l'exception des hapax, sont très significativement moins fréquents dans le quotidien, avec des LL très élevés, notamment autour de l'utilisation des pronoms de 1re personne (tableau 5). De plus, les résultats nous permettent de distinguer *Rue89* et *AgoraVox* selon le type d'EPS qui y est observé (colonne R vs. A, tableau 5). Bien que les LL soient beaucoup plus faibles que ceux mesurés entre les médias participatif et la presse traditionnelle, on peut observer une distinction entre d'une part des adverbiaux modalisateurs (ce qui corrobore le

significative a été observée dans les deux positions.

côté plus argumentatif d'*AgoraVox*) et des constructions impersonnelles dans les articles d'*AgoraVox*, et d'autre part, un style peut-être plus direct avec une forte utilisation de clivées dans les articles de *Rue89*.

Tableau 5 : LL pour les 5 catégories d'EPS

Phrases contenant	**R vs. A**	**A vs. S**	**R vs. S**
EPS :		3321 (A)	2830 (R)
Pronom 1re pers. + On	39 (R)	2 396 (A)	2 529 (R)
SN poss. 1re pers.		297 (A)	188 (R)
adv. modalisateurs	90 (A)	460 (A)	104 (R)
impersonnelles	152 (A)	499 (A)	69 (R)
clivées	31 (R)	24 (A)	97 (R)
Nombre d'hapax	300 (R)	26 (S)	139 (R)

Pour ce qui est de l'utilisation des hapax, il semble contre toutes attentes que ce soit *AgoraVox* qui en use le moins. Mais cette mesure est à prendre avec beaucoup de précaution en attendant une évaluation plus fine de la nature de ces hapax.

6. Conclusion

Nous avons présenté dans cet article une méthodologie d'analyse de corpus permettant de contraster différentes modalités d'expression de la subjectivité. Dans cette méthodologie, le corpus joue un rôle très important puisque c'est sa constitution même qui va permettre d'avancer dans la description de l'influence des modalités de rédaction sur l'expression de la subjectivité.

Cet article relate la première étape de notre étude qui consiste à valider l'idée que notre corpus représente bien trois modalités d'expression de la subjectivité, selon l'échelle proposée : *Le Soir* < *Rue89*< *AgoraVox*, du moins au plus subjectif. Les premiers résultats obtenus nous confortent dans cette validation. En effet, le large panel d'indices choisis pour mesurer la subjectivité des trois médias permet de les distinguer efficacement. D'un côté, comme attendu, la presse traditionnelle se distingue des nouveaux médias en ligne, avec une fréquence comparativement très faible d'indices de subjectivité. D'un autre côté et contre toute attente, il semble que l'implication de

l’auteur soit plus ouverte dans *Rue89* que dans *AgoraVox* si l'on considère le nombre significativement plus important des pronoms de 1re personne et des constructions clivées dans les articles de *Rue89*. Enfin, le caractère plus professionnel de *Le Soir* et *Rue89* est en partie validé par les LL calculés pour les indices d'organisation discursive en général (tableau 4), interprétés comme signe d'une organisation discursive plus construite et plus signalée à la surface du texte.

Nous avons choisi d'affronter « en masse » l'expression de la subjectivité en prenant en considération un très large panel d'éléments potentiellement subjectifs sans recourir à des ressources particulières. À l'heure de publication de cet article, des lexiques d'éléments potentiellement subjectifs sont à disposition et il serait tout à fait pertinent d'observer ce que donnerait la projection de tels lexiques sur notre corpus. De même, d'autres études contrastives pourraient être réalisées à partir de notre corpus de presse. C'est dans cet objectif que les sous-corpus *Rue89* et *AgoraVox* seront diffusés prochainement, afin de permettre le partage de telles ressources, méthodes et analyses.

Les caractéristiques linguistiques de nos trois médias vont maintenant servir de base à des analyses plus fines et davantage qualitatives pour décrire en précision les différentes modalités d'expression de la subjectivité selon les différents processus de rédaction que représentent nos trois média, en cherchant particulièrement à évaluer si la subjectivité varie en terme de fréquence et/ou de mode d'expression, de choix lexicaux, syntaxiques et/ou discursifs.

Références bibliographiques

Baayen H. (2001). *Word Frequency Distributions*, Dordrecht : Kluwer Academic.

Baayen H. & Sproat R. (1996). « Estimating lexical priors for low-frequency morphologically ambiguous forms », *Computational Linguistics* 22(2) : 155-166.

Bednarek M. (2006). *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*, London/New-York : Continuum.

Biber D., Johansson S., Leech G., Conrad S. & Finegan E. (1999). *Longman Grammar of Spoken and Written English*, London: Longman.

Bourigault D. (2007). *Un Analyseur Syntaxique Opérationnel : SYNTEX*, Ph.D. dissertation, Mémoire d'HDR en Sciences du Langage, CLLE-ERSS, Toulouse, France.

