Traitement de données issues d'un corpus écrit multilingue. Approche agile pour l'analyse du discours eurorégional

Marie-Hélène Hermand, Emmanuel Thouraud

Traitement de données issues d'un corpus écrit multilingue. Approche agile pour l'analyse du discours eurorégional

Data processing model on a multilingual written corpus. Agile approach to analyze the euroregional discourse

Marie-Hélène Hermand[1], Emmanuel Thouraud[2]

[1]*marie-helene.hermand@ulb.ac.be*
[2]*emmanuel.thouraud@corde.org*

Résumé. L'article présente quelques éléments de la procédure mise en place pour traiter un corpus écrit comportant 617 textes (près de 500 000 mots) relatifs aux eurorégions. Complexe et hétérogène à plusieurs titres (technique, linguistique, éditorial, générique, énonciatif), le corpus pose la difficulté majeure de l'appréhension de données multilingues (français, italien, espagnol, anglais, allemand, néerlandais). Sa manipulation a nécessité une réflexion adaptée et une démarche de modélisation que nous qualifions d'« agile » en raison de son caractère souple et itératif. La plateforme d'analyse élaborée permet de disposer de résultats utiles à l'analyse qualitative ultérieure du discours eurorégional. Elle articule un logiciel d'analyse morphosyntaxique éprouvé (TreeTagger) à des programmes (Perl) et à une base de données (SQLite) développés pour optimiser les requêtes multilingues simultanées et l'exportation automatique des résultats. Les fonctionnalités liées à la localisation contextualisée de mots-pivots, au recueil de dénominations et à la détection de segments répétés nous servent ici de guides pour exprimer les besoins de la recherche, les problèmes rencontrés et les solutions proposées. L'analyse d'observables récurrents, à savoir les notions de *décision* et de *responsabilité*, illustre le propos.

Abstract. The article presents some aspects of the model applied to a written corpus of 617 texts (around 500 000 words) relative to the Euroregions. Complex and heterogeneous in several respects (technical, linguistic, editorial, generic, enunciative), the corpus raises the major challenge of the analysis of multilingual data (French, Italian, Spanish, English, German, Dutch). This analysis required a suitable reflection and modeling process which we call "agile" because of its flexible and iterative character. The analysis platform can provide useful results for subsequent qualitative analysis of Euroregional discourse. It combines a proven part-of-speech tagger software (TreeTagger) with Perl modules and SQLite database developed to optimize simultaneous multilingual queries and automatic export of the results. The features related to the location of contextualized words, the collection of own names and the detection of repeated segments serve as guides to express the needs of research, problems and proposed solutions. The analysis of the repeated expressions of *decision* and *responsibility* in the corpus illustrates the model.

1 Problématique

Partant du constat de la visibilité croissante des régions frontalières en Europe, nous nous intéressons aux discours en ligne suscités par et au sujet des eurorégions. Ces organisations transfrontalières, définies comme des organisations de coopération transfrontalière formées le long des frontières

européennes[a], renvoient à un concept large et à des acteurs territoriaux émergents en Europe. Historiquement concentrées le long des frontières de pays ou régions francophones et germanophones, les eurorégions incarnent de manière spontanée les initiatives de réconciliation à la fin de la seconde guerre mondiale. Elles bénéficient du soutien explicite des institutions européennes depuis le milieu des années 1970 puis d'une attention communautaire accrue à partir du milieu des années 1990 (programme Interreg et subventions de projets transfrontaliers) (Alliès, 2011 [1]). Relativement méconnues des citoyens, elles présentent la particularité de bénéficier d'une reconnaissance institutionnelle partielle et croissante au sein de l'espace public européen grâce à l'instauration de statuts juridiques mis à leur disposition en 2006 (avec le GECT) et en 2013 (avec le GEC)[b]. Engagées dans une quête de légitimité, les eurorégions suscitent des dispositifs de communication et des discours toujours plus visibles et formalisés.

Cette accumulation d'indices historiques, institutionnels et communicationnels nous invite à penser qu'il existe une « culture discursive » (Claudel *et al.*, 2013 [2]) eurorégionale indépendante des lieux (pays) d'émission et des langues d'expression. Pour la caractériser, nous avons réuni des textes qui permettent d'analyser comment la communication transfrontalière européenne passe d'un stade empirique et informel à un stade professionnel. Ces écrits de natures très diverses prolifèrent sur le web depuis le début des années 2000 et témoignent de positions discursives variées en Europe.

Les textes ont été sélectionnés selon le critère principal de présence du terme « eurorégion » ou de ses nombreuses variantes (regio, euregio, euroregio, eurorégion etc). Figurent ainsi dans notre sélection 617 textes (près de 500 000 mots) plutôt courts (près de 80% des textes du corpus contiennent entre 100 et 1000 mots). Disponibles dans l'une des six langues dont la lecture nous est accessible (français, italien, espagnol, anglais, allemand, néerlandais), ces textes ne sont pas des traductions les uns des autres. Ils forment ainsi un corpus non parallèle qui présente des discours authentiques largement représentatifs de la thématique eurorégionale : 39% des textes du corpus concernent l'Europe du Nord-Ouest[c], 21% concernent la zone Alpes-Danube[d], 17% concernent l'Europe du Sud-Ouest[e], 8% concernent l'Europe centrale et orientale[f] et 7% concernent l'Europe du Nord[g]. Constitué sur une période d'environ six mois, le corpus nous permet de considérer 42 eurorégions sur une centaine recensée aujourd'hui (Morata, 2010 [3]).

