SHS Web of Conferences **20**, 01015 (2015)
DOI: 10.1051/shsconf/20152001015

Analyse d'un corpus multilingue : visualisations textométriques des convergences et divergences dans l'écriture journalistique

Analysis of a multilingual corpus: textometric visualizations of similarities and differences in journalistic writing

Mariola Moreno, Pascal Marchand et Pierre Ratinaud[1,a]

[1]Université de Toulouse ; Laboratoire d'études et de recherches appliquées en sciences sociales (LERASS) ; 115B route de Narbonne ; BP 67701 ; F-31077 Toulouse Cedex 9 ; France

Résumé. Le traitement de la crise économique par les agences de presse française (AFP) et espagnole (EFE) pose la question de la comparaison des corpus multilingues. Au premier abord, les analyses montrent une macrostructure thématique similaire dans les deux corpus. Mais une analyse plus approfondie, associant une classification du vocabulaire et une comparaison interprétative des structures lexicales, révèle une focalisation différente : lorsque l'AFP établit une distinction entre les aspects économiques (contexte mondial) et politiques (contexte national), l'EFE l'aborde davantage comme un problème politique en lien direct avec l'économie nationale et l'Europe.

Abstract. The economic crisis treatment by French (AFP) and Spanish (EFE) news raises the question about the multilingual corpus comparison. In a first approach, the analyses show a similar thematic macrostructure on both corpuses. However a deeper analysis, combining a classification of vocabulary and an interpretive comparison of lexical structures, reveals a different focusing: while the AFP makes a distinction between the economic aspects (global context) and political ones (national context), the EFE takes it up more strenuously as a political issue directly related to the national economy and Europe.

La façon dont la crise économique est traitée par les principales agences de presse en France et en Espagne, permet d'appréhender la genèse et la diffusion de représentations sociales dans le discours médiatique et d'en repérer les convergences et divergences de contenu, selon le contexte social, économique, politique et culturel dans lequel il est produit. Il s'agit donc d'aborder la signification au travers de l'organisation des formes lexicales et des relations qu'elles entretiennent pour fournir une représentation cohérente. L'analyse de l'activité de construction de schémas d'interprétation du monde social par les médias se fait ainsi par la mobilisation de *cadres* dans les discours médiatiques (Gamson et Modigliani, 1989 [7] ; Iyengar, 1991 [8] ; Entman, 1993 [2] ; Marty, 2010 [13] pour une revue de question) dont nous posons l'hypothèse qu'ils sont organisés lexicalement. La textométrie (employée ici dans une acception similaire à l'analyse statistique de données textuelles, la lexicométrie ou la logométrie), propose alors des procédures de tris et de calculs statistiques pour l'étude de gros corpus textuels numérisés (Lebart & Salem, 1994 [?]).

La question qui se pose à nous, est celle de la comparaison de corpus qui, pour mobiliser une même thématique, sont indépendants dans leurs conditions de production. C'est le cas lorsque l'on veut comparer, pour un même objet socio-représentationnel, la façon dont il est organisé par les médias, dans des discussions en ligne, dans des conversations spontanées (*focus-groups*), dans des

[a]Auteur de correspondance : mariolamorenocalvo@gmail.com
Ce travail a été réalisé dans le cadre du LABEX SMS portant la référence ANR-11-LABX-0066"

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20152001015

entretiens, etc. Compulser ces différentes traces verbales en un même corpus présenterait le risque de produire des différences artefactuelles davantage liées à leurs conditions de production et à leurs normes discursives (genres, styles) qu'aux prises de position effectives des locuteurs.

On peut noter que Marty, Marchand et Ratinaud (2013 [12]) ont proposé une méthode originale de comparaison de corpus issus de sources différentes (*i.e.* un débat en ligne et son traitement par la presse) par la méthode des *types généralisés* (Lamalle & Salem, 2002 [10]).S'agissant ici de la comparaison de discours multilingues et sur des corpus non-alignés (voir par exemple Fleury & Zimina, 2007 [6]), la structure lexicale et syntaxique des langues oblige à analyser indépendamment la distribution statistique des corpus et la méthode des *types généralisés* n'est pas utilisable de la même façon, à moins de passer par une phase de traduction dans l'une des deux langues, ce que nous avons écarté, tant d'un point de vue théorique (biais et distorsions introduits par la traduction) que méthodologique (faisabilité sur des corpus de très grande taille).

Au-delà d'un positionnement épistémologique que nous assumons, la nature de ce type de corpus rend donc indispensable une articulation qualitatif / quantitatif. La comparaison entre les analyses devra alors recourir à des procédures interprétatives et les procédures formelles devront être articulées avec des parcours plus interprétatifs, déterminants quant aux affinités possibles avec une théorie linguistique telle que la sémantique interprétative (Pincemin 2011 [15]).

