Remise en questions d'une lecture kuhnienne de la géographie française : réflexions épistémologiques entre sciences sociales, humanités numériques et données massives

Max Beligné

N° d'ordre NNT : 2023LYO20094

THÈSE de DOCTORAT DE L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 483

Sciences sociales

Discipline : Géographie Aménagement Urbanisme

Soutenue publiquement le 20 décembre 2023 par :

Max BELIGNÉ

Remise en questions d'une lecture kuhnienne de la géographie française.

Réflexions épistémologiques entre sciences sociales, humanités numériques et données massives.

Devant le jury composé de :

Thierry JOLIVEAU, Professeur des Universités, Université Jean Monnet Saint-Étienne, Président

Hervé THÉRY, Directeur de recherche émérite, CNRS, Rapporteur

Damon MAYAFFRE, Chargé de recherche HDR, CNRS, Rapporteur

Aurélien BERRA, Maître de conférences, Université Paris Nanterre, Examinateur

Isabelle LEFORT, Professeure des Universités, Université Lumière Lyon 2, Directrice de thèse

Sabine LOUDCHER, Professeure des Universités, Université Lumière Lyon 2, Co-Directrice de thèse

Contrat de diffusion

Université Lumière Lyon 2

École Doctorale n°483 Sciences Sociales

Remise en questions d'une lecture kuhnienne de la géographie française :

Réflexions épistémologiques entre sciences sociales, humanités numériques et données massives

Max BELIGNÉ

Discipline de doctorat :
Géographie-Aménagement

Sous la direction de :

Isabelle LEFORT et Sabine LOUDCHER

Présentée et soutenue publiquement le 20 décembre 2023

Devant le jury composé de :

Hervé Théry, Directeur de recherche émérite CNRS - Université de São Paulo	Rapporteur
Damon Mayaffre, Chargé de recherche CNRS - Université Côte d'Azur	Rapporteur
Boris Beaude, Professeur associé Université de Lausanne	Examinateur
Aurélien Berra, Maître de conférences Université Paris-Nanterre	Examinateur
Thierry Joliveau, Professeur des Universités Université de Saint-Étienne	Examinateur
Isabelle Lefort, Professeure des Universités Université Lumière Lyon 2	Directrice
Sabine Loudcher, Professeure des Universités Université Lumière Lyon 2	Directrice
Gabrielle Richard, Directrice Unité d'Appui et de Recherche *Persée*	Invitée

---

Remerciements

Ce travail a tout d'abord bénéficié pendant trois ans (2016-2019) d'une allocation doctorale de la Région Auvergne-Rhône-Alpes. Cette allocation relevait plus spécifiquement de la cinquième *Communauté de Recherche Académique* (ARC 5) intitulée « Cultures, Sciences, Sociétés et Médiations ». Sans cette aide financière, je n'aurais pas commencé cette aventure scientifique. Je tiens donc à remercier vivement les ex-dirigeants politiques qui avaient mis en place ce dispositif et Corinne Sainte-Colombe qui l'a animé.

Ma thèse s'est déroulée à l'université Lyon 2 au sein de deux organismes d'accueil différents : l'Unité Mixte de Recherche « Environnement Ville et Société » (UMR EVS) et le laboratoire « Entrepôt, Représentation et Ingénierie des Connaissances » (ERIC). Ces deux organismes correspondent aux institutions de référence de mes directrices de thèse : Isabelle Lefort et Sabine Loudcher. Plusieurs personnes m'ont particulièrement aidé à prendre mes marques :

- Au niveau de l'UMR EVS, j'ai reçu à plusieurs moments des réponses professionnelles et amicales de Christian Montès. Manuel Appert a réalisé un bon suivi de mon doctorat et je dois à Anne Honneger, le fait d'avoir été, dès ma première année de thèse, représentant des doctorants. Enfin, sur le plan administratif, je souhaite remercier Patrick Gilbert pour ses bons et loyaux services dans cette UMR.
- Au niveau du laboratoire ERIC, Jean-Hugues Chauchat m'a accompagné intellectuellement et aussi substantiellement (avec des aides efficaces pour obtenir plusieurs contrats). Mon retour dans le monde de la recherche lui doit beaucoup. Je l'en remercie grandement. Dans les projets réalisés avec ce laboratoire, j'ai particulièrement apprécié les impulsions intellectuelles de Julien Velcin et d'Adrien Guille. Je tiens aussi à remercier le directeur de l'époque, Jérome Darmont, ainsi qu'Osman Habiba, la responsable administrative et financière, pour leur accueil.

Par rapport à mes premières orientations de travail, j'ai eu la chance d'être gentiment accueilli par Roger Brunet, Paul Claval et Yves Lacoste à leur domicile. Il y a des moments que l'on n'oublie pas dans sa carrière de chercheur. Ces entretiens menés en début de doctorat avec ces acteurs majeurs de la géographie française en font partie.

Mes premières orientations de travail m'ont également conduit à recevoir de l'aide de deux ingénieures de la Maison des Sciences de l'Homme (MSH) de Lyon Saint-Etienne : Hélène Kieffer et Celine Faure. Même si je n'ai pas continué les pistes qu'on avait

commencé à élaborer ensemble, je les remercie de m'avoir aidé dans ces premiers temps si importants dans une recherche.

Au niveau de l'accès à mes données, il me faut évidemment remercier une partie de l'équipe de l'Unité d'Appui et de Recherche *Persée* : Hélène Bégnis, Viviane Boulétreau, Eric Astier, Gabrielle Richard et Nathalie Fargier. De plus, et ce, même si je n'ai pas utilisé directement leurs données pour cette thèse, je tiens à mentionner Jean-Baptiste de Vathaire et Marc Minon pour leur aide dans la récupération des données du portail *Cairn*.

Au niveau des traitements, il me tient à cœur de remercier fortement les milliers de personnes qui travaillent à proposer des solutions *open source* (*Ubuntu*, *LibreOffice*…), des langages de programmation (*Python*, *R*, *Javascript*…) et des librairies (*Lxml*, *Gensim*, *Django*…). Sans eux, tout ce travail n'aurait pas eu l'ampleur que j'ai pu lui donner. J'ai également bénéficié pendant quelques mois de l'aide d'un ingénieur d'études, Pierre Prablanc, pour améliorer plusieurs parties du code.

Je tiens à remercier également vivement tous les organisateurs de colloques auxquels j'ai pu participer. Ce sont toujours des moments d'échanges importants, que ça soit avec des personnes déjà connues ou par le biais de nouvelles rencontres. Après quelques-uns de ces colloques, certains acteurs se sont lancés dans la publication d'actes. J'ai pu constater le volume de travail bénévole que cela demande. Donc, un grand merci pour leur implication à Camille Mortelette et François Moullé (à la suite du colloque « Penser avec les discontinuités » à l'université d'Arras en 2018), à Léo Dumont, Octave Julien et Stéphane Lamassé (colloque « Histoire, langue et Textométrie » à la Sorbonne en 2019) et à Pierre Ratinaud (« Journées d'Analyse de Données Textuelles » à Toulouse en 2020). Je n'oublie pas non plus tous ceux qui ont relu patiemment et bénévolement tous mes articles (à la suite de colloques ou directement proposés à des revues) et qui m'ont permis de les améliorer.

Au niveau de l'aide informatique, je tiens à remercier Patrice Foxonet pour la mise en place de mon poste informatique de départ. Gérald Foliot, membre de l'infrastructure de recherche *Huma-Num*, a été d'un accompagnement particulièrement important dans la mise en place technique de l'application Web qui accompagne ma thèse.

J'ai particulièrement apprécié avoir quelques petits échanges avec ce que je considère comme des personnalités de l'analyse textuelle française : Bénédicte Pincemin, André Salem, Pierre Ratinaud, Laurent Vanni, Dominique Labbé, Serge Heiden et Jean-Marc Leblanc. Merci également pour toutes les aides ponctuelles reçues sur le chemin par une

diversité d'acteurs différents : Thomas Bourbon, Damien Petermann, François Briatte, Nicolas Dugué, Jean-Charles Lamirel…

Sur cette fin de thèse, je dois une fière chandelle à l'accompagnement de Pierre Mercklé, dans le TD « Observation quantifiée du monde social ». Ce travail m'a permis d'intégrer le 1er février 2023 le poste d'ingénieur d'étude au sein de la Plateforme Universitaire de Données de l'Université Grenoble-Alpes (PUD-GA). J'aimerais ici remercier l'ensemble de l'équipe de la MSH Alpes, l'infrastructure de recherche *PROGEDO* et Frédéric Gonthier pour l'accueil et l'enthousiasme !

Mon travail de doctorat doit bien sûr beaucoup au travail de thèse précédemment mené par Olivier Orain. Mon travail critique doit être vu comme une preuve de ma prise au sérieux de sa thèse. J'ai grandement apprécié nos discussions et ses invitations : tout d'abord, dans un cadre informel pour partager mes premiers résultats de recherche ; puis dans un séminaire de l'équipe *Épistémologie et Histoire de la GéOgraphie* (EHGO) pour présenter l'avancée de mes travaux ; et enfin, à collaborer au cinquantenaire de la revue *L'Espace Géographique*. Je lui suis très reconnaissant de n'avoir pas coupé les ponts du fait de ma critique scientifique.

Pendant tout ce doctorat, j'ai la grande chance d'avoir un très bon accompagnement de la part de mes deux directrices de thèse. La confiance, la bienveillance et l'autonomie qu'elles m'ont données sont vraiment les plus belles choses que l'on peut accorder à un doctorant. Merci Sabine pour ton ouverture d'esprit et ta rigueur ! Merci Isabelle pour ta générosité et cette intelligence des liens que tu partages ! Ce chemin ensemble a été une chance pour moi.

Dans la dernière ligne droite, les aides de Manon Boucharechas, Éric Mermet, Émilien Schultz, Mariannig Le Bechec et Julien Jacques ont été particulièrement bienvenues.

Pour finir, un immense merci à mes parents et à Élodie, ma compagne, qui m'a « supporté » durant toutes ces années. Sans ces proches et cette stabilité, il m'aurait été difficile de mener cette recherche aussi loin que j'ai pu le faire.

Et belle vie à mes deux filles qui ont commencé à grandir en même temps que cette thèse !

Résumé

Remise en questions d'une lecture kuhnienne de la géographie française : Réflexions épistémologiques entre sciences sociales, humanités numériques et données massives

Cette thèse est d'abord le récit d'un parcours de recherche. Il a fallu préciser la problématique pour justifier le corpus (revues des *Annales de Géographie* et de *L'Espace Géographique*) et l'orientation méthodologique (plongements de mots diachroniques). La problématique s'est focalisée sur un test de l'« hypothèse socio-sémantique » qu'Olivier Orain a réalisé dans sa thèse (2003) pour affirmer la pertinence de sa lecture kuhnienne de la géographie française.

Les résultats obtenus montrent que cette hypothèse socio-sémantique ne se vérifie pas. Les termes « espace » et « milieu » n'ont pas la même assiette sémantique dans la géographie française des années 1960. Durant la phase révolutionnaire des années 1970, « espace » n'acquiert pas un nouveau sens, mais plusieurs. De plus, ces sens ne sont pas circonscrits au paradigme théorico-quantitativiste. Enfin, dans les années 1980, il n'y a pas de nouvelle stabilité et de bascule sémantique affirmées.

Ces résultats ne permettent pas pour autant de remettre en cause directement la lecture kuhnienne d'Olivier Orain. Les développements de Jean-Claude Passeron dans *Le raisonnement sociologique* (2006, *1991*) fournissent une compréhension approfondie de cette situation : le pôle expérimental se suffit rarement à lui-même dans les sciences sociales. Une conséquence de cette situation est une opposition franche à toute lecture kuhnienne de ces sciences. Cette perspective passeronienne a conduit à examiner précisément de nombreux autres arguments utilisés par Olivier Orain et à mettre en lumière plusieurs limites par rapport au schéma initial kuhnien.

Dans un dernier temps, cette perspective critique est appliquée aux discours contemporains : humanités numériques et données massives. Une déconstruction des discours qui s'affirment régulièrement comme changement de paradigme est ainsi effectuée pour y substituer une réflexion sur les conditions de production et *in fine* d'enseignement des sciences sociales.

Abstract

Challenging a kuhnian interpretation of french geography: Epistemological reflections at the intersection of social sciences, digital humanities and big data

This thesis chronicles a research journey that necessitated a throrough problem formulation to substantiate the chosen corpus (*Annales de Géographie* and *L'Espace Géographique* journals) and methodological approach (utilizing word embeddings). The crux of the problem revolved around addressing the 'socio-semantic hypothesis' postulated by Olivier Orain in his thesis (2003) to validate the pertinence of his kuhnian perspective on French geography.