Disséminés dans des sites web très différents (sites institutionnels, sites d'entreprises, sites de presse, sites événementiels), ces textes en ligne sont soumis à des problèmes de volatilité familiers des philologues : disparitions, déplacements et privatisations des textes constituent des aléas récurrents qui ont imposé dès le départ le figement du corpus afin de disposer de son état initial tout au long de la recherche[h]. Équilibré pour représenter les principales sources d'émission des discours, le corpus est décomposé en sous-corpus (désormais SC) selon qu'il s'agit d'énonciateurs issus des mondes institutionnel (SC « institutions eurorégionales », SC « institutions européennes », SC « institutions universitaires »), économique (SC « acteurs économiques ») ou médiatique (SC « médias »). Rendu complexe par une hétérogénéité remarquable (linguistique, sémiotique, éditoriale, énonciative, technique), le corpus a rendu obligatoires un étiquetage des données permettant d'anticiper la description fine de ses contenus et la création d'une table de liaison en vue de faciliter le croisement

[a] Lexique de l'aménagement du territoire européen (Université de Paris VII-DATAR-CNRS). En ligne : http://www.ums-riate.fr/lexique/modeleterme.php?id=21 (consulté le 27 novembre 2015).

[b] GECT pour Groupement européen de coopération territoriale ; GEC pour Groupement eurorégional de coopération.

[c] Par exemple les eurorégions Enschede-Gronau, Lille-Courtrai-Tournai, Saar-Lorraine-Luxembourg, Meuse-Rhin, Scheldemond…

[d] Par exemple les eurorégions Alpes-Méditerranée, Bodensee, Inn-Salzbach, Insubrica, Tirol-Südtirol/Alto Adige-Trentino, Via Salina…

[e] Par exemple les eurorégions Aquitaine-Euskadi, Pyrénées-Méditerranée, Andalucia-Alentejo-Algarve, Galicia-Norte de Portugal…

[f] Par exemple les eurorégions Adriatico-Ionica, Egrensis, EuroBalkans, Viadrina…

[g] Par exemple les eurorégions Baltic, Helsinki-Tallinn, Barents, Karelia, Pomerania…

[h] Pour figer le corpus, nous avons utilisé le logiciel Evernote, bloc-note virtuel développé par la société éponyme, avec un compte Premium.

ultérieur des données. Enfin, la contrainte du multilinguisme s'est révélée la plus complexe à appréhender pour obtenir des données manipulables en vue de l'analyse qualitative du discours. C'est sur ce troisième aspect que nous proposons de nous concentrer dans cet article.

Après avoir présenté brièvement l'originalité et les contraintes du corpus, nous exposons la démarche de conception d'un modèle permettant de traiter des données en vue d'analyser l'objet construit que nous proposons d'appeler « discours eurorégional ». Dans des exemples volontairement restreints pour l'occasion, nous mettons à l'épreuve les fonctionnalités du modèle liées au traitement de trois éléments textuels utiles au repérage des traces de construction identitaire des eurorégions :

- les mots-pivots (Guilhaumou, 2006 [4]) ou plus largement les items (lemmes, expressions) (Pincemin, 2007 [5]) dont nous souhaitons repérer les occurrences contextualisées ;
- les dénominations, qui visent « l'institution entre un objet et un signe X d'une association référentielle durable » (Kleiber, 1984 [6]) ;
- les segments répétés, définis comme des « suites de formes comprises entre deux délimiteurs de séquence » (Lafon et Salem, 1983 [7]).

2 Contexte de la recherche

2.1 Pratiques discursives eurorégionales

Précisons que notre hypothèse de départ, qui suggère l'existence de régularités discursives eurorégionales indépendantes des langues, impose de reconstituer un univers de sens caractéristique de la thématique eurorégionale plutôt que de procéder systématiquement à des comparaisons linguistiques en fonction de telle ou telle implantation géographique. D'un point de vue théorique, nous considérons le discours eurorégional comme une « formation discursive » [Foucault, 1969 [8]), c'est-à-dire comme un ensemble de règles de formation du discours eurorégional. Décrire cette formation discursive revient à identifier ce qui peut et doit être dit lorsqu'il est question d'eurorégion. En nous inscrivant plus précisément dans une conception interdiscursive, c'est moins le genre discursif (message d'accueil, brochure de présentation, manuel de bonnes pratiques...) qui nous sert de repère déterminant que les ancrages idéologiques des différents émetteurs dans une conjoncture donnée (Haroche *et al.*, 1971 [9]) et l'interdépendance avec des discours transverses (Marandin, 1979 [10] ; Pêcheux, 1983 [11]).

L'étiquetage préalable des textes s'est avéré indispensable pour décrire le millefeuille énonciatif complexe qui entre en scène aux quatre coins de l'Europe lorsqu'il est question d'eurorégions : les Présidents et représentants officiels des eurorégions, les institutions européennes (Commission, Parlement, Conseil), les universités, les *clusters* économiques transfrontaliers, les organisations professionnelles (chambres de commerce et d'industrie, chambres d'arts et métiers), les sociétés d'audit et les médias (locaux, régionaux, nationaux, européens) mêlent en effet leurs voix pour présenter, expliquer, juger, défendre ou critiquer les eurorégions[i].