1 Corpus et méthode

L'objectif étant l'étude d'une information à la source, avant qu'elle ne passe par les processus éditoriaux des médias de diffusion, on a retenu deux agences de presse. L'agence France Presse (AFP) a succédé à Havas après la seconde guerre mondiale. Elle n'a pas d'actionnaire. L'EFE a succédé à l'agence Fabra en 1939 et se présente comme la quatrième agence de presse mondiale et principale agence hispanophone. Son actionnaire majoritaire est l'Etat espagnol. Les dépêches des agences française (AFP) et espagnole (EFE) ont donc été extraites de la base de données de presse *Factiva* (Dow Jones & Company), qui collecte les contenus de plus de 31 000 sourceshttps://fr.wikipedia.org/wiki/Factiva - cite_note-1 de presse (journaux, magazines retranscriptions radio et télévision, photos, etc..) provenant de 200 pays en 25 langues.

Un premier corpus a été constitué sur la base des mots clés « crise économique / crisis económica », de janvier 2008 jusqu'à mars 2014. La question se pose de limiter les dépêches à la locution « crise économique » plutôt que la forme simple « crise ». Cette décision a été prise après avoir observé la polysémie de « crise », particulièrement dans presse française, que ce soit pour qualifier d'autres contextes (crise sociale, crise de régime…), ou des emplois plus métaphoriques (crise de nerfs…).

Une opération de nettoyage et d'harmonisation des corpus a consisté à enlever les liens hypertextes et metadonnées, ajoutés par *Factiva* en début et fin de chaque dépêche. Il a fallu également supprimer les doublons et coder les métadonnées utiles à notre analyse (essentiellement la date, selon trois formats : en années, mois et jours).

Un travail a enfin dû être effectué sur le dictionnaire espagnol d'IRaMuTeQ pour rendre comparables les opérations de reconnaissance et lemmatisation des formes et locutions dans les deux langues.

Une fois ces différentes opérations réalisées, les corpus ont les caractéristiques suivantes (cf. **Tableau 1**).

La différence de taille entre les deux corpus (presque du simple au double) peut-être due à un effet de contexte (différence d'importance de la question de la crise économique dans chacun des deux pays) ou d'agenda médiatique (différence d'importance du traitement de cette question par les agences de presse). En revanche, la taille moyenne des dépêches est comparable pour les deux agences (le nombre moyen occurrences par dépêche est de 446 pour l'AFP et de 436 pour l'EFE).

Tableau 1. Caractéristiques lexicométriques des corpus « AFP » et « EFE »

	Agence France Presse (AFP)	Agence de presse espagnole (EFE)
Nombre de dépêches	17 850	32 504
Nombre de segments de texte	220 550	394 341
Nombre d'occurrences	7 962 140	14 183 078
Nombre de formes	91 920	119 069
Nombre de lemmes	66 579	94 634
Nombre de formes actives	64 438	91 922
Formes supplémentaires	2141	2665
Moyenne de formes par segment	36.1	35.9
CDH :		
Nombre de classes	11	9
Segments classés	204 770 (92.85%)	381 892 (96.84%)

2 Premiers résultats : Nuages de mots et Analyses de similitudes

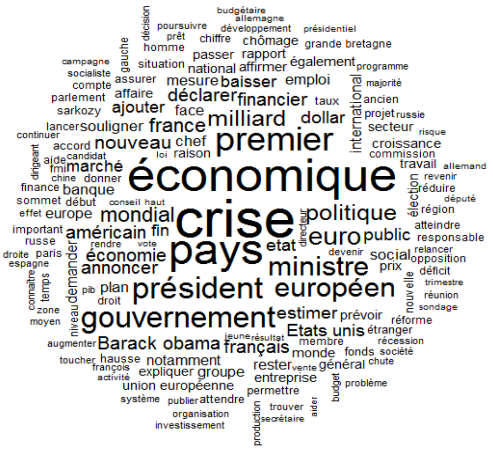

Figure 1. Nuage du corpus « AFP »

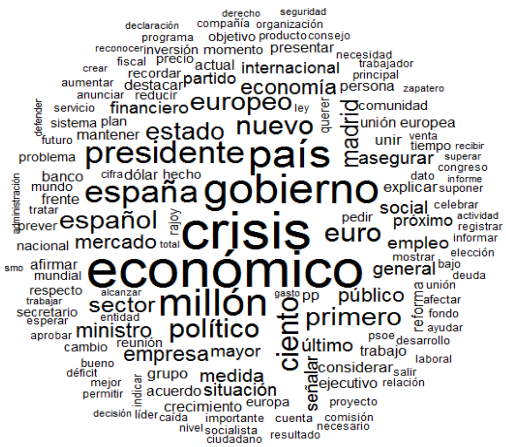

Figure 2. Nuage du corpus « EFE »