The ensuing results show this socio-semantic hypothesis does not withstand scrutiny. Most notably, the notions of 'space' and 'environment' exhibit distinct semantic foundations within french geography during the 1960s. In the tumultuous years of the 1970s, 'space' does not acquire a singular new meaning but rather multiple interpretations. Furthermore, these nuanced meanings extend beyond the confines of the theoretical-quantitative paradigm. Lastly, the 1980s do not usher in a newfound stability or clear semantic shift.

Still, it is worth stressing that these findings do not directly impugn Olivier Orain's kuhnian interpretation. Jean-Claude Passeron's developments in *Le raisonnement sociologique* (2006, *1991*) provide an in-depth understanding of this situation : the experimental framework is rarely sufficient in social sciences. One consequence of this situation is an outright opposition to any kuhnian reading of these sciences. This perspective from Passeron has led to a meticulous evaluation of various other arguments advanced by Olivier Orain, revealing several shortcomings when compared to the original kuhnian schema.

In the final phase of the thesis, this critical perspective is extrapolated to contemporary discourse, namely the realms of digital humanities and big data. A deconstruction of the narratives that frequently assert themselves as heralding paradigmatic shifts is carried out, supplemented by an introspective analysis of the conditions governing production and, ultimately, pedagogy of social sciences.

Sommaire

Notes de l'auteur

Une thèse reliée à une application web

En plus des habituels supports papier et PDF, les parties 1 et 2 de cette thèse sont accessibles via une application Web à partir du lien suivant : https://analytics.huma-num.fr/EtPistezMots/. Avant d'utiliser cette application, il est conseillé de lire sa présentation détaillée au début de la partie 1 (*cf.* section IntroPart1). Tous les liens internet et les renvois introduits par l'abréviation *cf.* peuvent être ouverts en cliquant dessus dans la version PDF de cette thèse.

Citations

Les citations sont délimitées par des guillemets français (« »). Dans les cas où une citation contenait des guillemets français, ces derniers ont été transformés en guillemets anglais (" ") pour qu'il n'y ait pas d'ambiguïté sur les débuts et les fins de citations. Pour aérer le texte, les citations de plus de quatre lignes ont fait l'objet d'une mise en page spécifique (police plus petite et centrage) et deux conventions ont été utilisées pour les références bibliographiques : (date, page)[1] et (auteur date, page)[2]. Dans quelques cas, les pages ne sont pas précisées, car les documents n'ont été trouvés que sous forme électronique sans mention de numéro de page. De plus, il m'a semblé parfois intéressant de préciser la date de la première édition d'une publication qui est alors indiquée en italique[3]. Enfin, comme pour les liens internet et les renvois, il est possible dans la version PDF de ce document de cliquer sur chaque citation pour consulter directement la référence bibliographique correspondante.

Ressources

En plus d'un index, un répertoire est disponible à la fin de ce document permettant de consulter les dates biographiques des géographes cités avec un classement par ordre alphabétique (*cf*. Répertoire Présent1) ou chronologique (*cf.* Répertoire Présent2).

[1] Par exemple : Kuhn (1983, 145).
[2] Comme (Kuhn 1983, 145).
[3] Comme (Kuhn 1983, *1962*).

Prologue

Discours de ma méthode

La reprise personnalisée du célèbre titre cartésien (1888, *1637*) renvoie évidemment à une moindre ambition ! Toutefois, cet intitulé octroie un statut particulier à ce prologue, allant au-delà d'une simple introduction. Si je me suis autorisé, avec l'accord de mes deux directrices de thèse, Sabine Loudcher et Isabelle Lefort, des développements d'une longueur inhabituelle pour ce type d'exercice, c'est bien parce que mon objectif est ici de mieux faire comprendre[4] l'intrication des dimensions intellectuelles et humaines dans la réalisation de ma recherche.

Si certains mettent trois ans pour réaliser l'ensemble de leur thèse, il m'a fallu cette même durée de travail intensif pour finir par en déterminer définitivement le sujet. Je ne regrette nullement cet investissement préalable tant cette étape a été déterminante pour la suite. Au risque d'effrayer les partisans de la thèse en trois ans pour tous, j'ajouterai que cette première période a été, dans mon cas, précédée d'une phase plus longue, lente et discontinue de maturation (pendant laquelle mon contrat doctoral n'avait pas encore commencé).

Le prochain chapitre intitulé « Entrée(s) en matière(s) » revient sur cette première phase en introduisant les éléments initiaux de contexte. Quant au chapitre suivant, intitulé « La restitution d'une trajectoire de recherche par trois projets successifs », il est consacré à mes trois premières années de contrat doctoral qui ont ainsi progressivement abouti à la définition de ma problématique et de mes axes de travail.

[4] Le sens étymologique de « comprendre » est ici particulièrement approprié : « cum "avec" et prehendere "prendre, saisir", littéralement "saisir ensemble, embrasser quelque chose, entourer quelque chose" d'où "saisir par l'intelligence, embrasser par la pensée" » d'après le *Centre National de Ressources Textuelles et Lexicales* : https://cnrtl.fr/etymologie/comprendre consulté le 09/05/2023.

Chapitre 1 :
Entrée(s) en matière(s)

Les entrées en matières suivantes visent à préciser le positionnement initial de cette thèse, à revenir sur mon parcours prédoctoral et à développer quelques caractéristiques de mon dispositif de recherche.

I. Positionnement initial

L'exercice de reconstitution rétrospective ici réalisé intègre plusieurs éléments autobiographiques. La décision de mettre en avant ce récit dans ce début de thèse est loin d'être anecdotique. La première fois que j'ai évoqué cette idée avec Sabine Loudcher qui a encadré ce travail du côté informatique, j'ai pu ressentir un certain étonnement de sa part. C'était incontestablement loin de sa culture professionnelle. Quelques mois après, quand elle m'a annoncé avoir demandé à un de ses autres doctorants d'effectuer ce même travail et qu'elle était très contente du résultat, je dois avouer avoir ressenti une certaine satisfaction de participer, même de manière très modeste, à la diffusion d'une telle pratique de recherche.

Durant toutes ces années de thèse, Sabine Loudcher m'a accompagné dans de nombreux apprentissages statistiques et informatiques : lecture d'Analyses Factorielles de Correspondance, conception de base de données… Ces échanges ont participé à mon avis de ce mouvement qu'a évoqué Michel Serres dans un de ses derniers entretiens :

> « Des informaticiens doivent apprendre à devenir un peu sociologues, un peu économistes, etc. Et les chercheurs en sciences humaines doivent devenir un peu informaticiens. C'est indispensable d'avoir les deux points de vue pour plonger dans le vrai monde » (Serres, Dowek, et Abiteboul 2018).

Cette dernière expression « pour plonger dans le vrai monde » peut être amplement discutée, car des informaticiens ne connaissant pas les sciences humaines sont loin de vivre dans un monde factice. Cette remarque est évidemment tout aussi valable pour des chercheurs en sciences humaines ayant une faible culture informatique. Il faut donc interpréter cette idée de Michel Serres comme une référence au fait que nous vivons dans un monde où les algorithmes sont de plus en plus utilisés (d'où la nécessité d'une culture informatique) avec des limites indiscutables quant à leur capacité à tout gérer et expliquer (d'où le besoin des connaissances de sciences humaines).

Ceci étant précisé, la question du niveau d'acculturation souhaitable ou nécessaire sous-entendu par l'expression « un peu » dans l'affirmation précédente de Michel Serres, mérite d'être posée. Ces multiples « un peu » sont-ils suffisants pour créer une compréhension mutuelle et un dialogue ? Concrètement et dans le détail, l'articulation entre la culture informatique et celle des sciences humaines est loin d'être évidente. Les difficultés peuvent être multiples, portant sur le vocabulaire utilisé, les modes de raisonnement ou encore les

objectifs poursuivis. De plus, il est important de préciser que les différences existantes sont plus complexes qu'une opposition entre deux mondes. Par exemple, par rapport à la proposition précédemment mentionnée, à savoir celle de rendre compte d'un parcours personnel dans un rendu scientifique, il n'existe pas de posture commune aux Sciences Humaines et Sociales (SHS). En effet, s'il est vrai que l'illusion d'une écriture impersonnelle[5] de la science a été levée depuis longtemps (Latour, Woolgar, et Biezunski 2006), il existe également une grande diversité de positionnements et de pratiques par rapport à la présence énonciative au sein des SHS (Tutin 2010).

La justification du positionnement adopté dans cette thèse sur ce sujet ne peut pas, par conséquent, se contenter de faire référence à une pratique usuelle dans ce champ de la connaissance. Il faut aussi prendre en compte et expliciter à ce propos l'objectif initial de ce doctorat. En effet, cette recherche visait initialement une mise en lumière du rôle des revues dans l'évolution de la géographie française. Pour être plus concret encore, Jacqueline Pluet-Despatin développe quelques exemples historiques pour l'ensemble des SHS :

> « Il est rare que le titulaire d'une chaire ne dispose pas d'une revue qu'il dirige ou dans laquelle il a ses entrées : la publication d'un article intervient opportunément pour appuyer une candidature. La participation à un jury de thèse mérite un compte-rendu élogieux, parfois exercice d'école exécuté par un assistant. Celui-ci paye ses chances de carrière par quelques ingrates besognes administratives dans une revue, prémices de contribution ou de fonction plus valorisantes. La cooptation d'un nouveau membre dans un comité de rédaction peut être également la contrepartie tactique d'une alliance à nouer pour une élection aux enjeux de carrière importante » (Pluet-Despatin 1992, 319).

L'orientation initiale de cette recherche visait à rechercher des exemples similaires dans l'histoire de la géographie et à mieux évaluer leur importance dans l'évolution de la discipline. Il s'agissait donc de réhabiliter une partie de cette dimension sociale qui est souvent ensuite effacée par l'histoire collective. Dans cette optique, il y avait une cohérence forte à ne pas, moi-même, occulter la dimension sociale de la recherche effectuée.

Cette dynamique a conduit à aller plus loin que la seule mention de modalités personnelles et interpersonnelles déterminantes dans la production de travail : l'objectif a été d'essayer de restituer *a posteriori* toute la trajectoire de recherche avec ses difficultés, ses errements et ses trouvailles. Il faut préciser que cette démarche est d'autant plus légitime que les premières recherches effectuées, même quand elles incluent des réflexions et des réalisations qui n'ont pas été poursuivies, ont joué un rôle majeur dans l'élaboration de la posture de recherche et la construction de la problématique de ma thèse.

[5] Écriture impersonnelle qui privilégie des présentations qui ne mettent en avant que des successions de choix rationnels.

Au moment de l'écriture, il y a eu toutefois un équilibre assez subtil à trouver. En effet, en rentrant dans les détails, le fil conducteur était particulièrement complexe à suivre du fait des nombreuses pistes qui ont été explorées et des multiples changements qui ont eu lieu. À l'inverse, en étant trop elliptique, une trop grande partie, de ce qui a animé et de ce qui permet de comprendre l'évolution de cette recherche, était passée sous silence. Pour débuter cet exercice délicat, la partie suivante présente quelques éléments biographiques antérieurs au début officiel de la thèse. Ces éléments ont été jugés assez importants pour être ici mentionnés.

II. Histoire de rencontres et naissance d'un sujet

En 2004, la réalisation d'un Master 2 Recherche (spécialité interface nature / société) à l'Université de Lyon, effectué suite à des études de mathématiques[6] et d'ingénierie de l'espace rural[7], m'a permis de découvrir l'épistémologie de la géographie. Les cours donnés par Isabelle Lefort et Philippe Pelletier m'ont fourni alors des éclairages réflexifs et critiques par rapport à l'évolution d'une discipline. Pour être plus précis, ces cours ont commencé à me faire comprendre ce qui a été historiquement dit et réalisé, mais aussi ce qui n'a pas été dit, les impasses, les tensions intellectuelles, les coups de force de certains acteurs… L'objectif global était de mieux appréhender les enjeux d'une discipline pour mieux situer et comprendre sa propre recherche. Ce travail, mené plusieurs années après, a hérité et reste dans la continuité de cette formation initiale.

En 2007, un « petit travail[8] » mené pour la revue *Géocarrefour* et l'Unité d'Appui et de Recherche (UAR) *Persée* comme annotateur a été fondamental dans la genèse de ma thèse. Pour bien comprendre son importance, il faut tout d'abord rappeler que cette UAR a ouvert en 2005 un portail[9] qui contient aujourd'hui plus de 300 revues numérisées, dont beaucoup de périodiques de référence pour les SHS francophones. Avant de mettre en ligne une revue, tout un travail est effectué en amont pour repérer et indexer les éléments structurants (titre, auteur, mot-clé, résumé, sous-titre, figure, note…) de tous les articles de la revue en question. C'est cette tâche que j'ai été amené à réaliser pour la revue *Géocarrefour*. Ce travail est plutôt fastidieux car il est nécessaire d'examiner chaque page une à une. Il a été sans conteste

[6] Une première année de Classe Préparatoire aux Grandes Écoles (CPGE) au lycée Champollion de 1999 à 2000.
[7] Trois années à l'École Supérieure Européenne d'Ingénierie de l'Espace Rural (IER) à Poisy de 2000 à 2003
[8] Le qualitatif de « petit » renvoie à la durée puisque ce n'était qu'un contrat à durée déterminée de 2 mois, mais aussi à la perception habituelle d'un tel travail plutôt peu valorisé.
[9] https://www.persee.fr/ consulté le 18/09/2023.

chez moi à l'origine d'une prise de conscience et d'une certaine connaissance de la masse de données produite par cette UAR.