2.2 Identification des besoins

Pour identifier les positions des divers énonciateurs, les données recueillies doivent être présentées par sous-corpus émetteurs (plutôt que par langues) et s'appuyer sur des « observables » tangibles [Moirand, 2007 [12]) immédiatement recontextualisés dans les textes intégraux. D'ampleur trop vaste pour un traitement manuel, le corpus nécessite un outillage informatique pour le traitement. Le premier outil utilisé - TXM[j] - pour appréhender le corpus a été choisi en fonction de son approche textométrique plutôt que lexicométrique (Pincemin, 2011 [13] ; Rastier, 2011 [14]). La possibilité de

[i] L'articulation théorique et énonciative du corpus a fait l'objet d'une communication au colloque « Textes et discours en confrontation dans l'espace européen. Pour un renouvellement épistémologique et heuristique » organisé en septembre 2015 par le CREM-Université de Lorraine.

[j] Logiciel mis à disposition par l'équipe du projet ANR Textométrie de l'ENS Lyon.

connecter TXM à l'analyseur morphosyntaxique multilingue TreeTagger[k] a également été déterminante pour interroger tour à tour chacune des six partitions linguistiques du corpus. Ce couple d'outils nous a permis :

- d'aboutir à une première vision du corpus par l'identification des lemmes les plus fréquents, la recherche de mots-pivots et de leurs cotextes, l'extraction d'adjectifs, d'adverbes ou de verbes ;
- de tester nos premières hypothèses de recherche en manipulant l'analyseur TreeTagger dans l'interface de TXM.

Au fil des hypothèses à tester, trois problèmes techniques majeurs ont émergé sans que nous réussissions à les surmonter avec le couple TXM-TreeTagger dans un délai raisonnable par rapport à notre objectif :

- les interrogations à formuler successivement dans les six partitions linguistiques du corpus sont chronophages, notamment en raison de la manipulation indépendante des grammaires de requêtes spécifiques aux différentes langues ;
- l'exportation des résultats obtenus dans les six langues nous pose des problèmes de reconnaissance de caractères (notamment en allemand) ;
- l'exportation des résultats dans des tableurs doit se faire langue par langue et se révèle chronophage.

Pour optimiser les requêtes multilingues et faciliter l'exportation automatique des résultats dans des tableurs aux possibilités de tris adéquates par rapport à nos objectifs, la modélisation d'une démarche de traitement s'est avérée nécessaire.

3 Modélisation agile pour le traitement des données

3.1 Modélisation : choix du principe agile

La modélisation implique de nous imposer des contraintes de cohérence lors des phases de préparation et de traitement des données issues de notre corpus complexe. Idéalement, elle vise la construction d'une démarche et l'obtention de résultats reproductibles pour analyser d'autres corpus multilingues au profil proche du nôtre. Afin de ne pas substituer les moyens (l'outillage informatique) à l'objectif de la recherche (la caractérisation de la formation discursive eurorégionale), nous avons opté pour le recours très ponctuel au développement informatique, à la double condition qu'il s'appuie sur un logiciel d'analyse morphosyntaxique éprouvé (TreeTagger en l'occurrence) et qu'il soit évolutif en fonction des besoins de la recherche. Dans ce souci d'adaptation permanente qui rappelle la logique du développement agile (Beck, 2001 [15]) – dont les principales caractéristiques sont la souplesse et l'itération –, nous avons fixé un cadre de contraintes restreint mais non négociable. Il s'agit d'être en mesure :

- d'effectuer des tests d'hypothèses de recherche sur l'ensemble du corpus multilingue autant de fois que nécessaire (par exemple sur les mots/expressions-pivots, les catégories morphosyntaxiques et les segments répétés) ;
- de trier les résultats en fonction des sous-corpus émetteurs (et pas seulement en fonction des langues) ;
- de situer immédiatement les occurrences dans leur contexte et de les attribuer à leur texte d'origine, quels qu'en soient la langue d'expression et le sous-corpus d'appartenance.

3.2 Besoins spécifiques : optimisation des requêtes et des exportations

Pour illustrer quelques fonctionnalités de la surcouche logicielle développée pour appréhender le corpus, nous prenons pour point de départ notre besoin de caractériser la dynamique d'affirmation identitaire des eurorégions. Parmi les données utiles, la détection et la localisation de la présence

[k] Mis à disposition par l'université de Stuttgart.

eurorégionale dans l'ensemble du corpus sont primordiales pour évaluer le positionnement eurorégional par rapport aux autres entités géopolitiques présentes en discours : l'Europe, l'État, la région, la ville. Puisque nous nous intéressons à des territoires impulsés par les instances européennes, soumis aux autorités nationales et liés aux régions qui les composent ainsi qu'aux villes qui les incarnent, il est incontournable de nous donner les moyens d'estimer si et dans quelle mesure le référent eurorégional parvient à s'imposer dans le corpus.

Dans cette optique, la localisation et la quantification des dénominations eurorégionales nous intéressent particulièrement, c'est-à-dire les noms propres (dénominations ordinaires) attribués par convention et les noms communs (dénominations métalinguistiques) utilisés de façon durable pour toute entité qui répond à la définition large que nous avons retenue des « eurorégions », à savoir :

> *des organisations européennes de coopération transfrontalière et transnationale, plus ou moins structurées, regroupant des autorités territoriales allant en général de la commune à la région ou à leurs équivalents, associées pour la réalisation d'actions et d'objectifs communs, en fonction d'intérêts partagés et dans le cadre de 'territoires de projets'.* (Perrin, 2013 [16])

Sans présenter le détail des résultats obtenus, nous décrivons ci-dessous quelques étapes de traitement mises en place à partir de trois fonctionnalités classiques offertes en textométrie et utiles à l'analyse du discours, à savoir :

- la détection et la contextualisation de mots-pivots, indispensables aux tests multiples d'hypothèses de recherche ;
- le recueil des noms propres, susceptible de faciliter la détection du référent eurorégional ;
- la détection de segments répétés, utiles à l'examen de formes figées éventuellement héritées d'autres discours.