Les nuages de mots ne reposent que sur la fréquence des formes lexicales et leur information est relativement faible. Toutefois, on observe que les formes de fréquence maximale (Fmax) sont assez comparables dans les deux nuages (cf. **Figure 1** et **2**):

AFP (Fmax) : *crise, économique, pays, premier, président, gouvernement, européen, ministre, euro, politique, milliard, mondial, france, nouveau, américain, économie, financier, Barack obama, déclarer, français, annoncer, Etats unis, Etat, estimer, dollar, public, fin…*

EFE (Fmax) : *crisis, económico, gobierno, país, millón, presidente, españa, ciento, político, euro, nuevo, europeo, español, primero, estado, madrid, sector, empresa, ministro, economía, público, mercado, medida, último, general, social, asegurar, situación, financiero…*

Plus précise que les nuages de mots, l'analyse de similitude (ADS) est une technique, reposant sur la théorie des graphes, classiquement utilisée pour décrire des représentations sociales sur la base de questionnaires d'enquête (Flament, 1962 [3] ; Flament, 1981 [4] ; Vergès & Bouriche, 2001 [18]). L'objectif de la méthode est d'étudier la proximité et les relations entre les éléments d'un ensemble, sous forme d'arbres. Le nombre de liens entre deux items évoluant « comme le carré du nombre de sommets » (Flament & Rouquette, 2003 : 88 [5]), l'ADS cherche à réduire le nombre de ces liens pour aboutir à « un graphe connexe et sans cycle » (Degenne & Vergès, 1973 : 473 [1]). Plus précisément, à partir des *cliques* de trois items du corpus, l'arbre maximum est créé en éliminant les arêtes les plus faibles à l'intérieur des *cliques*. C'est l'arbre le plus simple que l'on peut obtenir, mais c'est aussi le plus lourd (en termes d'information).

L'analyse de similitude d'une matrice textuelle a été intégrée au logiciel *IRaMuTeQ* et permet de décrire des classes lexicales, des profils de spécificités ou même des corpus entiers (Marchand & Ratinaud, 2012 [11]). Pour accroître la lisibilité des graphes, on sélectionne les mots à représenter sur la base de leur fréquence et on applique un algorithme de détection des communautés pour visibiliser des champs lexicaux.

L'arbre maximum pour le corpus AFP a été construit avec les 91 mots de fréquence supérieure à 3000 (cf. **Figure 3**) :

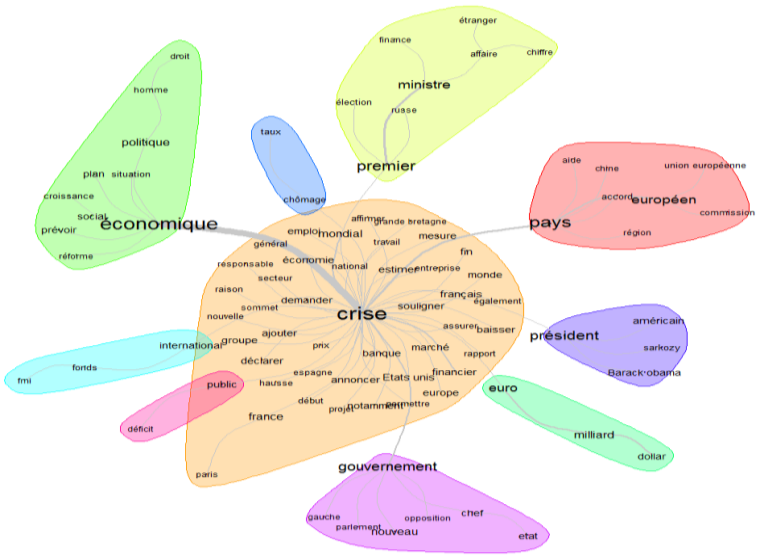

Figure 3. Graphe Analyse de Similitude « AFP ».