Le constat, plusieurs années plus tard, d'une sous-exploitation du potentiel scientifique de ces données, a été d'autant plus important du fait de cette expérience professionnelle préalable. De plus, il est utile d'évoquer, sans rentrer ici dans les détails, qu'il existe quelques différences entre les données produites au moment de l'annotation et celles effectivement visibles par un utilisateur habituel du portail. Cette connaissance de données spécifiques cachées et cette conscience d'un potentiel scientifique largement inexploité ont été des facteurs déterminants dans ma décision de réaliser un doctorat. Cependant, ce projet ne s'est pas formalisé directement à la suite de ce premier petit travail. D'un point de vue scientifique, une analyse *a posteriori* me conduit à penser qu'il me manquait alors tous les outils techniques pour exploiter ces données et pouvoir développer une recherche à partir de leur utilisation.

Cet apprentissage méthodologique a, quant à lui, débuté à partir de 2013 à la suite d'un travail proposé par Isabelle Lefort et Jean-Hugues Chauchat[10]. Ma première mission a consisté à travailler statistiquement sur un corpus d'entretiens réalisés par Yann Calbérac (2010) pour sa thèse de géographie. Ce recrutement était dû à ma bonne connaissance de ce corpus que j'avais en grande partie retranscrit plusieurs années avant. Pendant cette mission, les méthodologies employées m'ont fortement intéressé du fait de mon tropisme initial pour les mathématiques. L'idée de m'investir dans la recherche, voie que j'avais eu la chance de connaître et d'apprécier dix ans auparavant en réalisant un mémoire de Master 2[11] sous la direction d'Isabelle Lefort, s'est alors précisée jusqu'à devenir un objectif central. Dans cette perspective, j'ai eu la chance d'être intégré comme force vive dans un projet de détection automatique de métaphores dans des textes de géographie avec la même équipe et une partie du laboratoire ERIC. J'ai pu alors découvrir des méthodes statistiques de plus en plus perfectionnées, ce qui explique une partie des orientations de cette thèse. Ce moment m'a aussi permis de mieux connaître plusieurs personnes de ce laboratoire, et notamment Sabine Loudcher.

C'est dans ce contexte que s'est précisé le projet de ce doctorat en co-direction. Quelques grandes caractéristiques du dispositif de recherche qui s'est alors mis en place sont détaillées dans la section suivante.

---

[10] Professeur émérite de statistique et de mathématique à l'Université de Lyon.

[11] M Beligné, *Proposition d'un modèle d'analyse pour une éducation géographique, réflexions à partir du « territoire »,* Master 2, 2004, non publié.

III. Un dispositif partagé, interdisciplinaire et transversal

Cette présentation n'est pas exhaustive : elle a seulement pour objectif de développer quelques caractéristiques permettant de mieux saisir certaines spécificités de la recherche effectuée. Cet exercice m'a demandé un effort d'objectivation dont j'assume évidemment les limites qui sont celles de toute subjectivité de chercheur.

1. Mutualisme

Le choix de ce premier terme peut étonner. Dans sa définition zoologique, mutualisme signifie : « association de deux animaux d'espèces différentes qui retirent des bénéfices réciproques de cette union sans vivre aux dépens l'un de l'autre » (Dictionnaire *Le Robert* 2002). Cette métaphore biologique doit se comprendre comme une mise en abyme, à la fois humoristique et signifiante, pour désigner le rapport entretenu par ce doctorat avec le projet de détection automatique de métaphores dans des textes de géographie. Les réunions de ce projet qui ont eu lieu pendant mes deux premières années de thèse m'ont permis de partager régulièrement toute une partie de mon travail avec des chercheurs au-delà de mes deux directrices, notamment avec Jean-Hugues Chauchat et Julien Velcin[12].

L'écriture d'un article[13] a conduit à ma participation à la conférence TALN (Traitement Automatique du Langage Naturel) au début de mon doctorat. Ce moment a été un temps d'acculturation important. Par rapport à ce doctorat, il faut préciser également que des méthodes, alors particulièrement utilisées par cette communauté de chercheurs – les plongements de mot[14] – ont donné ensuite matière à une présentation d'Adrien Guille[15] dans le cadre d'une réunion du projet de détection de métaphores. À partir de ces expériences, il y a eu de ma part une familiarisation progressive avec ces outils. Ce processus a sans conteste rencontré et participé à l'évolution de ma problématique (*cf.* Chap2).

Sur un autre plan, ce projet de détection automatique des métaphores a utilisé le corpus construit pour ce doctorat et a même manifesté la volonté de l'élargir à d'autres revues. Cet objectif a abouti à l'embauche d'un ingénieur d'études pour réaliser une application de préparation des données, travail que j'ai encadré avec Sabine Loudcher. Ces synergies, autant sur le plan intellectuel que pratique, ont joué un rôle important dans l'évolution de mon travail.

---

[12] Professeur d'informatique au laboratoire ERIC.
[13] *Cf.* Bibliographie : (Beligné *et al,* 2017).
[14] Traduction de *word embedding*.
[15] Maître de Conférences au laboratoire ERIC.

Le terme suivant « interdisciplinarité » permet incontestablement d'aller plus loin qu'une simple approche mutualiste, mais requiert de faire un point sur quelques-unes de ses acceptions.

2. Interdisciplinarité

La réalisation d'une interdisciplinarité a commencé dans ce travail avec une volonté forte dès le départ de faire des réunions à chaque fois avec mes deux directrices en même temps. Il nous est arrivé dans la pratique de faire quelques séances en dissocié, mais ce qui a été globalement visé a été une mise en commun de l'ensemble des problématiques. L'objectif a été d'éviter qu'une difficulté spécifique, qu'elle soit d'ordre statistique, informatique ou relevant des SHS, ne soit exposée et traitée qu'avec la spécialiste du domaine en question. Cette modalité de travail me conduit ici plutôt à privilégier le terme d'interdisciplinarité par rapport à celui de pluridisciplinarité qui sous-entend une segmentation plus forte (Resweber 2011). La collaboration au long cours qui a été précédemment explicitée, a joué un rôle non négligeable dans le processus d'acculturation et de travail en commun. Si l'ouverture de mes deux directrices à des sujets qui ne sont pas initialement les leurs a été un élément fondamental, le fait de pouvoir mener personnellement des travaux autant sur les deux versants a été aussi un facteur déterminant dans l'élaboration de cette interdisciplinarité.

Bien qu'un des cadres revendiqués de ce travail soit les humanités numériques qui se sont définies initialement comme « une transdiscipline, porteuse des méthodes, des dispositifs et des perspectives heuristiques liés au numérique dans le domaine des SHS »[16], il ne m'a pas semblé opportun de choisir le terme de « transdisciplinarité » pour caractériser ce travail. En effet, il y a eu dès le départ, non pas une remise en cause des cadres disciplinaires, mais, bien au contraire, une inscription forte dans ces derniers. En effet, l'objectif initial de ce doctorat était de relire l'histoire d'une discipline. L'évolution de la problématique qui a ensuite eu lieu (*cf.* Chap2), ne me semble aucunement remettre en cause cette inscription première dans des matrices disciplinaires.

Il est également exact, comme le souligne le terme suivant de « transversalité », que les réflexions menées ont fait appel à plusieurs champs disciplinaires. Cette dynamique est inscrite par certains dans un contexte plus global de remise en cause des disciplines (Fabiani 2012). Les sous-entendus idéologiques possibles liés à cette dynamique, notamment celui

[16] *Manifeste des Digital Humanities* (Collectif 2010).

d'un néolibéralisme s'attaquant à un « âge disciplinaire qui serait celui de la *Structure des révolutions scientifiques* de Kuhn » (*Ibid* 2012, 137) au profit d'un « monde fluide des configurations provisoires » (*Ibid* 2012, 139), me poussent à préciser que cette recherche ne revendique aucunement une quelconque fin des disciplines. Le positionnement scientifique qui s'est affirmé au cours du temps est même tout à fait opposé à cette perspective : la contextualisation des recherches dans leur(s) champ(s) d'origine est ici revendiquée comme une pratique tout à fait nécessaire pour mieux comprendre les références mobilisées et les enjeux de tout travail intellectuel. Dans cette perspective, c'est davantage une lecture en termes d'approche transversale qui doit être privilégiée.

3. Transversalité

L'aspect transversal de ce travail provient d'une nécessité qui s'est fait sentir dès le début de la recherche. En effet, il n'existe pas un champ à proprement parler des études sur les revues (Boure 1993), ce qui a amené à réaliser des recherches dans des disciplines multiples. Ensuite, quand la problématique a évolué en se recentrant plus précisément sur la thèse d'Olivier Orain (2003), il m'a fallu explorer et mobiliser les champs de la sociologie, voire de la philosophie, pour réfléchir la thèse de cet auteur.

De plus, ce travail, par sa double affiliation, géographie et informatique, s'inscrit dès le départ dans les problématiques du champ des humanités numériques. C'est ainsi que j'ai eu la chance, dès le début de ma thèse, d'être invité à une rencontre[17] organisée par l'infrastructure de recherche *Huma-Num* et de pouvoir échanger avec de nombreux acteurs de ce champ.

Enfin, la réflexion sur mon sujet m'a conduit à rencontrer et à lire plusieurs chercheurs[18] ne se réclamant pas directement des humanités numériques, mais traitant de la problématique de l'évolution des SHS dans le contexte des données massives[19]. L'objectif, en explorant ces différents champs de recherche, a été d'alimenter une démarche réflexive à visée épistémologique en suivant une problématique qui a évolué et s'est précisée au cours du temps.

---

[17] Rencontre *Huma-Num* 2016 à Valpré.

[18] Tout d'abord avec une participation au colloque *Science XXL : Ce que l'abondance et la diversité des données font aux sciences sociales,* Paris, 2017. Puis à travers la lecture de nombreux auteurs comme Boris Beaude, Bruno Bachimont, Jean-Philippe Cointet…

[19] *Big data* en anglais. Cette expression désigne l'explosion quantitative des données numériques disponibles et leurs traitements associés. Elle est traitée de manière plus détaillée au chapitre 13.

Notons enfin qu'au sein de mes deux laboratoires de rattachement, il n'a pas existé durant les premières années de ma thèse, de volonté de développer un pôle « humanités numériques » à EVS tandis que du côté d'ERIC, s'il a existé (et existe encore) un véritable intérêt pour la fouille de données textuelles et les humanités numériques, les réflexions épistémologiques étaient très peu développées. Ces précisions permettent de comprendre que mon travail s'est situé en périphérie, à la marge de ses deux rattachements institutionnels. Cette situation explique en partie une forte dimension d'auto-apprentissage et d'autonomie dans les travaux effectués. Le chemin de recherche suivi découle d'une volonté continue d'adaptation aux difficultés et aux questionnements rencontrés successivement.

La trajectoire scientifique réalisée est décomposable en trois grands projets scientifiques qui se sont succédé au cours des années de travail et qui sont présentés dans le chapitre suivant.

Chapitre 2 :
La restitution d'une trajectoire de recherche par trois projets successifs

Les appellations des trois projets scientifiques suivants ont été formulées *a posteriori* pour mieux donner à comprendre les grandes directions de travail dans lesquelles je me suis engagé. Ces trois projets sont des reconstitutions d'étapes entre lesquelles il a existé parfois des transitions plus lentes et floues que ce rendu écrit ne le présente. L'objectif n'a pas été d'être le plus précis possible, mais de restituer la teneur et les raisons du parcours effectué.

I. Le projet initial : la géographie par les revues

Le premier projet de travail pour cette thèse s'appuyait sur l'ensemble des revues numérisées de géographie française et l'utilisation de méthodes statistiques pour analyser ces données. Bien que cette recherche provienne en partie d'idées conçues lors d'une expérience professionnelle en lien avec le portail *Persée* (*cf.* section Chap1.II), le projet initial ne s'est pas limité aux revues qu'il héberge. Plusieurs raisons expliquent ce choix : tout d'abord, certains périodiques comme *Hérodote*, dont l'importance est reconnue dans le champ géographique, ne sont pas hébergés par le portail *Persée*. De plus, les parties contemporaines des revues, après les années 2000 (avec des dates variables selon les périodiques), ne sont pas hébergées par le portail *Persée* mais par d'autres plateformes comme *Cairn*[20] et *OpenEdition Journals*[21].