Il convient de préciser l'importance de deux étapes préalables à l'élaboration de notre modèle :

- *étape préalable n°1* : nous exportons manuellement les textes du corpus, initialement figés dans un bloc-notes virtuel (Evernote), en 617 fichiers textes (au format UTF8) afin de rendre possible leur manipulation ultérieure ;
- *étape préalable n°2* : nous créons des programmes Shell qui transforment ces 617 fichiers textes UTF8 en autant de fichiers analysés par le logiciel TreeTagger afin de bénéficier d'un étiquetage morphosyntaxique multilingue éprouvé (reconnaissance automatique des noms propres, des noms communs, des verbes, des adjectifs, des adverbes…)[1].

3.2.1 Localisation et contextualisation de mots-pivots

En amont de tout traitement informatique, les premières lectures linéaires du corpus ont abouti à deux constats. Le premier concerne l'importante instabilité de la dénomination *eurorégion.* Avec ou sans majuscule (au début ou à l'intérieur du mot) et déclinées dans toutes les langues considérées, les occurrences varient : *eurorégion, euroregio, euregio, europaregion*… Le second constat pointe la prolifération de mots nouveaux ou de constructions comprenant *euro* et *région* : *eurocampus, euroMOTNetz, euroregionenews, euroBIOrregión*…

Pour relever précisément ces éléments dans l'ensemble du corpus, nous développons :

- un programme Perl qui crée un fichier de résultats (Excel) listant toutes les occurrences comprenant *eur* entourées de leurs cotextes gauche et droit. Après un nettoyage manuel, le résultat aboutit à un recensement qui contextualise et comptabilise la diversité des dénominations rencontrées. En nous limitant ici aux variantes du mot *eurorégion* apparaissant plus de 5 fois, voici un extrait du recensement obtenu **(Tableau 1)**.

[1] TreeTagger a été choisi car il était livré avec des modèles pour toutes les langues mobilisées dans notre corpus.

Tableau 1. Extrait (hors contextualisation) du recensement des variantes du mot *eurorégion* rencontrées dans l'ensemble du corpus[m].

Sous-corpus « institutions eurorégionales »	**Sous-corpus « acteurs économiques »**	**Sous-corpus « médias »**
(fr) Eurorégion (87), Euregio (21), eurorégion (6) **(it)** Euroregione (97), Euregio (20), euroregione (6) **(es)** Eurorregión (125), eurorregiones (9), Euroregión (5) **(en)** Euregion (67), Euroregion (50), **(de)** EUREGIO (105), EURORÉGION (40), Euregio (40), Euroregion (39), Europaregion (25), EuRegio (17), Euregios (17) **(nl)** Euregio (48), euregio (8)	**(fr)** Eurorégion (14), EuroRégion (5) **(it)** Euregio (111), Euroregione (89) **(es)** Eurorregión (83), Eurorregión (7) **(en)** Euregio (9), Euroregion (3) **(de)** Euregio (41) **(nl)** Euregio (22)	**(fr)** Eurorégion (129), Euregio (19), euro-région (7), eurorégions (6) **(it)** Euroregione (130), euroregione (14), Euregio (7) **(es)** Eurorregión (81), eurorregión (26) **(en)** Euroregion (18), Euroregions (13), Euro-region (10), Euregio (5) **(de)** Euregion (40), Europaregion (16), Euroregion (9) **(nl)** Euregio (44)
Sous-corpus « institutions académiques »	**Sous-corpus « institutions européennes »***	
(fr) Eurorégion (50), Eurorégions (9) **(it)** Euroregione (23), Euregio (7), EUREGIO (6) **(es)** Eurorregión (23), Euroregión (7) **(en)** Euroregion (11) **(de)** EUREGIO (28), Euregio (6), Euregio (1), euregio (1), Euroregion (1) **(nl)** Euregio (47)	**(fr)** eurorégions (31), Eurorégion (28), Eurorégions (12), eurorégion (9) * textes entrés dans le corpus en version française	

- un programme Perl qui recherche les mots comprenant *euro* ou *regio* à travers tout le corpus et en fournit la fréquence. Un extrait très réduit des occurrences recensées donne une idée des résultats obtenus pour ce qui concerne le sous-corpus constitué de textes médiatiques (**Tableaux 2 et 3**).

Tableau 2. Extrait (hors contextualisation) des occurrences comportant *euro* (tri par fréquences relatives)

Composant recherché	**Langue**	**Préfixe**	**Suffixe**	**Fréquences relatives**
euro	it	-	regione	114
euro	fr	-	péen	98
euro	fr	-	région	93
euro	es	-	rregión	86
euro	en	-	pean	80
euro	en	-	pe	72
euro	it	-	peo	67

[m] Dans le tableau et dans la suite de l'article, nous abrégeons : fr=français, it=italien, es=espagnol, en=anglais, de=allemand, nl=néerlandais.