L'arbre maximum pour le corpus EFE a été construit avec les 110 mots de fréquence supérieure à 4500 (cf. **Figure 4**):

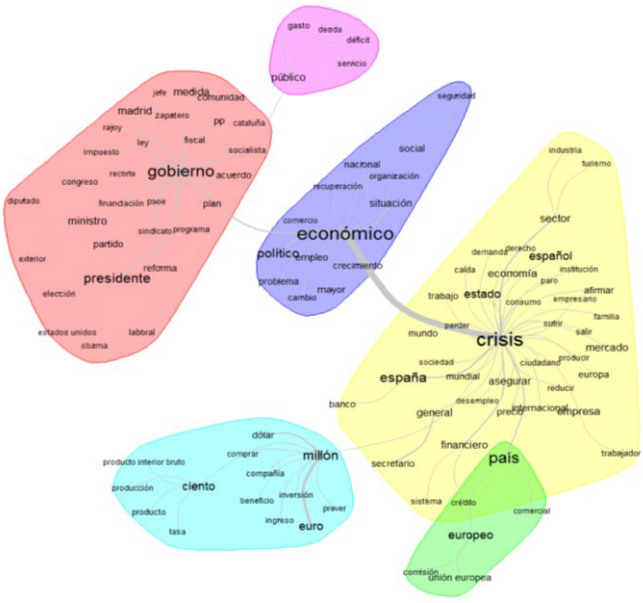

Figure 4. Graphe Analyse de Similitude « EFE ».

Les graphes de similitudes donnent essentiellement à voir les convergences entre les deux corpus. De façon prévisible, les deux graphes s'organisent autour du mot-clé « crise » qui a déterminé le corpus, et fortement connecté à l'adjectif « économique ». Mais « économique » centralise un champ lexical autour de notions plus générales *social/social, situation/situación, plan/proyecto* ou d'indicateurs (*chômage / desempleo*).

Les autres champs lexicaux définis par les relations entre formes de plus fortes fréquences sont très convergents entre les deux analyses :

- Un champ centralisé par pays/país où se connectent des considérations nationales et internationales (*Europe/europeo*)
- Un champ centralisé par la politique gouvernementale (*gouvernement/gobierno, président/presidente, ministre/ministro*)
- Un champ autour des chiffres dans un lexique plus financier (*euro/euro, dollar/dolar, ciento, millón, milliard*…)

Il semblerait donc, à partir de cette analyse macrostructurale, que les traitements de la crise économique par les agences de presse des deux pays mobilisent, peu ou prou, les mêmes cadres interprétatifs. Quel que soit le pays, un tel moment médiatique (Moirand, 2007 [14]) est décrit en termes économiques et financiers et abordé selon son impact sur la politique intérieure et les relations nationales et internationales. L'uniformisation des traitements de l'information dans un système médiatique de plus en plus mondialisé conduisant des pratiques routinières et une convergence d'analyse pourrait ainsi être évoquée. Il faut néanmoins envisager que ces traitement, pour instructifs qu'ils puissent paraître, sont limités aux formes les plus fréquentes (Fmax) et aux relations de plus forte similitude entre ces formes. Nous préférerons donc les utiliser davantage pour valider la constitution des deux corpus et la pertinence de leur comparaison : ils parlent bien de la même chose et globalement dans les mêmes termes, mais l'analyse comparative ne peut en rester là.

3 Approfondissement : classification lexicale

Nous utilisons une méthode proposée par Reinert (1983 [17]) et intégrée au logiciel *IRaMuTeQ* (Ratinaud & Marchand, 2015 [16]). Il s'agit de construire un tableau binaire croisant les formes lemmatisées et des segments de textes définis par la ponctuation forte et une taille d'une quarantaine de formes lexicales. (voir **Tableau 1**) L'algorithme de Classification Descendante Hiérarchique (CDH) permet de définir des classes lexicales, en partant du corpus global pour descendre progressivement vers les classes « terminales ». Chacune des classes peut alors être décrite selon le vocabulaire qui lui est corrélé.

La classification lexicale permet d'identifier onze classes pour l'AFP (92,85% de segments classés) et neuf pour EFE (96,84% de segments classés). Les graphiques suivants (cf. **Figure 5** et **6**) sont les dendrogrammes de classification auxquels on ajoute le vocabulaire spécifique des classes (par chi² décroissant).

Pour ce qui concerne l'AFP (**Figure 5**), une première partition distingue d'abord les considérations économiques (classes 5, 6, 7, 8, 11), qui se divisent en deux types de vocabulaire : l'économie financière (classes 7, 8 et 11 sur des indicateurs financiers, chiffrés, impliquant les échanges internationaux) et l'activité économique (classe 5 sur l'industrie et classe 6 sur les services). D'un autre côté, ce sont d'abord des considérations politiques (classes 1, 2, 9, 10) et sociales (classes 3 et 4) qui sont identifiables.