D'un point de vue méthodologique, parallèlement à la réalisation d'analyses quantitatives à partir des données numériques issues des revues, des entretiens plus qualitatifs avaient été prévus avec les acteurs des périodiques en question (directeur, comité de rédaction…). L'objectif central était de mettre à jour l'importance qu'ont eue les jeux d'acteurs dans les coulisses des revues sur la production scientifique. Les entretiens devaient permettre de rentrer dans le détail de ces jeux d'acteurs. Les analyses quantitatives avaient pour but de mettre en lumière l'évolution de la production scientifique de ces revues au cours du temps. En articulant ces deux méthodologies, l'hypothèse était que la production d'une relecture à la fois justifiée et originale de la discipline était possible.

Plusieurs difficultés ont été rencontrées lors de l'exécution de ce premier projet. Ces difficultés ont donné lieu à de multiples travaux qui relèvent de quatre problématiques différentes : le choix du corpus, l'obtention et la préparation des données, le choix des méthodologies d'analyse et enfin les entretiens avec l'articulation quantitatif-qualitatif.

[20] https://www.cairn.info/ consulté le 18/09/2023.
[21] https://journals.openedition.org/ consulté le 18/09/2023.

1. Choix du corpus

Le périmètre de ce premier projet avait été réduit dès le départ à l'ensemble de la géographie française à l'exception de la branche se consacrant exclusivement à la géographie physique. Cette réduction peut se justifier en mentionnant des dynamiques de production et de diffusion de connaissances différentes entre les SHS et les sciences physiques (Larivière *et al.* 2006). Malgré cette précision apportée au périmètre de la recherche, la problématique du choix du corpus et de sa représentativité vis-à-vis de la discipline a été une difficulté centrale. Une première réponse assez opportuniste consiste à chercher et à exploiter tout ce qui est disponible. Toutefois, cette technique soulève de nombreuses questions. Par exemple, que faire des nombreux périodiques interdisciplinaires ? À partir de quand une revue peut-elle être considérée de géographie ? À partir de quand doit-elle au contraire être exclue du corpus choisi ?

Ces questions interrogent la définition même et les limites de la discipline. De plus, en utilisant cette technique, certains manques sont difficilement justifiables. Par exemple, la revue *Géographie et Culture*, qui a été sans conteste un périodique important dans l'histoire et la structuration du champ disciplinaire, n'a pas numérisé ses numéros de 1993 à 2005. S'il est décidé de réaliser une exception pour cette revue en s'attelant à la numérisation de ses numéros manquants, comment justifier cette exception par rapport aux autres revues non numérisées ? Face à ce problème, deux grandes directions de recherche ont été testées pour essayer de mieux justifier le périmètre de la recherche.

La première s'appuie sur la notion de « corpus réflexif » (Mayaffre 2002). Cette notion renvoie au fait que les différents constituants d'un tel corpus « renvoient les uns aux autres pour former un *réseau sémantique* performant dans un tout (le corpus) cohérent et auto-suffisant » (*Ibid* 2002, 5). Concrètement, le processus pour créer un corpus réflexif n'est pas vraiment totalement explicité. Cette approche a conduit à la formulation d'un ensemble de sous-problématiques : quels ont été les positionnements des revues régionales apparues dans les années 1920-1930 (R*evue de Géographie Alpine*, *Études Rhodaniennes* et *Revue géographique des Pyrénées et du Sud-Ouest Européen*) avec la revue générale de référence, les *Annales de Géographie* ? De la même manière, *Norois*, *Méditerranée*, la *Revue Tiers Monde* et la *Revue Géographique de l'Est* (*cf.* Tableau 1) semblent former une deuxième vague régionale plus tardive, qu'il est intéressant de comparer à la première. S'agit-il seulement de nouvelles initiatives locales qui complètent le paysage des revues régionales ou participent-elles des mutations disciplinaires qui aboutiront à la « nouvelle géographie » des années 70 portée par la triade *L'Espace Géographique* - *Espace Temps* - *Hérodote* ? Comment cette période de renouvellement disciplinaire a-t-elle « travaillé » et a-t-elle été

« travaillée » par les revues déjà présentes ? Les nouvelles revues thématiques des années 1980-1990 (*Espace Populations Sociétés*, *Géographie et Cultures*, *M@ppemonde*) ont-elles modifié le centre de gravité des revues déjà existantes ? En quoi *Cybergéo*, avec une logique de publication s'affranchissant des dossiers thématiques devenus dominants par rapport aux varias, change-t-elle la donne ?

Cette étape a abouti à l'établissement d'une liste de revues pouvant servir à construire un premier corpus :

Date création	**Nom de la revue**	**Groupe d'appartenance**
1892	*Annales de Géographe*	Générale 1ère génération
1913	*Revue de Géographie Alpine*	Régionale 1ère génération
1926	*Études Rhodaniennes, Revue Géographie de Lyon* puis *Géocarrefour*	Régionale 1ère génération
1930	*Revue géographique des Pyrénées et du Sud-Ouest Européen*	Régionale 1ère génération
1954	*Norois*	Régionale 2ème génération
1960	*Méditerranée*	Régionale 2ème génération
1960	*Revue Tiers Monde*	Régionale 2ème génération
1961	*Revue Géographique de l'Est*	Régionale 2ème génération
1972	*L'Espace Géographique*	Générale 2ème génération
1975	*Espaces Temps*	Générale 2ème génération
1976	*Hérodote*	Générale 2ème génération Thématique
1983	*Espace Populations Sociétés*	Thématique
1986	*Mappemonde puis M@ppemonde*	Thématique
1992	*Géographie et cultures*	Thématique
1996	*Cybergéo*	Générale 3ème génération

Tableau n°1 : La liste de revues retenues à partir de la notion de corpus réflexif.

La méthode du corpus réflexif a eu l'avantage d'apporter une première réponse par rapport à la question délicate de la représentativité disciplinaire. Mais dès lors qu'une sous-problématique est creusée, ces questions du choix du corpus et de la représentativité redeviennent centrales. Par exemple, en prenant celle du rôle des revues régionales de seconde génération dans les mutations disciplinaires qui ont abouti à la « nouvelle géographie », ne serait-il pas plus pertinent d'intégrer la revue des *Travaux de l'Institut de Géographie de Reims* fondée en 1969 par Roger Brunet ? Cette question amène à préciser qu'au moment de la constitution de ce corpus, cette revue ne faisait pas encore partie des collections disponibles sous forme numérique. Cette remarque rappelle l'aspect opportuniste et problématique de la construction de ce premier corpus.

Face à ces doutes, une exploration a été commencée sous la forme d'une enquête visant à soumettre ce problème directement aux chercheurs en géographie. Un questionnaire en ligne a été créé avec comme question centrale : « Pouvez-vous citer entre 5 et 10 revues scientifiques de géographie universitaire française que vous jugez comme les plus importantes pour comprendre le champ disciplinaire et ses évolutions depuis les années 1890 (sans prendre en compte les revues consacrées uniquement à la géographie physique) ? ». Un ensemble d'autres questions (sur la spécialité de l'enquêté, son âge, sa formation…) était présent pour permettre une analyse détaillée des résultats. L'enquête a été construite en utilisant le logiciel *LimeSurvey* avec l'aide technique de la MSH Lyon Saint-Etienne. Elle a commencé à être testée par des utilisateurs en petit nombre.

Après avoir compris l'effort nécessaire pour obtenir un grand nombre de résultats et après avoir perçu que ce procédé reportait le problème de la représentativité du corpus sur celui de la construction d'un échantillon, j'ai volontairement arrêté cette expérimentation. Sur le fond, une telle enquête aurait été très intéressante, mais cela ne m'a pas semblé pouvoir constituer une justification pertinente d'un corpus de revues représentatif de la géographie française. Il faut préciser qu'en parallèle d'autres problèmes étaient apparus, m'invitant à prendre cette décision quelque peu radicale. Cette enquête aurait sans aucun doute demandé un travail important pour une solution qui restait finalement insatisfaisante. Parmi ces problèmes parallèles, il y avait notamment celui de l'obtention et de la préparation des données.

2. Obtention et préparation des données

Une convention a été signée tout d'abord avec l'UAR *Persée* pour obtenir leurs données comme le préconise la loi « Pour une République Numérique » adoptée en 2016. Il faut

souligner le caractère novateur de la démarche qui a obligé l'équipe de *Persée* à établir un premier modèle de convention qui d'ailleurs n'a pas été validé par les services juridiques. Nathalie Fargier, alors responsable de l'UAR *Persée,* a accepté de me laisser tout de même l'accès à leurs données avant même la signature officielle, ce qui a été une aide non négligeable. D'autres conventions ont été également signées par la suite avec les équipes d'*OpenEdition Journals* et *Cairn* pour obtenir les numéros les plus récents qui ne sont pas sur le portail *Persée*. Mes expériences dans ce domaine m'amènent à penser que ces procédures ne sont jamais de simple formalité. Ce problème de l'accès aux données constitue à mon avis une forte limitation des recherches dans ce domaine. Un long processus de rencontres, d'explications, de formalisation et de relances a été indispensable en amont.

Par rapport à la liste de périodiques précédemment présentée (*cf.* Tableau 1), un premier problème pris en charge a été celui de la revue *Géographie et cultures*. Tous les numéros de 1993 à 2005 n'existant pas en version numérique, j'ai eu la chance de bénéficier d'un don de Paul Claval pour remédier à cela. Les numéros ont été ensuite numérisés et traités en utilisant un logiciel de Reconnaissance Optique de Caractères (OCR) avec l'aide de la MSH Lyon Saint-Etienne. Toutefois, le résultat obtenu au terme de ce processus est le texte brut des articles. Ce détail est loin d'être négligeable, car en parallèle, tout un travail de nettoyage a été pensé à partir des annotations réalisées par l'UAR *Persée* pour améliorer les données. En effet, il est pertinent, par exemple, d'exclure des éléments comme les bibliographies[22] ou les notes de bas de page qui obéissent à des règles spécifiques de construction avant d'effectuer certaines analyses textuelles. Or, ce travail de nettoyage et d'adaptation des analyses en fonction des éléments n'a de sens que s'il est réalisé sur l'intégralité du corpus. Malheureusement, le travail effectué de récupération des données a permis de s'apercevoir que ce problème n'était pas limité à la revue *Géographie et Cultures.* Dans le cas par exemple de la revue *Hérodote*, il est possible de récupérer facilement tous les fichiers PDF directement à partir du portail *Gallica*[23] mais la situation est ensuite similaire. Il n'existe pas d'annotation des différents éléments structurant les articles (résumé, notes de bas de page, figures, bibliographie…).

Dans le cadre d'une thèse, il est difficile d'envisager de faire tout un travail manuel d'annotation sur une si grande masse de données. Soit, il faut oublier tout le travail d'annotation réalisé par l'UAR *Persée*, soit, il faut essayer d'ajouter de manière automatisée les informations manquantes. Un outil nommé *Grobid*[24] (*GeneRation Of BIbliographic*

[22] Des outils ne relevant pas du domaine de l'analyse textuelle comme les méthodes bibliométriques peuvent être utilisées par ailleurs.
[23] https://gallica.bnf.fr consulté le 18/09/2023.
[24] https://github.com/kermitt2/grobid consulté le 18/09/2023.

Data), notamment utilisé par une autre plateforme de diffusion de littérature scientifique, à savoir *ISTEX*[25], a été testé. Les données d'entraînement de cet outil sont constituées d'articles plutôt récents de sciences dures. Cette situation explique une obtention initiale de mauvais résultats, car les articles étaient dans mon cas assez différents de ceux utilisés pour entraîner le modèle. Une phase de génération de nouvelles données d'entraînement sur des articles de la revue *Hérodote* a été réalisée, mais ce travail est particulièrement long. Après deux semaines, les résultats obtenus avec ces nouveaux articles annotés n'étaient pas vraiment satisfaisants au sens où il y avait toujours beaucoup d'erreurs d'annotations automatiques. Il est évident que les données d'entraînement produites n'étaient pas assez nombreuses, mais Patrice Lopez, le créateur de *Grobid*, qui a été consulté suite aux résultats de cette expérimentation, a mis aussi en avant une qualité de numérisation insuffisante en entrée.

Pour les revues nativement numériques hébergées par les portails *OpenEdition Journals* ou *Cairn*, il existe des données structurées permettant de différencier les différents éléments des articles (titre, résumé, figure…). Il faut tout de même souligner que les schémas informatiques[26] utilisés pour encoder ces données structurées ne sont pas identiques suivant les portails. Si l'optique est de travailler sur l'ensemble de ces données, il y a par conséquent un travail d'homogénéisation non négligeable à fournir. Ce dernier suit la logique du plus petit dénominateur commun. Afin d'expliquer cette expression, imaginons qu'il soit décidé d'enlever les références biographiques des auteurs (souvent au début ou à la fin de chaque article). En effet, ces références entraînent des répétitions de termes, comme « université » ou « laboratoire », qui peuvent ensuite apparaître statistiquement sans être vraiment très signifiantes. Si cette action n'est possible que sur une partie du corpus, il est alors préférable de ne pas la réaliser. En effet, une suppression partielle risque de créer une forte dissymétrie statistique : dans la partie du corpus où cette action n'a pas été effectuée, ces termes vont être artificiellement surreprésentés. Une solution possible est d'enlever ces mots spécifiques après coup, mais définir le périmètre des termes à enlever est toujours une opération délicate. Quand on additionne toutes les opérations qui viennent d'être mentionnées, il est facile de comprendre pourquoi la tâche de préparation des données est vite devenue problématique.