Tableau 3. Extrait des occurrences comportant *regio* (tri par fréquences relatives)

Composant recherché	Langue	Préfixe	Suffixe	Fréquences relatives
regio	it	Euro	ne	104
regio	de	-	n	54
regio	nl	Eu		36
regio	it	-	nale	24
regio	de	Groß	n	19
regio	de	Metropol	n	13
regio	de	Europa	n	13
regio	it	Macro	ne	11
regio	en	Euro	ns	11

L'ensemble des résultats fournit des données utiles :

- à l'analyse de la stabilité de la dénomination eurorégionale à partir du recensement des variantes attestées du néologisme *eurorégion*, mot-valise formé de la troncation du mot *Euro[pe]* associée au suffixe *région* ;
- à l'analyse de la créativité lexicale du corpus eurorégional à partir des mots comprenant *euro* ou *regio*, en vue de la situer par rapport à la tradition eurolectale et/ou régiolectale (Raus, 2014 [17] ; Bozhinova, 2011 [18] ; Goffin, 1994 [19]).

3.2.2 Le recueil des noms propres

La détection du référent eurorégional passe en outre par la reconnaissance des noms propres (désormais NP) attribués par convention aux eurorégions. Trois phases sont nécessaires afin de disposer des données utiles :

- *étape NP n°1* : nous développons des programmes Perl qui chargent les 617 fichiers analysés par TreeTagger dans une base de données (désormais BDD SQLite) ;
- *étape NP n°2* : nous développons des programmes Perl qui, à partir de la BDD SQLite, créent des fichiers de résultats (au format Excel) permettant la quantification et le tri des données selon de nombreux critères (sous-corpus émetteurs, genres discursifs, numéros d'identifiant des textes, langues...) ;
- *étape NP n°3* : en nous concentrant sur la reconnaissance morphosyntaxique des NP opérée par TreeTagger (lors de l'*étape préalable n°2*, décrite ci-dessus dans le paragraphe 3.2) et exportée (lors de l'*étape NP n°2*), nous procédons au nettoyage manuel des erreurs éventuelles de reconnaissance commises par TreeTagger et obtenons 3 908 occurrences pertinentes de NP. Nous les qualifions manuellement selon qu'il s'agit d'un nom d'eurorégion ou du nom *Europe*, d'un nom de pays, de région, de ville ou encore de délimitation naturelle (NP de mer, de montagne, de fleuve).

Comme en témoigne le graphe en radar (**Figure 1**), les données récoltées facilitent la détection des sous-corpus les plus enclins à mobiliser (SC institutions eurorégionales) ou à exclure (SC institutions européennes) la dénomination eurorégionale. Elles permettent aussi la détection des lieux de concurrence entre les dénominations eurorégionales et celles des autres entités géopolitiques (SC médias, acteurs économiques, institutions universitaires).

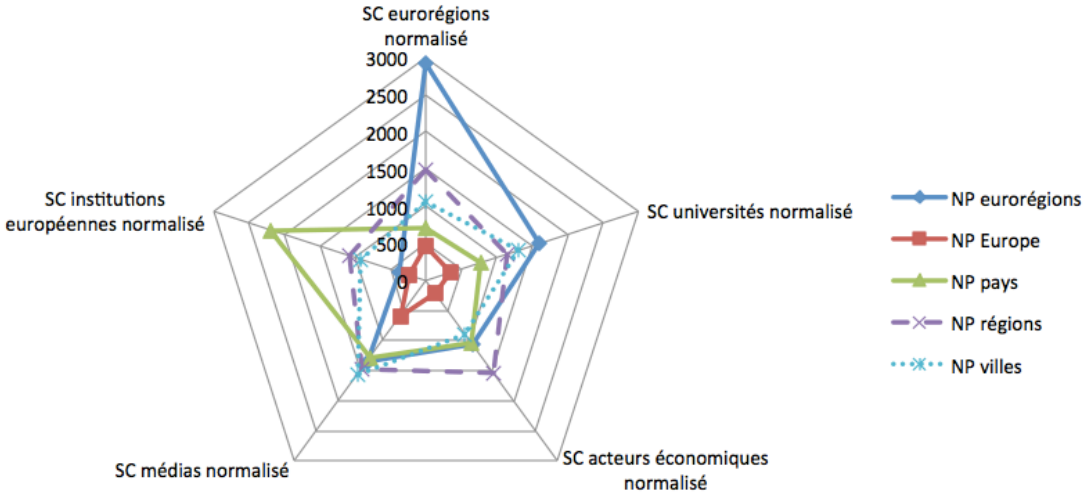

Figure 1. Fréquences relatives des noms propres (NP) détectés par sous-corpus (SC).

3.2.3 La détection des segments répétés

Nous souhaitons considérer les occurrences de NP d'eurorégions recensées comme des pivots dont le contexte permettra d'éclairer la démarche d'affirmation des eurorégions en discours. Nous procédons comme suit :

- *étape NP n°4* : nous développons un programme Perl capable d'effectuer une recherche (simultanément dans toutes les partitions linguistiques du corpus) de segments répétés accompagnés de leur cotexte ;
- *étape NP n°5* : nous utilisons ce programme pour chercher les cotextes gauche et droit des dénominations eurorégionales dans toutes les langues. Comme le nom d'une même eurorégion peut présenter des variantes au sein d'une même langue et être traduit dans plusieurs langues, nous avons exécuté le programme autant de fois que nécessaire (125 fois) pour retrouver toutes les occurrences. Dans l'exportation des résultats, les dénominations repérées pour une même eurorégion sont entourées de leurs cotextes (dont la présentation est réduite ici par manque de place)[n]. Nous surlignons dans l'exemple ci-dessous quelques variantes repérées pour la dénomination d'une même eurorégion (**Tableau 4**).

Tableau 4. Extrait de variantes dénominatives pour une même eurorégion, entourées de leur contexte.