On retrouve ici les grands champs lexicaux qui émergeaient déjà des analyses de similitude mais on voit apparaître, d'une part des sous-champs (la politique, par exemple, se subdivise en considérations électorales, législatives et internationales, essentiellement américaines ici), et d'autre part des champs nouveaux, notamment autour des conséquences de la crise sur les mouvements sociaux.

Pour ce qui concerne l'EFE (**Figure 6**), on retrouve encore une structuration forte autour des considérations économiques financières (classes 4, 5, 7) qui s'opposent à des considérations politiques (classes 1, 6, 8, 9). On confirme ici également les grands champs lexicaux dessinés par l'analyse de

similitude, mais on voit émerger une dimension sociale et culturelle (classe 2) et une dimension internationale (classe 3).

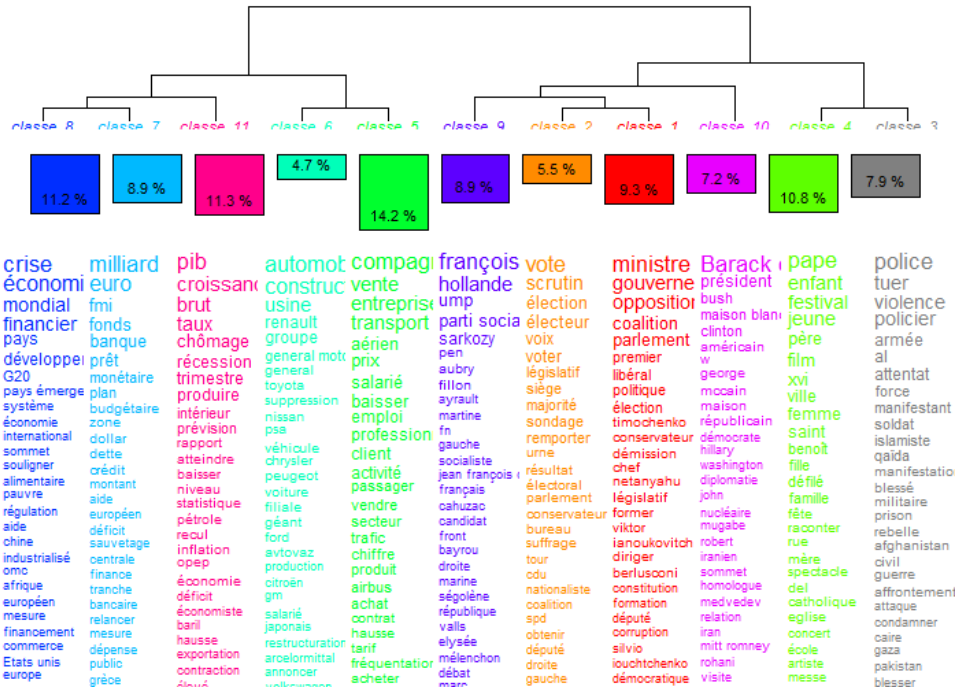

Figure 5. Graphe classes issues de la CDH du corpus « AFP ».

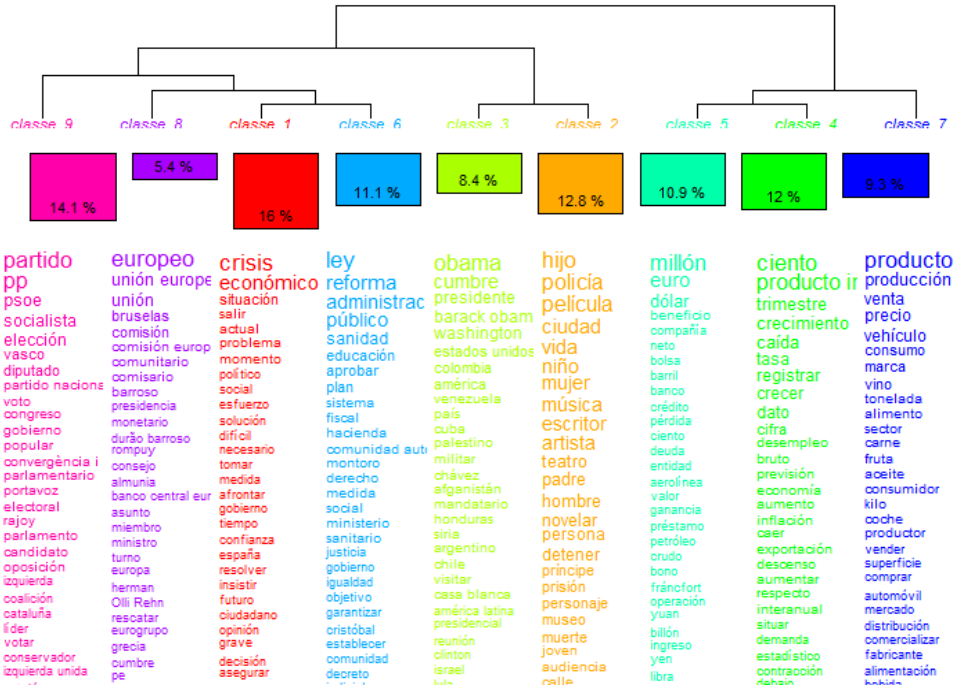

Figure 6. Graphe classes issues de la CDH du corpus « EFE ».