[25] https://www.istex.fr/ consulté le 18/09/2023.

[26] Le terme schéma fait ici référence à l'organisation des données, à la structure employée pour les stocker. Derrière un même nom de format comme le « xml », il peut exister des schémas différents.

3. Choix des méthodologies d'analyse

En parallèle, un travail a été effectué pour définir plus précisément ce qui était cherché. Ce qui avait été défini dans le projet initial était d'étudier les évolutions disciplinaires, mais une telle perspective est particulièrement étendue. De manière sous-jacente, cela pose la question complexe de la définition d'une discipline, sujet sur lequel il existe de multiples approches, des plus réalistes aux plus post-modernistes remettant en question l'existence même des disciplines (Fabiani 2012). Pour essayer d'avancer concrètement, nous avions défini avec mes directrices de thèse cinq grands angles d'attaque :

- La structuration sociale
- Les objets et les thèmes
- Les méthodes et les argumentaires
- Les modes d'expression : la textualité et les autres figures
- Les lieux

Il y a eu ensuite une réflexion pour réaliser une construction par articulation : entre, d'un côté, ce qu'il est possible de calculer avec la définition d'indicateurs et, de l'autre, ce qu'il est le plus intéressant de relire avec la constitution de thèmes d'intérêt. Le Tableau n°2 est présenté pour illustrer le début de travail réalisé. L'ensemble est organisé suivant les cinq angles d'attaque précédemment définis. Cette présentation n'est pas aboutie au sens où les indicateurs sont parfois des pistes méthodologiques assez imprécises. Si j'ai choisi de diffuser tout de même ce travail inabouti, c'est surtout parce qu'il donne une idée de l'ambition qui guidait alors la recherche.

Angle d'attaque	Thème d'intérêt	Indicateur
1	Entre soi et échange entre les revues	Nombre d'auteurs en commun entre les deux revues sur différentes périodes
1	Sous-groupes au sein de chaque revue	Communautés dans réseaux de co-auteurs
1	Effet générationnel	Âge moyen des auteurs au moment de la publication
1 et 2	Réseau épistémique[27]	Graphe bi-partite des auteurs et des thèmes
2	Degré de spécialisation	Nombre médian de mots dans le titre
2	Variation lexicale	Mots spécifiques par période et revue
2	Variation thématique	Thématiques spécifiques par période et revue
3	Évolution des mots de coordinations	Mots de coordination les plus fréquents par période et revue
3	Évolution des plans	Analyse d'échantillon
3	Références bibliographiques	Références les plus citées par période et par revue
4	Richesse du vocabulaire	Nombre de mots utilisés par période et revue
4	Variation des figures	Type de figure (photos, cartes) et place dans le texte
5	Évolution des lieux étudiés	Détection des entités de lieux et évolution par revue et période

Tableau n°2 : Les premières pistes de travail et indicateurs retenus.

Les numéros des angles d'attaque renvoient à l'ordre de leur présentation sur la page précédente

Chaque piste ouvre une exploration bibliographique et des difficultés méthodologiques. Au niveau des indicateurs définis, certains sont faciles à calculer comme le nombre médian de mots dans les titres. D'autres indicateurs, comme l'évolution des lieux ou des bibliographies, sont beaucoup plus complexes à étudier. Des systèmes de détection doivent être utilisés en amont. Il existe déjà sur ces sujets de nombreux travaux et ces champs de recherche sont dynamiques. Une première difficulté a été de faire face à cette profusion de pistes et de gérer celles qui étaient les plus complexes. Dans cette perspective, il est vite

[27] *Cf.* Bibliographie : (Roth 2008).

apparu qu'une thèse n'est sûrement pas un support pertinent pour présenter des expérimentations nombreuses, mais partielles. Ce cadre implique en effet la nécessité de réaliser en amont pour chaque piste un état des lieux et de justifier à chaque fois le choix méthodologique retenu. Avec ces contraintes, un tel projet scientifique n'était pas tenable à l'échelle d'une thèse menée par une seule personne.

À la suite de ce constat, il a été décidé d'effectuer une réduction du projet. En allant au-delà de la question de la faisabilité de chacune des pistes, comment justifier le choix de se consacrer à certaines et non à d'autres ? À ce stade de la recherche, cette question est restée pendant un long moment sans réponse. Il y a eu un tiraillement fort entre la richesse pressentie de ces multiples pistes méthodologiques et la nécessité d'en choisir une plus spécifiquement. Mon orientation méthodologique initiale était axée sur des outils d'analyse textuelle du fait de mon parcours antérieur (*cf.* section Chap1.II). Cependant, l'ouverture et l'explicitation de toutes ses possibilités d'analyses ont entraîné un élargissement vécu dans un premier temps comme une chance et un enrichissement. La prise de conscience qu'il fallait *a contrario* aller dans le sens d'une réduction ne s'est faite que progressivement. La longue phase d'indécision quant au choix à effectuer entre ces différentes pistes méthodologiques a entraîné un certain malaise, car il s'agissait d'arriver à définir *in fine* le cœur du projet : ce qui était précisément cherché.

Enfin, pour achever cette section sur les difficultés liées à la réalisation du projet initial, il est nécessaire d'aborder le projet d'entretiens avec les différents acteurs des revues et la problématique de son articulation avec les analyses quantitatives.

4. Les entretiens : quelle articulation quantitatif-qualitatif ?

Un travail a été mené avec Isabelle Lefort pour construire un guide d'entretien. L'objectif global étant de comprendre le rôle des revues dans la construction et l'évolution de la géographie française, plusieurs thématiques ont été définies :

- L'intégration et la première découverte du monde des revues.
- Les avantages et inconvénients liés au fait d'être un membre actif dans un ou des périodique(s).
- Le fonctionnement avec les processus de prise de décision et de renouvellement des comités de rédaction et de direction.
- Les polémiques et les débats de fond.
- L'évolution de la ligne éditoriale.
- La problématique de la concurrence et de la complémentarité avec les autres revues.

- Le thème de la matérialité de la fabrication : que se passe-t-il concrètement après l'envoi d'un article à la revue ? Qui le reçoit et comment est-il partagé entre les différents membres afin d'être évalué ?
- Le poids des éditeurs dans l'évolution de la revue.
- La question de l'impact du lectorat sur les choix éditoriaux réalisés.
- Les évolutions plus récentes notamment avec le passage au numérique.

Puis, nous avons effectué un tour d'horizon des personnes à interroger. Il y a eu une réflexion pour choisir des personnes aux statuts variés : directeur / membre du comité de rédaction / secrétaire de rédaction. Ce choix reposait avant tout sur les connaissances d'Isabelle Lefort et de son réseau professionnel. Cette méthodologie, insatisfaisante d'un point de vue statistique, provient du fait que la définition d'un échantillon représentatif est loin d'être simple sur ce sujet. Nous avons décidé, pour commencer, d'interroger les grands acteurs encore vivants de la génération qui a participé à la « crise de la géographie » et qui ont fondé des revues majeures dans le champ de la géographie : Yves Lacoste, Roger Brunet et Paul Claval.

Il en a résulté des entretiens semi-directifs assez longs (près de deux heures) avec quelques éléments intéressants, mais une véritable insatisfaction de ma part sur plusieurs plans. En effet, ces expériences m'ont appris que beaucoup de détails des controverses ont été gommés par le temps. Le point, sûrement le plus crucial par rapport au travail initialement pensé, a été la difficulté de penser et de construire l'articulation avec la dimension quantitative. Il avait été initialement prévu de partager avec les acteurs des revues quelques résultats chiffrés obtenus à partir des indicateurs précédemment explicités. L'idée était de les faire participer aux processus de réflexion et d'interprétation que peuvent susciter de tels résultats. De nouvelles pistes d'analyse étaient ainsi susceptibles d'apparaître. Dans certains cas, celles-ci auraient pu être ensuite testées quantitativement. Or, les nombreuses difficultés précédemment détaillées expliquent que durant une période initiale assez longue, il ne m'a pas été possible de discuter de résultats puisque ces derniers n'existaient pas. Par conséquent, des entretiens sans cette interaction avec la dimension quantitative ont été réalisés dans un premier temps. Par la suite, plus le travail avançait, plus une certaine inquiétude s'installait, car cette interaction entre résultats quantitatifs et entretiens des acteurs qui constituait un des cœurs du projet initial, n'arrivait pas à prendre forme et à être mise en œuvre.

Toutes ces difficultés sont à l'origine d'une redéfinition et d'une réorientation profonde de la problématique après plus de huit mois intenses d'investigation. Ce processus m'a conduit concrètement à abandonner une grande partie du travail entrepris. Une valorisation

de ce travail a eu lieu avec la rédaction d'un article (Beligné, Lefort, et Loudcher 2020b) publié dans la revue *Humanités numériques*. Avant que cette dernière ne prenne l'initiative de changer le titre de ce texte, cette publication était intitulée : « Pour un renouvellement épistémologique disciplinaire à partir des revues francophones de sciences sociales : potentialités et verrous ». Ce qui est explicité dans cet article reprend quelques problématiques développées ici, mais sous un angle différent. Il y a eu notamment une tentative de montée en généralité puisque le propos ne se limite pas au cas particulier de la géographie.

De manière inverse, l'écriture du présent développement a été l'occasion de rendre compte de manière plus spécifique et précise du défrichement effectivement réalisé. C'est cette phase de découverte d'un « terrain », avec ses problèmes d'accessibilité, ses potentialités et ses contraintes que cette écriture permet d'appréhender. La tension, entre ce qui est théoriquement prévu et ce qui est effectivement réalisable dans le temps imparti, est un élément essentiel pour comprendre toute recherche. La gestion de cette friction sur plusieurs années constitue sans aucun doute un apprentissage majeur de cette période de doctorat. Face aux difficultés rencontrées, il a finalement été décidé de modifier en profondeur le projet.

II. Le projet central : mise à l'épreuve d'une proposition épistémologique par des outils d'analyse textuelle

La dénomination de « projet central » a été ici retenue. Ce choix provient de raisons qui vont au-delà de l'argument de sa position entre le projet initial et celui finalement retenu. En effet, contrairement au projet précédent, ce projet central n'a pas été profondément remis en cause. Il y aura certes des ajustements non négligeables avant de définir le projet finalement retenu (*cf.* section Chap2.III), mais le cadre théorique et problématique de ce projet central ne sera pas foncièrement modifié lors de cette évolution.

La solution trouvée face aux difficultés précédemment explicitées a consisté à définir une problématique plus précise en partant d'une lecture épistémologique déjà connue. Ce faisant, ce doctorat participe d'une pratique continuée dans la production scientifique qui consiste toujours à s'adosser et à prendre en charge des travaux déjà réalisés. La référence initiale, ici choisie, est une construction intellectuelle développée par un géographe, Olivier Orain, dans sa thèse intitulée *Le plain-pied du monde : postures épistémologiques et pratiques d'écriture*

dans la géographie française au vingtième siècle[28] (2003). Le cadre théorique qu'il mobilise est celui de *La structure des révolutions scientifiques* (Kuhn 2008, *1962*) élaboré pour penser l'évolution des sciences de la nature. Il est ici nécessaire tout d'abord de présenter plus en détail ce modèle explicatif kuhnien en rappelant son contexte d'apparition, son contenu, sa réception et son évolution.

1. Présentation du modèle kuhnien

L'ouvrage le plus connu de Thomas Kuhn est la *Structure des révolutions scientifiques* (SRS) paru en 1962 dans lequel il expose les bases de son modèle explicatif des dynamiques scientifiques.

a. Contexte de la SRS et son contenu

Le contexte intellectuel des années 1950 à 1970, période de l'élaboration et des premières réceptions de ce livre, est notamment marqué par le structuralisme. Ce courant provenant de la linguistique saussurienne se diffuse alors dans l'ensemble des sciences humaines (Claude Lévi-Strauss pour l'anthropologie, Jean-Pierre Vernant pour l'histoire des religions, Roland Barthes pour la sémiologie, plus tardivement Roger Brunet pour la géographie…). L'objectif de ce courant intellectuel est de rechercher des invariants permettant d'expliquer une variété de situations dans un domaine donné. L'exemple le plus connu est l'ouvrage *Les Structures élémentaires de la parenté* (Lévi-Strauss 2002, *1949*). La proposition de Thomas Kuhn avec la SRS s'inscrit dans cette dynamique puisqu'elle permet de penser de manière structurale le développement des sciences de la nature.