Sous-corpus	Identifiant texte	Cotexte gauche	Pivot (segment répété)	Cotexte droit
carnet-1-1-fr	c_1-1_t27.txt	Créée le 28 janvier 2008, l'	***Eurométropole** Lille-Kortrijk-Tournai*	est un Groupement Européen de Coopération Territoriale.
carnet-3-1-fr	c_3-1_t37.txt	Ce projet vise à aller à la rencontre des habitants de l'	***Eurorégion** Lille-**Tournai-Kortrijk***	via ces émissions sur les actualités politiques, économiques et culturelles de cette région.
carnet-3-6-nl	c_3-6_t10.txt	Hij verwijst daarbij naar de "	***Euregio Rijsel-Kortrijk-Doornik***	", waar gemeenten en steden uit zuid-West-Vlaanderen,...

[n] La présentation de notre tableau de résultats s'inspire de l'interface mise en œuvre dans le logiciel TXM.

- *étape NP n°6* : pour une manipulation plus commode, nous développons un programme Perl qui fusionne les 125 fichiers de résultats générés à l'*étape NP n°5* en un seul fichier permettant de consulter et de trier l'ensemble des résultats.

Les données recueillies facilitent l'observation lexicographique systématique des dénominations eurorégionales afin d'y déceler des valeurs connotatives (telles que l'ostentation, la revendication ou la réminiscence) ou des variantes dans l'usage (toponymes alternatifs) auxquelles s'intéresse l'analyse du discours. Les possibilités de recherche de segments répétés permettent aussi d'évaluer la stabilité des dénominations de ces eurorégions en cours d'installation dans un panorama européen mouvant. L'on en perçoit aussi l'intérêt lorsqu'appliquées à d'autres mots du corpus, elles ouvrent la voie à la détection des « formules » en circulation dans le corpus[o] [Krieg-Planque, 2009 [20]).

3.3 Mise à l'épreuve du modèle

Nos premières analyses du corpus ont montré de nettes convergences lexicales dans les désignations et dans les descriptions des eurorégions indépendamment de leur localisation géographique (Hermand, 2014 [21]). En présentant les eurorégions comme des « espaces de coopération cohérents » ou des « territoires de références », le volet institutionnel du discours propose de les considérer comme une évolution « naturelle » et idéalisée de la configuration européenne. Cette évolution nécessite un « engagement » multiforme de tous les acteurs eurorégionaux. Observable dans l'ensemble du corpus (lexique, marqueurs énonciatifs, modalisations), cet engagement est attesté par l'entrée des acteurs eurorégionaux dans un processus d'apprentissage de leur rôle de nouveaux « modèles » européens chargés de modifier l'état mental des citoyens et d'inciter à l'action collective (Hermand, 2014 [22]). Si la politique eurorégionale menée au niveau communautaire impose un tel engagement, elle requiert aussi la mise en œuvre de décisions communes qui ne vont pas de soi. Afin d'évaluer l'appropriation de cette gouvernance coopérative par les divers locuteurs, nous partons de l'hypothèse selon laquelle l'apprentissage de la responsabilité et l'exercice de la décision sont des éléments susceptibles d'éclairer le processus d'affirmation discursive du référent eurorégional.

En privilégiant l'analyse des notions de *responsabilité* et de *décision*, nous organisons un recueil de données qui doit nous permettre, d'une part, de localiser ces deux notions dans l'ensemble du corpus et, d'autre part, d'en préciser l'étendue et les limites. L'objectif consiste à ouvrir des pistes d'analyse afin d'estimer dans quelle mesure les divers énonciateurs prennent part au processus décisionnel eurorégional.

3.3.1 Détection des données utiles à l'analyse du processus décisionnel eurorégional

Afin de disposer d'un premier repérage, nous élaborons un cadre de requêtes automatisé par regroupements de formes permettant de balayer simultanément les six partitions linguistiques du corpus en détectant les notions de :

- *responsabilité* : (fr) respons* ; (it) respons* ; (es) respons* ; (en) respons* ; (de) verantw* ; (nl) verantw* ;
- *décision* : (fr) décid*, décis* ; (it) decid*, decis* ; (es) decid*, decis* ; (en) decid*, decis* ; (de) entsch* ; (nl) besliss*.

L'exportation automatique des résultats en tableur (Excel) localise toutes les occurrences entourées de leurs cotextes (gauche et droit). Elle identifie aussi les lemmes associés, dont nous donnons la liste nettoyée ci-dessous (**Tableau 5**), et compte les items associés :

[o] À titre d'exemple, la détection des segments répétés « cohésion territoriale » ou « coopération transfrontalière » observables dans les six langues permet non seulement de tisser un « réseau sémantique » (Mayaffre et Mellet, 2002 [23]) propre au corpus eurorégional mais aussi d'établir une filiation avec des textes programmatiques européens.