Ces résultats confirment donc les analyses de similitudes mais les précisent (distinctions à l'intérieur d'un registre lexical) et les complètent (apparition de nouveaux champs lexicaux). Mais la comparaison des deux dendrogrammes apporte des informations nouvelles et attirent notre attention sur les cadres interprétatifs mobilisés par les journalistes à propos de la crise économique. Cette comparaison, on l'a vu, doit recourir à une lecture plus interprétative.

4 La comparaison : une lecture interprétative

Encore une fois, les analyses des deux corpus d'agences française et espagnole convergent fortement non seulement dans le vocabulaire mobilisé, mais également dans la définition de champs lexicaux assez comparables à première vue. Pourtant, l'organisation de ces champs lexicaux n'est pas la même dans chacun des corpus.

La dimension économique est, logiquement, fortement structurante dans les deux cas. Aux trois classes financières espagnoles (4, 5, 7) semblent bien répondre trois classes françaises (7, 8, 11). En élargissant aux classes 5 et 6 du corpus français, qui proviennent de la même classe-mère, on intègre également, pour les deux pays, les effets de la crise sur l'activité économique de l'industrie et des services (classe 7 du corpus espagnol). De même, *Pétrole* et *baril* figurent dans le vocabulaire économique des dépêches des deux agences, compte tenu des fortes hausses de prix des carburants fossiles et de leur répercussion sur plusieurs aspects économiques. En revanche, les moteurs économiques ne sont pas les mêmes dans les deux pays. Dans le cas français il s'agit davantage d'une économie basée sur le grand secteur industriel de l'automobile (*Renault, Nissan, Peugeot* sont présentes dans la classe 6) et de l'aviation (transport, aérien, *Airbus*, dans la classe 5). En Espagne, la classe 7 de l'EFE cite davantage le secteur agricole (*vino, carne, aceite* …). Mais l'intérêt de l'approche textométrique est bien de porter sur des co(n)textes qui sont moins sensibles aux variations des formes particulières et permettent de retrouver une cohérence thématique y compris lorsque les formes lexicales diffèrent.

Plusieurs résultats marquent pourtant une grande différence. Le premier concerne le terme même de « crise économique » : s'il est bien corrélé à la finance dans le corpus français (classe 8), et plutôt dans un co(n)texte mondialisé, il n'est pas lié au même champ lexical dans le corpus espagnol, où on va le retrouver avec un vocabulaire plus général et plus socio-politique (classe 1). Il est plus intéressant encore de noter que cette classe 1 de l'EFE partage la même classe-mère que les considérations sur les lois de réforme engendrées par la crise (classe 6). Le terme même de « crise économique » n'est donc pas lexicalisé de la même façon en France et en Espagne : davantage liée aux décisions politiques pour la presse espagnole, à la finance mondialisée pour la presse française. C'est peut-être la raison pour laquelle on trouve, dans l'analyse de l'EFE, une classe dédiée aux mesures d'austérité qu'ont adoptées les deux gouvernements espagnols pour sortir de la crise, ce qui n'est pas le cas dans l'analyse du corpus AFP.

La question de la politique internationale, ensuite, est présente dans les deux corpus, mais à y regarder de plus près, il ne s'agit pas des mêmes positionnements. Dans le corpus espagnol, deux lexiques sont concernés (classes 3 et 8). Un lien est établi entre Europe et crise (classe 8 : Commission européenne, Bruxelles, Barroso, Rumpoy…). Ce n'est pas étonnant pour un pays qui a obtenu des aides économiques du Fonds Monétaire International (FMI) et de la Banque Centrale Européenne (BCE) et qui doit intégrer les obligations que ces organismes lui imposent. D'autre part, on évoque (classe 3) les Etats-Unis (Obama, Estados Unidos, Washington), l'Amérique du Sud (Venezuela, Cuba, Colombie, Argentine, Chili) et les pays en crise (Afghanistan, Palestine). En France, la crise est davantage traitée comme un problème mondial (classe 8 : mondial, G20, Grèce) et la question européenne est présentée comme moins centrale, plus diluée dans les divers registres lexicaux. Les Etats-Unis ont une importance comparable, mais sous l'angle des acteurs (Obama, Bush, Clinton…) et en lien avec d'autres grands acteurs du moment et d'autres crises (Russie, Iran…).