Dès sa préface, Thomas Kuhn précise que c'est en partant de l'opposition entre sciences sociales et sciences de la nature qu'il a bâti son concept central de « paradigme » :

> « la pratique de l'astronomie, de la physique, de la chimie ou de la biologie ne fait pas naître de ces controverses sur les faits fondamentaux qui semblent aujourd'hui endémiques parmi les psychologues ou les sociologues. C'est en essayant de découvrir l'origine de cette différence que j'ai été amené à reconnaître le rôle joué dans la recherche scientifique par ce que j'ai depuis appelé les paradigmes, c'est-à-dire les découvertes scientifiques universellement reconnues qui, pour un temps fournissent à une communauté de chercheurs des problèmes types et des solutions » (Kuhn 2008, 11).

[28] Cette thèse peut être lue à partir du lien : https://theses.hal.science/tel-00082408/file/Orain_these.pdf consulté le 18/09/2023.

À partir de ce concept et en s'appuyant sur plusieurs cas historiques (du géocentrisme de Ptolémée à l'héliocentrisme de Copernic, de la gravitation de Newton à la relativité générale d'Einstein…), Thomas Kuhn élabore un modèle de développement de la science comportant plusieurs étapes.

La première phase de ce modèle correspond à une période de « pré-science » durant laquelle plusieurs écoles s'affrontent sans qu'aucune ne soit vraiment dominante. Ce n'est qu'avec la reconnaissance par une communauté de chercheurs de la supériorité d'une théorie sur les autres qu'une « science adulte » advient. Une phase de « science normale » se met alors en place. Durant cette période, l'objectif est alors surtout de résoudre des problèmes d'une façon qui est déjà déterminée par le paradigme en cours. Pourtant, des résultats qui ne sont pas forcément attendus dans ce cadre sont également produits. Ces derniers sont souvent déconsidérés, car ils sont métaphoriquement effacés par le paradigme dominant qui tend à exclure tous les éléments susceptibles de le remettre en cause. Il arrive toutefois qu'une de ces incohérences prenne de l'importance. Il y a alors prise de conscience de ce que Thomas Kuhn a appelé une « anomalie » : « c'est-à-dire l'impression que la nature, d'une manière ou d'une autre, contredit les résultats attendus dans le cadre du paradigme qui gouverne la science normale » (Kuhn 2008, 83).

Si l'anomalie persiste en continuant à être explorée, un état de « crise » s'installe : plusieurs propositions théoriques s'affrontent. Il y a une période d'insécurité scientifique qui prend le nom de « science extraordinaire » dans la dénomination kuhnienne. Au terme de cette période de crise, un phénomène que cet auteur appelle « conversion » a lieu. Cette dénomination renvoie au fait que, pour Thomas Kuhn, ce sont des éléments dépassant un cadre strictement rationnel qui finissent par emporter l'adhésion de la communauté de chercheurs confrontés à la crise : « la concurrence entre paradigmes n'est pas le genre de bataille qui puisse se gagner avec des preuves » (Kuhn 2008, 204). Pour faire comprendre plus finement le changement, Thomas Kuhn emprunte des arguments à la psychologie de la forme (*Gestalt theory*, d'où le nom de renversement gestaltique). Ce courant met en avant qu'une même forme peut être vue successivement par une même personne de manière totalement différente (par exemple, l'image célèbre du canard-lapin[29], notamment utilisée par Wittgenstein).

Au terme d'une conversion du regard, une personne peut voir la forme qu'elle ne voyait pas au départ. Pourtant, elle ne peut pas voir les deux en même temps. Il y a pour Thomas Kuhn une incommensurabilité des paradigmes. La conversion est définitive, au sens où il ne

[29] https://fr.wikipedia.org/wiki/Canard-lapin consulté le 18/09/2023.

peut pas avoir de retour en arrière[30] : il est impossible pour un chercheur d'être à la fois dans le paradigme antérieur et le suivant. La phase révolutionnaire se termine donc quand un nouveau paradigme s'impose parmi les différentes théories qui étaient jusqu'ici concurrentes face à l'anomalie. Une nouvelle phase de « science normale » commence alors. Cette période dure jusqu'à ce qu'une autre « anomalie » apparaisse. Le processus qui vient d'être décrit recommence alors de manière cyclique.

Dans la thèse kuhnienne, comme le processus d'abandon d'un paradigme ne résulte pas directement d'une réfutation au sens poppérien du terme, des explications exogènes (et non plus strictement endogènes à la science) peuvent être mobilisées pour rendre compte de changement. Cette voie, initiée mais peu développée par la SRS, a constitué historiquement une innovation[31] remarquée. Elle explique autant un accueil enthousiaste de l'approche kuhnienne par certains chercheurs (Barber 1963) que l'existence d'une forte opposition (Lakatos et Musgrave 1970). Plusieurs points de la théorie kuhnienne ont été très débattus : la notion d'incommensurabilité entre deux paradigmes, le caractère relativiste de cette théorie, l'imprécision dans la définition de la notion de paradigme (Masterman 1970)… Globalement, le schéma kuhnien a connu une grande popularité du fait de la nouveauté qu'il a apportée dans la compréhension des dynamiques scientifiques, mais a suscité aussi de fortes polémiques. Ces dernières ont d'ailleurs conduit Thomas Kuhn à publier une postface à son ouvrage en 1969.

b. La postface de la SRS

Cette postface apporte plusieurs précisions importantes. Thomas Kuhn y détaille notamment la définition du concept de paradigme en proposant deux acceptions majeures et distinctes :

- La première renvoie à ce que cet auteur appelle une « matrice disciplinaire ». Cette dernière est composée et définie pour Thomas Kuhn par quatre éléments :
    - Les généralisations symboliques qui sont directement exprimées par des formules mathématiques[32] ou par des « formes verbales »[33] (Kuhn 2008, 249).

[30] Sur ce point, l'image du canard-lapin ne fonctionne plus car il est toujours possible dans cet exemple de revoir la forme vue antérieurement.

[31] Il existe bien entendu des auteurs ayant esquissé cette perspective avant Thomas Kuhn, notamment (Fleck 2008, 1935), mais elle a été sans conteste plus fortement discutée et diffusée après la SRS.

[32] Comme « f = ma » (*Ibid* 2008, 249).

[33] Par exemple, « les éléments se combinent dans des rapports de poids constants » (*Ibid* 2008, 249).

- La partie métaphysique est constituée par des croyances[34] qui forment des modèles ontologiques ou heuristiques[35]. Cet élément fournit à la communauté scientifique « des métaphores et des analogies préférées ou permises » (*Ibid* 2008, 252).
- Les valeurs : celles, ayant le plus de force selon Thomas Kuhn, sont celles prédictives et quantitatives. Il existe aussi des valeurs non quantitatives[36] utilisées pour juger des théories.
- Les exemples communs (aussi appelés « exercice-type ») qui sont des problèmes[37] dont la solution est connue et sert de base d'apprentissage pour les étudiants.

- La seconde acception de paradigme est un sens réduit au quatrième élément de cette matrice, c'est-à-dire les exemples communs. C'est en effet pour Thomas Kuhn l'élément le plus important de sa théorie. C'est par la réalisation de ces exemples communs que les étudiants assimilent « une manière de voir autorisée par le groupe et éprouvée par le temps » (*Ibid* 2008, 258).

Plusieurs autres points sont abordés dans cette postface. S'il n'est pas possible de tous les développer ici, il faut souligner qu'à la métaphore visuelle du changement gestaltique originellement mis en avant pour expliquer l'incommensurabilité, Thomas Kuhn ajoute une explication s'appuyant davantage sur l'apprentissage d'un langage différencié. Cette évolution a pu être identifiée comme l'origine d'un « tournant linguistique » (Leduc 2007, 14) qui ne cessera de s'affirmer dans les réflexions ultérieures de cet auteur (Kuhn 2000).

Un autre point abordé dans cette postface, important pour ma recherche, est la question du transfert du modèle au-delà de l'explication des dynamiques des sciences de la nature. Thomas Kuhn aborde cette problématique dans une optique qui n'est pas spécifiquement tournée vers les sciences humaines, mais qui traite plus globalement du transfert pour n'importe quel champ d'activité. Après avoir fait part de son étonnement vis-à-vis de ces applications qu'il n'avait pas prévues et anticipées[38] en écrivant son ouvrage, il explique ce phénomène en affirmant :

---

[34] Comme « la chaleur est l'énergie cinétique des parties constituantes des corps » (*Ibid* 2008, 250).
[35] Par exemple, « le circuit électrique peut être considéré comme un circuit hydrodynamique en état d'équilibre » (*Ibid* 2008, 250).
[36] Comme « la cohérence interne » (*Ibid* 2008, 252).
[37] Par exemple, le problème du « plan incliné » (*Ibid* 2008, 258).
[38] « Je vois ce qu'ils [les personnes ayant appliqué ses idées à d'autres domaines que les sciences dures] veulent dire et ne voudrais pas décourager leur désir d'élargir ma thèse, mais leur réaction m'a néanmoins surpris » (*Ibid* 2008, 282).

> « Dans la mesure où ce livre décrit le développement scientifique comme une succession de périodes traditionalistes, ponctuées par ruptures non cumulatives, ses thèses sont sans aucun doute applicables à de nombreux domaines » (Kuhn, 2008, 282).

Toutefois, il finit par rappeler quelques différences et particularités du domaine scientifique que son schéma visait plus spécifiquement à expliquer : « absence ou tout au moins rareté relative des écoles concurrentes dans les sciences développées » (*Ibid* 2008, 283), auto-évaluation d'une communauté et nature de la formation axée sur la résolution d'énigmes.

De manière globale, s'il est vrai que cette postface apporte des réponses à plusieurs questions et critiques, elle n'a pas mis fin aux interrogations et aux débats. Il n'est pas si facile de résumer l'ensemble des critiques, car elles viennent de plusieurs fronts :

- Tout d'abord, des épistémologies internalistes, soucieuses des procédures cognitives de production des connaissances et d'une certaine objectivité : l'ouverture à des explications de type externaliste et le caractère relativiste de sa théorie (que conteste Thomas Kuhn dans sa postface) expliquent un débat fort sur la définition et les conditions de possibilité de la science.
- Ensuite, des épistémologies plus externalistes : parmi les chercheurs se revendiquant de ces épistémologies, de nombreux sociologues, notamment ceux orientés vers un « programme fort » (Bloor 1976), pensent que les explications socio-politiques restent insuffisamment mises en avant dans l'approche kuhnienne.
- Enfin, de ceux qui essayent de penser un entre-deux dans ce dilemme internaliste et externaliste de manière un peu différente que le fait Thomas Kuhn : ces réflexions peuvent donner lieu à la naissance d'autres concepts comme celui de « programme de recherche » (Lakatos 1994). Il faut souligner que l'articulation entre approches internaliste et externaliste est loin d'être clarifiée chez Thomas Kuhn. Alors que de nombreux auteurs insistent sur le côté externaliste (Hacking 2008), cela n'empêche en rien des lectures plus internalistes, comme celle réalisée par Olivier Orain dans sa thèse (2003).

Thomas Kuhn n'a cessé de faire évoluer sa théorie tout au long de sa carrière, mais notons qu'Olivier Orain s'appuie surtout, et de manière détaillée, sur la SRS et sa postface. Il se réfère occasionnellement aux derniers travaux de Thomas Kuhn, uniquement pour reprendre et proposer un développement du tournant linguistique. Quelques-unes des propositions kuhniennes ultérieures méritent cependant d'être brièvement développées dans l'optique de mes réflexions de thèse.

c. Quelques propositions kuhniennes post-postface

Une première évolution de la pensée kuhnienne concerne la notion d'incommensurabilité qui devient moins radicale et plus localisée (Kuhn 2000) : contrairement à ce qui avait été affirmé pour répondre aux premières controverses de la SRS, il est alors reconnu explicitement par Thomas Kuhn que les problèmes de traduction entre deux paradigmes peuvent ne porter que sur un nombre limité de termes. Cette inflexion théorique donne de plus en plus d'importance, dès les années 1980, aux notions de taxinomie[39] et d'interprétation (Kuhn 1982). Si un nouvel élément conduit à modifier la structure taxinomique liée à une théorie, il y a alors impossibilité stricte de traduction et l'apparition d'une incommensurabilité. Il est possible de comprendre d'anciennes théories, mais cela se fait au prix d'un travail non plus de traduction, mais d'interprétation pour se réapproprier une taxinomie passée.

Thomas Kuhn ne réduit pas cependant le changement de paradigme à un changement taxinomique comme le montre son opposition à la perspective nominaliste présentée par Ian Hacking (1993). Ce dernier, pour expliquer l'expression de « changement de monde », argue que les scientifiques sont obligés de regrouper les individus en espèces pour comprendre le monde. Un changement de paradigme correspond alors pour Ian Hacking à un changement de classement des individus. L'opposition de Thomas Kuhn à cette conception s'appuie sur des arguments concrets[40], mais relève aussi d'un positionnement plus profond. En effet, par rapport à cette position nominaliste, Thomas Kuhn s'est qualifié lui-même de « réaliste non converti »[41] (1993, 415).