Tableau 5. Lemmes associés à l'expression de la *responsabilité* et de la *décision* (fréquences absolues).

responsabilité	***décision***
(fr) responsable (31), responsabilité (15) **(it)** responsabile (7), responsabilità (7), irresponsabilità (1), responsabilizzazione (1) **(es)** responsable **(en)** responsability (18), responsible (18), response (3) **(de)** verantwortlich (10), Verantwortung (2), verantwortbewusst (1) **(nl)** verantwoordelijk (6), verantwoordelijkheid (3)	**(fr)** décider (28), décision (27), décisionnel (3), décideur (3), décisif (2), codécision (1) **(it)** decisione (21), decidere (19), autodecisione (4), decisionale (4), decisivo (3), decisore (4), deciso (1) **(es)** decidir (6), decisión (6), decisivo (1) **(en)** decision (19), decide (15), decision-making (5), decision-maker (2), decision-taking (1), decisively (1) **(de)** Entscheidung (11), entscheiden (10), entscheidend (9), Entscheider (3), Entscheidungs- prozess (2), -spielraum (2), -träger (2), -gremium (1), instanz (1), **(nl)** beslissing (4), beslissen (1)

3.3.2 Mobilisation de la responsabilité *et de la* décision

Pour localiser et évaluer la mobilisation des deux notions dans le corpus, nous lançons ensuite un programme Perl développé dans le but de calculer les statistiques de répartition des occurrences considérées au sein des différents sous-corpus, indépendamment des langues. Cette information nous semble davantage exploitable qu'une présentation en seuls rangs de fréquences, laquelle n'est techniquement possible que par langues (et donc moins utile par rapport à notre objectif) et ne révèle pas forcément la régularité (ou l'irrégularité) de la diffusion au sein des textes. Notre calcul de répartition[p] aboutit à une présentation de statistiques simples mais vérifiables grâce à un archivage automatique des requêtes :

Tableau 6. Répartition de la notion de *responsabilité* par sous-corpus (fréquences relatives).

Sous-corpus émetteur considéré	**% de textes mobilisant la notion de *responsabilité***
institutions européennes	36%
institutions académiques (universités)	13%
institutions eurorégionales	12%
acteurs économiques	10%
médias	8%

Tableau 7. Répartition de la notion de *décision* par sous-corpus (fréquences relatives).

Sous-corpus émetteur considéré	**% de textes mobilisant la notion de *décision***
institutions européennes	45%
médias	20%
institutions eurorégionales	14%
acteurs économiques	10%
institutions académiques (universités)	9%

Les résultats obtenus, dont nous avons présenté ici seulement quelques tendances, permettent d'approfondir deux principales pistes d'analyse. D'un côté, le relevé des occurrences précisément contextualisées ouvre la voie à la caractérisation de la responsabilité (qui sont les responsables mentionnés ? sur quoi portent les responsabilités eurorégionales ?) et à la description d'un processus décisionnel complexe, tantôt détenu par les eurorégions *(autodecisione)* ou collaboratif *(codécision)* et tantôt concentré entre les mains de quelques-uns (*decisionmaker, Entscheidsgremium,*

[p] Performance indicative sur un MacBook Air 13-inch, Early 2014 : 15 secondes pour obtenir et exporter l'ensemble des données multilingues du corpus relatives aux notions de *responsabilité* et de *décision.*

Entscheidsträger...). Ces données permettent d'approfondir la piste de la dispersion et de la dilution d'un processus décisionnel eurorégional par ailleurs vanté pour sa commodité et sa proximité avec le citoyen européen. D'un autre côté, les statistiques de répartition permettent d'identifier les sous-corpus émetteurs qui mobilisent plus volontiers les deux notions envisagées. Il est intéressant de constater leur plus forte mobilisation dans les textes produits par les institutions européennes au sujet des eurorégions et leur décrochage dans les textes produits par les institutions eurorégionales elles-mêmes. Ces constats invitent à se pencher notamment sur les relations de dépendance discursive instaurées entre les textes programmatiques de l'Union européenne et le corpus eurorégional (Hermand, à paraître [24]).

4 Conclusion

L'utilisation du qualificatif « agile » pour décrire notre approche évolutive du traitement du corpus eurorégional amène la question de la portabilité de cette démarche pour analyser d'autres corpus. Afin de dégager des invariants, nous faisons le point sur les caractéristiques du corpus et l'identification préalable des besoins, puis sur l'optimisation du traitement et la présentation des résultats.

Pour ce qui concerne les caractéristiques du corpus, le modèle est susceptible d'être appliqué à un corpus comparable, c'est-à-dire numérique (issu du web) et multilingue non parallèle, hétérogène du point de vue énonciatif (avec une identification possible de communautés discursives à l'origine du discours), étiqueté (selon des critères techniques, éditoriaux et discursifs) et exprimé dans des langues romanes et/ou germaniques déjà modélisées par un analyseur morphosyntaxique éprouvé (en l'occurrence TreeTagger). La démarche de recherche qui préside à l'élaboration de la méthode répond au double besoin de faciliter la manipulation d'un corpus multilingue en vue d'une analyse qualitative de discours internationaux (européens) encore peu pratiquée (Gobin et Deroubaix, 2012 [25]) et d'entamer la description de cultures discursives en évitant de la situer au niveau des communautés ethnolinguistiques (Claudel *et al*., 2013 [2]).

Pour ce qui concerne l'optimisation du traitement et la présentation des résultats, l'expérience actuelle porte sur dix-huit mois de tests itératifs et d'adaptation aux besoins de manipulation du corpus eurorégional et d'exploitation des données récoltées. Parmi les étapes automatisées qui nous semblent réutilisables, citons :

- la transformation des fichiers textes du corpus (codés en UFT8) en fichiers analysés par TreeTagger et leur chargement dans une base de données ;
- à partir d'un cadre de requêtes élaboré par le chercheur, la génération d'un premier fichier de résultats (en tableur) listant toutes les occurrences contextualisées d'une notion reconnue dans tout le corpus par TreeTagger (avec les erreurs de reconnaissance inhérentes au logiciel) ;
- l'ajout d'une colonne identifiant le lemme correspondant à chaque occurrence pour faciliter le tri manuel des résultats ;
- à partir du tri manuel, la génération d'un cadre de requêtes restreint aux seules occurrences jugées pertinentes par le chercheur ;
- la génération d'un fichier de résultats définitifs (en tableur) présenté selon le critère de la communauté discursive (un onglet par sous-corpus émetteur) et doté d'une large possibilité de tris ;
- la génération d'un fichier de statistiques de répartition de la notion considérée dans les textes, organisé par sous-corpus émetteurs ;
- l'archivage des programmes dans un logiciel de gestion de versions ;
- la documentation des requêtes formulées grâce à l'ajout d'un onglet « historique de la requête » dans chaque fichier de résultats, de manière à anticiper la constitution d'un cahier d'expériences.