Une thématique prégnante dans les deux classifications lexicales concerne la question politique. Mais, ici encore, la politique n'est pas traitée d'une façon similaire dans les deux agences de presse.

L'agence espagnole développe un point de vue plus concentré (classe 9) et plus institutionnel, donnant plus d'importance aux partis politiques (parti, PP, PSOE, socialiste). La politique espagnole est évoquée directement en lien avec les problématiques économiques, la responsabilité vis-à-vis des réformes.

Du côté français, au contraire, la politique et la crise économique apparaissent dans des espaces lexicaux très distincts, comme si les deux n'avaient pas de rapport. Il y a trois classes liées au politique (classes 1, 2, 9). Il y a d'abord une tendance à la *personnalisation* (classe 9) mettant en scène des individualités que l'on met en confrontation (Hollande, Sarkozy, Aubry, Fillon…). Et ce mode de traitement se retrouve également au niveau international (Obama, Bush, Clinton, McCain…). Il y a ensuite une forte référence aux sondages d'opinion (classe 2) et à la projection obligatoire de toute analyse sur les élections à venir. Pour la presse française, quel que soit le moment et quel que soit l'objet, toute personnalité politique est immédiatement traitée comme un-e candidat-e potentiel-le dans une échéance électorale plus ou moins lointaine. Une partie de la « politique internationale » est également traitée au travers des rencontres avec nos propres dirigeants.

Enfin, des thématiques apparaissent spécifiquement liées aux pays. En France, les dépêches mobilisent des questions liées aux mouvements sociaux (classe 3 : manifestation, violence, police…), qui n'apparaissent pas aussi structurants en première analyse du corpus espagnol, et ce, malgré les mouvements des *indignados* qui ont pu marquer l'actualité espagnole, mais semblent être moins directement mis en relation avec la crise économique. En revanche, l'agence espagnole mobilise un lexique culturel (música, escrito, artista, teatro) que l'on ne trouve pas dans l'agence française.

5 Conclusion

Notre analyse propose une lecture à la fois qualitative et quantitative d'un corpus complexe, puisqu'il implique deux langues, les deux permettant d'accéder à des informations spécifiques qui enrichissent les connaissances pour décrire et expliquer la formation et la diffusion des représentations médiatiques. (Laflamme, 2007 [9]).

La construction du corpus sur la base de mots-clés le limite aux dépêches mentionnant explicitement la « crise économique », à l'exclusion d'éléments plus implicites, ce qui pourrait être envisagé comme une limite. Mais c'est au contraire dans l'explicitation des liens entre le fait de nommer la crise économique et d'autres concepts que se situe, selon nous, les différences entre les pays.

Les nuages de mots et les analyses de similitude des deux corpus, qui en révèlent la macrostructure, montrent des proximités flagrantes qui pourraient renvoyer à des cadres interprétatifs similaires. Mais la classification lexicale permet d'accéder à des espaces discursifs différents (ou organisés différemment) et ces microstructures montrent que le traitement et l'interprétation de la crise économique n'est pas similaire dans les deux agences.

Cette étude montre qu'il est nécessaire de ne pas en rester à des macrostructures interprétatives, mais d'approfondir les traitements, non seulement en multipliant les classes lexicales mais, surtout, en s'intéressant à leur organisation interne (liens entre les classes) et les reliant aux connaissances que nous avons sur les conditions de production des corpus. C'est bien un aller-retour permanent entre la production de mesures et les démarches interprétatives qui permet de révéler les processus de construction et de diffusion des représentations médiatiques.

L'ensemble de ces différences révèle la diversité qui existe dans deux pays avec une culture différente, déterminée par leur histoire. Cela justifie l'étude comparative de ces deux pays qui traversent un évènement commun, comme c'est le cas de la crise économique. Bien qu'ils ne soient pas touchés de la même manière, ils doivent faire face à ce problème global. Et le traitement médiatique de cette thématique globale va mobiliser différents points de vue et, surtout, va les organiser différemment.

Lorsque l'AFP fait un traitement lié aux questions plus économiques, l'EFE met la crise en relation avec des aspects politiques. Cette différence de traitement peut être rapportée aux différentes situations de chaque pays. On trouve ainsi, dans l'agence espagnole, une insistance plus grande sur les

mesures prises pour sortir de la crise et sur l'importance de l'Europe. En revanche, les mouvements sociaux et les questions de sécurité sont davantage traités par l'AFP que par l'EFE, même s'il y a eu des grèves nationales importantes en Espagne (le mouvement des *Indignados* n'apparaît pas lié au mot-clé « crise économique »).