Charaudeau P. (2006) « Discours journalistique et positionnements énonciatifs. Frontières et dérives », *Semen* 22.

Charles M. (2003). « 'This mystery...': A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines », *Journal of English for Academic Purposes* 2 : 313-326.

Charles M. (2006). « Construction of stance in reporting clauses: A cross-disciplinary study of theses », *Applied Linguistics* 27(3) : 492-518.

Conrad S. & Biber D. (2000), *Evaluation in text. Authorial ttance and the construction of discourse*, Oxford: Oxford University Press.

Degand, A. (2009). La presse en crise conventionnelle. État des lieux de l'intégration du journalisme en ligne en Belgique francophone, Colloque *Médias 09*, ISIM, Aix-en-Provence, 16-17 décembre 2009, URL http://www.medias09.univ-cezanne.fr/programme/a8/degand.html.

Dugué B. (2008). « Rue89 et AgoraVox sont alternatifs, mais sont-ils différents dans l'esprit ? », *AgoraVox* 7 janvier 2008, URL http://www.AgoraVox.fr/actualites/medias/article/Rue89-et-AgoraVox-sont-alternatifs-33894.

Halliday, M. (1985). *An Introduction to Functional Grammar*. London, Arnold.

Harwood N. (2005a). « 'Nowhere has anyone attempted...in this article I aim to do just that': A corpus-based study of self-promotional I and we in academic writing across four disciplines », *Journal of Pragmatics* 37 : 1207-1231.

Harwood N. (2005b). « 'We do not seem to have a theory...the theory I present here attempts to fill this gap': Inclusive and exclusive pronouns in academic writing », *Applied Linguistics* 26(3) : 343-375.

Ho-*Dac,L-M. (2007). La position initiale dans l'organisation du discours : Une exploration en corpus*, Thèse de doctorat. Université Toulouse-le Mirail.

Ho-Dac L.-M. (2010). « An exploratory data-driven analysis for describing discourse organization », in A. Sánchez & M. Almela (eds.) *A Mosaic of Corpus Linguistics. Selected Approaches*, Frankfurt/Berlin: Peter Lang, 79-100.

Hyland K. (1994). « Hedging in academic writing and eap textbooks », *English for Specific Purposes* 13 : 239-256.

Hyland K. (1996).. « Writing without conviction? Hedging in science research articles », *Applied Linguistics* 17 : 433-454.

Hyland K. (1998). « Persuasion and context: The pragmatics of academic metadiscourse », *Journal of Pragmatics* 30 : 437-455.

Hyland K. (2005). « Stance and engagement: A modal of interaction in academic discourse », *Discourse Studies* 7(2) : 173-192.

Hyland K. & Tse P. (2004). « Metadiscourse in academic writing: A reappraisal », *Applied Linguistics* 25(2) : 156-177.

Kuo C.H. (1999). « The use of personal pronouns: Role relationships in scientific journal articles », *English for Specific Purposes* 18(2) : 121-138.

Lakoff G. (1972).. « Hedges: A study in meaning criteria and the logic of fuzzy concepts », *in* P. Peranteau, J. Levi, and G. Phares (eds.) *Papers from the Eighth Regional Meeting*, Chicago Linguistics Society (CLS 8) : 183-228.

Lambrecht K. (2001). *Language typology and language universals*, Berlin/New York: Mouton de Gruyter.

Lyons J. (1982). « Deixis and subjectivity: Loquor ergo sum », in R. Jarvella and W. Klein (eds.), *Speech, place and*

action: Studies in deixis and related topics. Chichester/New York, 101-124.

Quirk R., Greenbaum S., Leech G. & Svartvik J. (1985). *A comprehensive grammar of the English language*, London: Longman.

Rayson P. & Garside R. (2000). *Comparing corpora using frequency profiling*, Proceedings of the workshop on Comparing Corpora, held in conjunction with ACL 2000. 1-8 October 2000, Hong Kong, 1-6.

Stubbs M. (1986). « A matter of prolonged fieldwork: Notes towards a modal grammar of English », *Applied Linguistics* 7 : 1-25.

Thompson G. & Hunston S. (2000). *Evaluation in text. Authorial stance and construction of discourse*, Oxford: Oxford University Press.

Traugott E. (1995). « Subjectification in grammaticalisation », in D. Stein and S. Wright (eds.), *Subjectivity and subjectivication*, Cambridge: CUP, 31-54.

Weeber M., Vos R. & Baayen R.H. (2000). « Extracting the lowest-frequency words: Pitfalls and possibilities », *Computational Linguistics* 26(3) : 301-318.

White P. (2001). *Appraisal Outline*, [En ligne] URL : www.grammatics.com/appraisal.

Wiebe J., Wilson T., Bruce R., Bell M. & Martin M. (2004). « Learning subjective language », *Computational Linguistics* 30(3) : 277-308.