L'intérêt de la démarche réside enfin dans la possibilité de développer des modules complémentaires au fil de la recherche, par exemple en vue d'améliorer l'interface d'utilisation ou la présentation des résultats (génération de graphiques) ou d'ajouter de nouveaux sous-corpus (par exemple, celui des contre-discours émis par les syndicats interrégionaux au sujet des eurorégions).

5 Références

1. P. Alliès, « La notion d'Eurorégion et sa mise en œuvre dans l'Union européenne », C.E. Pachado Amaral (éd.), *Autonomie régionale et relations internationales. Nouvelles dimensions de la gouvernance multilatérale*, Paris, L'Harmattan, pp. 245-255 (2011).
2. C. Claudel, P. von Münchow, M.P. Ribeiro, F. Pugnière-Saavedra et G. Tréguer-Felten G., *Cultures, discours, langues : nouveaux abordages*, Limoges, Lambert-Lucas, p. 35 (2013).
3. F. Morata, « Euroregions i integració europea », *Documents d'Anàlisi Geogràfica*, **56**, n°1, pp. 41-56 (2010).
4. J. Guilhaumou, *Discours et événement. L'histoire langagière des concepts*, Besançon, Presses universitaires de Franche-Comté, p. 15 (2006).
5. B. Pincemin, « Concordances et concordanciers : de l'art du bon KWAC », F. Rastier, M. Ballabriga, C. Duteil-Mougel, B. Fouquié, *XVII*ème *colloque d'Albi Langages et signification - Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation*, Juillet 2006, Albi, France. CALS-CPST, pp.33-42 (2007).
6. G. Kleiber, « Dénomination et relations dénominatives », *Langages*, **76**, pp. 77-94 (1984).
7. P. Lafon et A. Salem, « L'inventaire des segments répétés d'un texte », *Mots. Les langages du politique,* **6**, pp. 161-177 (1983).
8. M. Foucault M., *L'archéologie du savoir*, Paris, Gallimard, p. 57 (2008, éd. originale 1969).
9. C. Haroche, P. Henry, M. Pêcheux, « *La sémantique et la coupure saussurienne : langue, langage, discours* », *Langages,* **24**, pp. 93-106 (1971).
10. J.-M. Marandin, « *Problèmes d'analyse du discours. Essai de description du discours français sur la Chine* », *Langages,* **55**, pp. 17-88 (1979).
11. M. Pêcheux, « *Analyse de discours. Trois époques* », D. Maldidier (éd.), *L'Inquiétude du discours. Textes de Michel Pêcheux,* Paris, Éditions des Cendres, pp. 295-302 (1990).
12. S. Moirand, *Les discours de la presse quotidienne : Observer, analyser, comprendre*, Paris, Presses Universitaires de France (2007).
13. B. Pincemin, « Sémantique interprétative et textométrie », *Corpus*, **10**, pp. 259-269 (2011).
14. F. Rastier, *La mesure et le grain. Sémantique de corpus*, Paris, Honoré Champion, p. 49 (2011).
15. K. Beck *et al.*, « Manifesto for Agile Software Development » (2001).
16. T. Perrin, *Culture et eurorégions. La coopération culturelle entre régions européennes,* Bruxelles, Presses universitaires de Bruxelles, p. 7 (2013).
17. R. Raus, « L' "eurojargon" et sa variante française », *ARGOTICA*, **vol. 1, n°2**, pp. 383-394 (2014).
18. K. Bozhinova, « La terminologie eurolectale en usage dans les relations européennes », *Revue internationale d'études en langues modernes appliquées,* **4**, pp. 175-188 (2011).
19. R. Goffin, « L'eurolecte : oui, jargon communautaire : non », *Langages*, **76**, pp. 77-94 (1994).
20. A. Krieg-Planque, *La notion de formule en analyse du discours. Cadre théorique et méthodologique*, Besançon, Presses universitaires de Franche-Comté (2009).
21. M.-H. Hermand, « Le discours eurorégional. Indices convergents de légitimation d'un espace institutionnel », *Mots. Les langages du politique,* **106**, pp. 71-85 (2014).
22. M.-H. Hermand, « Eurorégions, émergence d'une culture discursive exemplaire », *Le discours et la langue. Revue de linguistique française et d'analyse du discours,* **6.2**, pp. 191-208 (2014).
23. D. Mayaffre, S. Mellet, « Les corpus réflexifs entre architextualité et hypertextualité », *Corpus*, **1** (2002).
24. M.-H. Hermand, « Affirmation des eurorégions en discours. Engagement dans l'apprentissage de la responsabilité », *Communication & Organisation,* **48** (à paraître en 2015).
25. C. Gobin et J.-C. Deroubaix, « L'analyse du discours des organisations internationales. Un vaste champ encore peu exploré », *Mots. Les langages du politique,* **94**, pp. 107-114 (2012).