Nous voudrions noter à quel point la comparaison des corpus français et espagnol fait émerger des effets qu'une analyse indépendante n'aurait pas permis d'observer. Bien au-delà d'une hypothèse interculturaliste que l'on pourrait qualifier de « faible », il s'agit d'utiliser les résultats de chacune des deux analyses pour nourrir l'interprétation de l'autre et percevoir des effets représentationnels qui dépassent ce qui est dit pour intégrer ce qui n'est pas dit ou ce qui est dit différemment, dans une organisation différente.

L'interprétation des classifications reposait ici essentiellement sur des indices statistiques (formes caractéristiques, réponses modales, cooccurrences et distributions…) et sur une lecture interprétative d'analyses indépendantes. Bien sûr, ces analyses demanderaient à être affinées et éclaircies par un recours plus précis aux contextes nationaux, mais elles permettent de commencer à décrire la façon dont les agences de presse intègrent la crise économique dans des représentations et des attentes différentes. Une prochaine étape devra consister à approfondir ces interprétations en faisant intervenir les contextes de production (chronologie, événements, histoires des pays…). Ainsi pourrons-nous mettre en relation textes, cotextes et contextes de la crise économique, ce qui constitue, pour nous, l'objectif ultime d'une approche textométrique.

6 Bibliographie

1. A. Degenne & P. Vergès, Introduction à l'analyse de similitude. *Revue française de sociologie*, **14 (4),** 471-511 (1973)
2. R.M. Entman, Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, **43(4),** 51-58 (1993)
3. C. Flament, L'analyse de similitude. *Cahiers du centre de recherche opérationnelle*, 4 (1962)
4. C. Flament C, L'Analyse de Similitude, Une Technique pour les Recherches sur les Représentations Sociales. *Cahiers de Psychologie Cognitive*, **1**, 375- 395 (1981)
5. C. Flament & M-L. Rouquette, *Anatomie des idées ordinaires, comment étudier les représentations sociales*. Paris, Armand Colin (2003)
6. S. Fleury & M. Zimina, Exploring Translation Corpora with mkAlign. *Translation Journal*, **11(1)** (2007) Disponible en ligne : http://accurapid.com/journal/39mk.htm.
7. W. Gamson & A. Modigliani, Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach. *American Journal of Sociology*, **95(1)**, 1-38 (1989)
8. S. Iyengar, *Is Anyone Responsible? How Television Frames Political Issues*. Chicago: University of Chicago Press (1991)
9. S. Lafflame, Analyses qualitatives et quantitatives : deux visions, une même science. *Nouvelles perspectives en sciences sociales : revue internationale de systémique complexe et d'études relationnelles*, **3(1)**, 141-149 (2007)
10. C. Lamalle & A. Salem, *Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels*. In Actes des sixièmes Journées d'Analyse des Données Textuelles, Saint-Malo : 403-12 (2002)
12. A. Lebart & A. Salem. *Statistique textuelle*. Dunod, Paris (1994)
11. P. Marchand & P. Ratinaud, *L'analyse de similitude appliquée aux corpus textuels : les primaires socialistes pour l'élection présidentielle française* (septembre-octobre 2011). Lexicometrica JADT (2012) (en ligne sur http://lexicometrica.univ-paris3.fr/jadt/jadt2012/)
12. E. Marty, P. Marchand & P. Ratinaud, Les médias et l'opinion- Eléments théoriques et méthodologiques pour une analyse du débat sur l'identité´ nationale. *Bulletin de Méthodologie Sociologique*, **117**, 46–60 (2013)

13. E. Marty, *Journalismes, discours et publics : une approche comparative de trois types de presse, de la production à la réception de l'information*. Thèse NR en SIC de l'Université de Toulouse (2010)
14. S. Moirand, *Les discours de la presse quotidienne. Observer, analyser, comprendre*. Paris, Presses Universitaires de France (2007)
15. B. Pincemin, Sémantique interprétative et textométrie – Version abrégée. *Corpus*, **10**, 259-269 (2011)
16. P. Ratinaud & P. Marchand, Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). *Mots*. Les langages du politique, **108**, 57-77 (2015)
17. M. Reinert, Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les Cahiers de l'analyse des données*, **VIII(2)**, 187-198 (1983)
18. P. Vergès & B. Bouriche, L'analyse des données par les graphes de similitude. *Sciences Humaines* (2001) Disponible en ligne : http://www.scienceshumaines.com/textesInedits/Bouriche.pdf