Il ne remet pas en cause la théorie de la référence, c'est-à-dire l'idée qu'il existe des relations entre les mots et des réalités au moins en partie extra-linguistiques. Dans le même temps, il est vrai que Thomas Kuhn n'est pas revenu sur l'idée qu'il a exprimée dans la postface de la SRS, c'est-à-dire qu'il n'y a pas selon lui d'« adéquation entre l'ontologie d'une théorie et sa contrepartie *réelle* dans la nature » (Kuhn 2008, 280). Ce positionnement explique une forte tension intellectuelle qui permet de mieux comprendre une recherche qui n'a jamais été complètement stabilisée. Pour Pierre Leduc qui a étudié en détail le problème

[39] Le mot *taxinomie* provient des mots grecs t*axis* (« placement », « classement », « ordre ») et *nomos (*« loi », « règle »). Un exemple classique de taxinomie est le classement biologique : Règne, Classe, Cohorte, Ordre, Famille, Tribu, Genre, Espèce. La pensée kuhnienne sous-entend qu'une partie au moins du lexique d'une science spécialisée peut être structurée sous la forme d'une taxinomie.

[40] Par exemple, l'idée qu'il n'est pas pertinent pour lui de considérer la construction d'un concept comme celui de force comme un regroupement d'individus (Leduc 2007, 148).

[41] Traduction personnelle de « unregenerate realist » (Kuhn 1993, 415).

de l'incommensurabilité chez Thomas Kuhn, le problème reste irrésolu et, par conséquent, à la charge de ceux qui revendiquent des lectures kuhniennes.

> « S'il est vrai, comme le remarque Claude Panaccio, que " l'enjeu de la discussion philosophique n'est pas autre chose, bien souvent, que de savoir à qui revient le fardeau de la preuve », alors je crois bien que le fardeau repose sur les épaules de ceux qui adoptent le point de vue de Kuhn sur l'incommensurabilité" » (Leduc 2007, 359).

Ces quelques précisions étant apportées, il faut souligner qu'Olivier Orain n'est pas le premier géographe à essayer d'importer le concept de paradigme pour penser les dynamiques de cette discipline. Il y a même sur la question plutôt un historique non négligeable à considérer.

2. Importations et usages du concept kuhnien de paradigme en géographie

a. Du côté de la géographie anglo-saxonne

La première utilisation célèbre du schéma kuhnien en géographie est celle de David Harvey dans son livre *Explanation in Geography* (1969). S'appuyant sur la confrontation des années 1950 entre Richard Hartshone (représentant de la position idiographique) et Fred K Schafer (représentant de la position nomothétique), l'utilisation des modèles comme porteur d'une nouvelle dimension scientifique pour la géographie est affirmée par William Bunge (1962) et ensuite reprise par David Harvey (1969). Il faut souligner qu'il existe dans sa démarche une forte cohérence à utiliser le schéma kuhnien puisqu'il participe à un mouvement transversal à toutes les sciences sociales revendiquant la mise à jour d'invariants sous la forme de structure ou de modèle (*cf.* section II.1.a). La conclusion de Stephen Gale (1972) dans le compte-rendu qu'il réalise du livre de David Harvey permet de mieux contextualiser ce dernier :

> « En dernière analyse, je vois *Explanation in Geography* comme une présentation de la philosophie et de la méthodologie de la géographie des années 1960 ; heureusement, les chercheurs férus de méthodologies des années 1970 poursuivront le débat et apporteront des éclaircissements supplémentaires sur ces questions importantes. La question centrale de ce débat sera décisive : dans quelle mesure les méthodologies développées pour les mathématiques et les sciences physiques peuvent-elles être appliquées aux sciences sociales ? »[42] (Gale 1972, 317).

[42] Traduction personnelle de « In the final analysis, I see *Explanation in Geography* as a presentation of the philosophy and methodology of geography in the 1960s; hopefully, the methodologists of the 1970s will continue the debate and bring about a further clarification of these important issues. The central question in

Dans cette perspective d'un transfert méthodologique restant en suspens, il est facile de comprendre pourquoi la difficulté d'utiliser un modèle, conçu initialement pour penser l'évolution des sciences de la nature[43], n'a pas été profondément prise en charge par les premiers chercheurs soucieux surtout de promouvoir fortement la géographie théorique et quantitative. Le schéma kuhnien leur a permis principalement de revendiquer de manière performative une « révolution scientifique » en cours.

La « new geography » (Gould 1968) s'affirme et se développe à travers la revendication de cette rupture légitimée par la théorie kuhnienne. Une perspective critique n'émerge qu'à partir de la fin des années 1970 avec l'expression des premiers doutes sur la capacité du modèle kuhnien à rendre compte des transformations de la discipline (Johnston 1978; Graves 1981; Stoddart 1981). Face à ce mouvement, certains chercheurs tentent de réhabiliter Thomas Kuhn en apportant des précisions sur sa théorie (Mair 1986). D'autres se tournent vers d'autres horizons épistémologiques comme les programmes de recherche de Lakatos (Wheeler 1982).

b. Du côté de la géographie française

Parallèlement à la diffusion de la « nouvelle géographie » en France dans les années 1970, le terme de paradigme a lui aussi été importé. S'il existe une attraction indéniable pour le modèle des sciences de la nature, il faut également souligner assez rapidement l'existence chez certains chercheurs d'une perspective critique par rapport à l'utilisation du modèle kuhnien. Ce double processus est par exemple très présent dans un article d'Antoine Bailly et Jean-Bernard Racine paru en 1978 dans *L'Espace Géographique*. Un peu piqués au vif par l'affirmation de Jean Piaget sur l'absence d'épistémologie de la géographie, les auteurs reconnaissent tout de même :

> « Comment ne pas remarquer, à travers l'analyse des pratiques, l'imprécision des raisonnements, l'absence d'épistémologie propre ? En fait, pragmatique par essence, notre discipline ne s'est pas encore développée comme une activité scientifique "normale", puisque, après les phases empiriques et inductives, elle n'est pas arrivée aux phases théoriques, déductives et axiomatiques. Cet arrêt dans l'évolution est-il inévitable, fondé sur la nature de la discipline ? Ce serait alors à la réflexion épistémologique d'en rendre compte, et de le légitimer » (Bailly et Racine 1978, 8).

---

this debate will be decisive: "To what extent can the methodologies developed for mathematics and the physical sciences be applied to the social sciences ?" » (Gale 1972, 317).

[43] Évolution des sciences de la nature reconnue par Thomas Kuhn comme fondamentalement différente de celle des sciences de la société.

Cette citation peut donner lieu à de multiples questionnements sur différents sujets : y a-t-il vraiment une nature de la discipline ? Comment faut-il interpréter cette transition : « En fait, pragmatique par essence » ? Enfin et surtout, s'il n'est pas possible de fonder un paradigme, ne faut-il pas réfléchir au moyen d'une autre épistémologie ?

Les auteurs ne répondent pas directement dans le texte à cette question. En revanche, une proposition sur ce sujet peut être lue dans l'article de Jean-Paul Ferrier, Jean-Bernard Racine et Claude Raffestin, *Vers un paradigme critique : matériaux pour un projet géographique*, paru à la fin de l'année 1978 dans *L'Espace géographique*. Les auteurs affirment :

> « en moins d'une génération, on pourrait croire, à lire ceux qui l'écrivent, que la géographie (quantitative) a subi plus d'une demi-douzaine de révolutions : l'initiale d'abord, la "quantitative", puis la révolution "méthodologique", "conceptuelle", "statistique", "révolution des modèles", révolution "behaviorale", "radicale", et tout dernièrement "axiomatique"… Le concept de "révolution" n'est-il pas dans ces conditions un tantinet abusif ? N'est-ce pas le signe que toute maîtrise d'une nouveauté est un moyen de fonder un pouvoir ? Faut-il lui préférer, ou plutôt lui associer, le concept de paradigme ? » (Ferrier, Racine et Raffestin 1978, 292).

La solution trouvée par ces auteurs est assez paradoxale puisqu'ils proposent finalement d'adopter un nouveau paradigme qu'ils appellent « le paradigme critique » (*Ibid* 1978). Le terme de « critique » est suffisamment large dans sa dénomination pour être difficilement remis en question. Les auteurs ne manquent pas d'ambition à son égard :

> « La géographie, par son incapacité à construire son propre langage, est une sorte de Pologne satellisée, partagée, souffrante, toujours vivante, mais pantelante. Quant au paradigme critique, il est une forme de libération par la résistance active » (*Ibid* 1978, 296).

La rhétorique révolutionnaire de l'époque est utilisée pour affirmer de manière forte une refondation de la géographie par le langage.

À partir des années 1980, une partie des géographes français suit[44] l'évolution anglo-saxonne s'orientant vers une approche plus critique du modèle sous-jacent des sciences dures et du transfert du schéma kuhnien. Cette évolution est bien soulignée par Paul Claval :

> « L'ère des paradigmes ambitieux paraît close en géographie : personne ne croit plus sérieusement que l'on puisse interpréter l'évolution contemporaine de la discipline à partir du schéma trop simple de Thomas Kuhn […] Du coup, les géographes se détournent un peu de l'épistémologie générale et se préoccupent davantage du processus interne de développement de leurs idées et de leurs méthodes » (Claval 1985, 173).

[44] Le compte-rendu de Paul Claval (1981) du travail de D R Stoddart montre qu'il y a eu une connaissance de ces travaux par quelques-uns et une volonté de transmission à une partie de la communauté française.

Face à cette évolution décrite par Paul Claval, il est nécessaire de distinguer deux positionnements :

- Le premier consiste à ne plus utiliser le schéma kuhnien et à se revendiquer d'autres épistémologies. C'est par exemple le cas de Vincent Berdoulay. Son article, *Géographie : lieux de discours,* comporte notamment un paragraphe intitulé « le paradigme introuvable » (Berdoulay 1988, 246). Une approche contextuelle, et non plus structurelle, de l'histoire de la géographie est fortement revendiquée. Le modèle kuhnien est totalement rejeté par Vincent Berdoulay.

- Le second positionnement consiste à reprendre le schéma kuhnien mais de manière plus raffinée. C'est par exemple le cas d'Olivier Soubeyran qui s'en sert dans une approche valorisant la notion d'imaginaire géographique (Soubeyran 1997). C'est également le cas de Marie-Claire Robic (1991) qui se sert du concept de paradigme non plus pour justifier la « nouvelle géographie », mais pour penser la géographie classique. Elle met ainsi en avant un « paradigme du mixte » (Robic 1991) qui aurait pour origine l'œuvre de Paul Vidal de la Blache. Ce paradigme aurait permis « une identité scientifique symbolique tout en permettant un jeu extrême entre ses deux limites, qui sont aussi les horizons, ou les pôles du champ épistémologique, entre l'empirisme et le constructivisme : la description pure et l'explication » (*Ibid* 1991, 57). Cette utilisation du concept de « paradigme » constitue un renversement quelque peu ironique de l'histoire, car la revendication initiale portée par certains « nouveaux géographes » insistait plutôt sur l'absence de paradigme dans la géographie vidalienne. Leur objectif était d'arriver enfin à fonder un paradigme par opposition à cet héritage intellectuel de Vidal de la Blache (1945-1918) et de ses successeurs dont la scientificité était contestée.

Dans une analyse très générale, Olivier Soubeyran met en avant l'idée d'un renversement profond des positionnements épistémologiques par rapport à ce concept de paradigme dans la géographie française des années 1970-1980 :

> « Ce rejet en bloc de Kuhn est bien sûr à la mesure de l'engouement dont ses thèses avaient fait l'objet en géographie (et ailleurs !). Il y a encore quelques années il fallait s'exprimer par "paradigme" sous peine de ne pas être compris et de ne pas être pris au sérieux. Aujourd'hui c'est carrément l'inverse : il faut montrer "patte blanche" lorsque l'on parle de paradigme (en géographie plus qu'ailleurs) comme s'il y avait toutes les chances que personne ne vous comprenne (Ah, ce "paradigme" qui veut tout dire et n'importe quoi !), ne vous prenne plus au sérieux » (Soubeyran 1988, 237).

Par rapport à cette présentation et à l'état des lieux ici réalisé sur la question, un tel renversement est perceptible, mais il n'est pas présent de manière aussi affirmée dans la littérature. La forme d'écriture utilisée par Olivier Soubeyran peut laisser penser que ce

dernier fait aussi référence à des réactions ayant eu lieu dans les colloques. Si cela demeure une hypothèse, ce qui est certain par rapport à l'utilisation de la notion de paradigme, c'est que cette dernière n'a pas été complètement rejetée et a largement survécu aux critiques.

Les années 1990 - 2000 sont en effet marquées par le maintien d'une utilisation du terme, mais sous une forme plus consensuelle. Il y a encore quelques débats sur ce sujet, mais ils sont sûrement moins systématiques et nombreux que ceux de la période précédente. « Changement de paradigme » devient pour certains chercheurs une manière de dire « un changement important » sans vraiment plus d'approfondissement et de plus-value heuristique sur ce point. D'autres épistémologies, notamment les « turns and studies »[45] (Naylor *et al*. 2018) se sont développées, renforçant la mise au second plan de la théorie kuhnienne précédemment amorcée (Claval 1985). Une grande partie des débats s'est déplacée vers d'autres problématiques comme celles liées au post-modernisme (Collectif 2004) ou au tournant spatial (Besse *et al.* 2017). Le paradigme n'est plus un objet chaud de l'épistémologie contemporaine.

Après avoir explicté cette évolution, il est désormais possible de mieux situer le travail d'Olivier Orain.

3. Les usages heuristiques d'Olivier Orain du modèle kuhnien

a. Un positionnement spécifique propice aux débats

Spécifique, tout d'abord parce que le travail d'Olivier Orain peut être vu comme une exception. En effet, dans les années 2000, l'épistémologie kuhnienne est loin d'être au centre des préoccupations. Pour comprendre ce *revival*, il est nécessaire à mon avis de replacer la thèse d'Olivier Orain dans une dynamique plus large, celle de l'équipe *Épistémologie et Histoire de la GéOgraphie* (EHGO) de l'UMR 8504 Géographie-Cités. Dans ce cadre, le rôle de Marie-Claire Robic qui a été la directrice de thèse d'Olivier Orain, mais aussi la directrice de cette équipe depuis 1991, est central. En reprenant le concept de paradigme, Olivier Orain a prolongé une partie de son travail[46] en essayant d'appliquer plus strictement le modèle kuhnien. Il s'est attelé au travail sans conteste le plus difficile : dé-montrer la pertinence d'utiliser ce modèle, ce qui peut être vu comme une tentative de fondation (ou de

[45] Littéralement en français : les tournants et les études.

[46] Cette manière de faire participe de dispositifs et de dynamiques scientifiques courantes dans un laboratoire de recherche. La structure de laboratoire est ainsi propice à l'élaboration et la proposition de nouveaux courants, voire d'écoles.

refondation) d'un courant kuhnien dans le champ de l'épistémologie disciplinaire géographique.

La thèse d'Olivier Orain s'inscrit tout à fait dans la lignée d'un usage raffiné de la théorie kuhnienne. Il utilise ainsi ce modèle comme un outil pour relire le corpus de la géographie française de 1910 à 1985 avec comme prisme « la question du rapport connaissance / réalité » (Orain 2003, 18). Toutefois, cet usage n'est pas sans soulever des interrogations qui continuent de faire débat, notamment quand Olivier Orain affirme dans la conclusion de sa thèse :

> « Nous avons la conviction que la *thick description*[47] kuhnienne "fonctionne" remarquablement bien pour penser les pratiques cognitives et contenus de la géographie française jusqu'au milieu des années 1980 » (*Ibid* 2003, 353).

Il valide ainsi de manière forte l'utilisation du modèle kuhnien qui permet, selon lui, de lire de manière très convaincante la géographie française sur la longue période qu'il a étudiée en détail. Il faut souligner qu'avant cette conclusion, Olivier Orain reconnaît à plusieurs reprises quelques adaptations[48] du modèle par rapport aux sciences de la nature. Toutefois, son propos conclusif tend à aller largement au-delà d'un usage analogique du concept de paradigme. Ce positionnement est affirmé d'une manière assez subtile :

> « On est en droit de se demander si l'analogie est tentante pour des raisons de stricte homologie ou s'il ne faudrait pas reprendre les constructions kuhniennes dans une autre perspective que celle des sciences expérimentales, dans un projet de théorie nominaliste de la connaissance » (*Ibid* 2003, 353).

Le premier point à souligner est que les limites du modèle sont largement mises de côté dans cette conclusion. Si Olivier Orain avait développé ces limites, l'expression employée de « stricte homologie » aurait été difficilement soutenable. De plus, l'idée avancée par Olivier Orain de « reprendre les constructions kuhniennes dans une autre perspective que celle des sciences expérimentales, dans un projet de théorie nominaliste de la connaissance »

---

[47] Cette expression a été employée à l'origine en 1971 par Gilbert Ryle pour la distinguer des descriptions minces *(thick description*) qui en restent aux faits de surface alors que les descriptions épaisses prennent en compte le contexte (Ryle 1971). Elle a été reprise par C Geertz pour situer le travail de l'anthropologue dans une optique très interprétative (Geertz 2008). En utilisant cette référence, Olivier Orain se place entre dans une lignée intellectuelle qui joue entre structuralisme et herméneutique sans vraiment clarifier sa position. Ce jeu a été décrypté par Isabelle Lefort (2003) sur *Le Déchiffrement du Monde* de Roger Brunet (2017).

[48] Par exemple, à la page 120 : « Malgré tout, l'amorce du transfert que nous venons d'opérer met à jour une difficulté récurrente que nous aurons à régler : le modèle kuhnien a été construit pour penser des sciences fort éloignées de la géographie (et des sciences humaines en général). Tant par leurs fonctionnements matériels (vie de laboratoire, etc.) que par leurs valeurs, les sciences dites "dures " offrent une base empirique " radicalement différente de celle que nous fournit la géographie classique (sauf en géomorphologie). On ne peut donc transférer de manière simpliste le modèle "paradigmes" (avec toutes ses implications), mais l'adapter, en ajustant un certain nombre d'attendus et en reconsidérant la liste de ses éléments » (Orain 2003, 120). Il faut souligner que la proposition relative « que nous aurons à régler » a disparu dans la version réécrite pour l'édition sous forme de livre (Orain 2009, 113).

est finalement peu développée laissant un certain flou. Il n'est pas question de développer dans cette introduction des hypothèses explicatives sur ce sujet complexe, mais cette revendication d'un usage du concept de paradigme au-delà d'une simple métaphore ou même d'une analogie[49] doit être soulignée. En effet, c'est une partie importante du problème qui anime cette thèse : le recours au modèle kuhnien pour lire la géographie française est-il seulement une image pour ouvrir des pistes de réflexion ou le rapprochement va-t-il au-delà d'une simple image ? Est-il utile de rester dans une position médiane, mal définie et floue ?

Enfin, un autre positionnement d'Olivier Orain mérite ici d'être explicité : contrairement à la plupart des lectures précédemment qualifiées de « raffinées » (qui se sont détachées de la perspective initiale très problématique consistant à utiliser le concept de paradigme pour valider la géographie théorique et quantitative comme « révolution scientifique »), Olivier Orain renoue avec cet objectif originel. Même si cet auteur reconnaît que le nouveau paradigme est incomplet, il n'en reste pas moins que sa lecture valorise très fortement l'importance intellectuelle qu'a eue la géographie théorique et quantitative.

Après ces éléments de contextualisation et de cadrage, il est temps de donner un aperçu détaillé des propositions intellectuelles contenues dans sa thèse. Ce lien[50] est une invitation à la consulter en ligne, car il est vrai qu'un résumé est toujours partiel et partial. Cet exercice est tout de même effectué et présenté dans la section suivante avec un effort pour essayer de synthétiser et rendre au plus juste les principaux éléments de cette thèse.

b. Économie générale de la thèse d'Olivier Orain : du paradigme réaliste aux approches constructivistes

Dans son introduction, Olivier Orain précise qu'il a construit son analyse par rapport à une dichotomie *réalisme* versus *constructivisme* qui « semblait particulièrement parlante et féconde, à condition de ne pas tout rabattre dessus et de ne pas l'absolutiser » (*Ibid* 2003, 19). Cette présentation reprend donc cette approche dichotomique en essayant de rendre compte des nuances apportées par cet auteur.

Concernant le « réalisme », ce paradigme commence pour Olivier Orain après la retraite de Paul Vidal de la Blache (1909). Pour justifier scientifiquement cette borne chronologique au-delà de cet évènement marquant[51], « l'idée d'un effort de rationalisation ou de

[49] L'analogie suppose une égalité des rapports dans les termes comparés alors que la métaphore permet des transferts beaucoup plus souple et imparfait.

[50] https://theses.hal.science/tel-00082408/file/Orain_these.pdf consulté le 18/09/2023.

[51] Paul Vidal de la Blache est le fondateur de ce qui a été appelé l'école française de géographie.

systématisation » (Orain 2003, 37) est avancée. Une « codification » et une réduction effectuée par les post-vidaliens[52] sont également mises en avant par rapport à une géographie vidalienne beaucoup plus littéraire et plurielle dans ses voies de recherche. Pour appuyer cet argument, une référence aux travaux de Marie-Claire Robic est utilisée, accréditant l'idée qu'à partir de la géographie vidalienne d'autres géographies auraient été possibles, notamment une guidée par « une problématique de la "position relative des lieux" » (Orain 2003, 36). C'est donc, par rapport au modèle kuhnien, qu'après l'élimination de ces propositions concurrentes que cette période, correspondant à la phase de pré-science[53], se termine.

Un autre registre d'explication est également avancé par Olivier Orain : selon lui, Paul Vidal de la Blache peut être rapproché des « écrivains réalistes ou naturalistes du second XIX^{ème} siècle, c'est-à-dire avec une conscience de l'effectuation scripturaire, plutôt qu'en vertu d'une clause épistémologique qui viendrait peser sur la liberté du dire » (Orain 2003, 38). Cette formulation complexe sert à avancer l'idée d'une rupture dans l'écriture. D'un jeu littéraire assumé avec une certaine liberté pour rendre compte de la complexité du monde chez Paul Vidal de la Blache, la géographie française serait passée à une normalisation pour donner à voir le réel à tout prix en oubliant ce qui relevait de la richesse et de la subtilité littéraire. Pour appuyer le propos, Olivier Orain mobilise les travaux linguistiques de Roman Jakobson[54]. Dans ce cadre, la géographie vidalienne faisait beaucoup plus de place aux fonctions expressive, conative et poétique du langage par rapport à ce qui a été développé par ses successeurs.

Les éléments qui caractérisent le paradigme classique sont un inductivisme fort (valorisation de ce qui vient avant tout du terrain), une recherche d'exhaustivité, une tendance au « plan à tiroirs » (c'est-à-dire à réaliser une partie pour chaque facteur d'explication : relief, climat…), une écriture avant tout descriptive et transparente ainsi qu'une posture réflexive réduite ne remettant pas en cause le caractère donné des faits. L'expression de « plain-pied » est utilisée pour résumer ce positionnement. Plusieurs auteurs faisant figure d'exceptions précoces comme Camille Vallaux (1870-1945) ou encore Jean Gottmann (1915-1994) sont détaillés.

Ensuite, pour Olivier Orain, la montée d'un malaise ne s'affirme dans la géographie française qu'à partir des années 1960 et l'« anomalie », au sens kuhnien du terme, n'a lieu

---

[52] Par cette expression, nous entendons les géographes venant après Paul Vidal de la Blache.

[53] Olivier Orain, habilement, n'utilise pas cette expression polémique de « pré-science ».

[54] Célèbre linguiste du XX^{ème} siècle (1896-1982) qui a posé les premières pierres du structuralisme. Il distingue six fonctions du langage : référentielle (représentative de l'état des choses), expressive (le sujet exprime quelque chose de personnel), conative (pour faire agir l'interlocuteur), phatique (qui maintient le contact entre le locuteur et l'interlocuteur), métalinguistique (qui fait référence au langage lui-même) et poétique.

qu'avec la rencontre avec la scène aménagiste. Le fort débat ayant eu lieu entre la géographie appliquée (incarnée par Michel Philipponneau) et la géographie active (portée par Pierre George) pendant la décennie 1960 est rappelé dans cette optique. Il y aurait eu alors, selon la thèse du plain-pied[55], une forte résistance du paradigme classique détaillée à travers des auteurs comme Pierre George ou Jean Labasse.

Pour Olivier Orain, la phase révolutionnaire ne commence qu'à partir du début des années 1970 : une diversité de courants théoriques notamment portés par plusieurs revues (*L'Espace Géographique*, *Hérodote* et *Espace-Temps)* s'affronte. L'expression historiquement utilisée de « crise de la géographie » est réutilisée dans un sens kuhnien. Des postures constructivistes s'affirment via la demande généralisée d'avoir des recherches explicitant leur(s) problématique(s) ou à travers des évènements plus spécifiques comme la rencontre *Géopoint* de 1978.

Olivier Orain reconnaît que le nouveau paradigme ne s'est pas complètement imposé. Ce dernier « n'est partagé que par une fraction restreinte de la communauté géographique française, n'a valeur 'normale' que dans certaines universités parisiennes et de l'Est et du Sud-Est de la France (sur le territoire des *Dupont*), et encore est-ce à concurrence d'autres discours » (Orain 2003, 291). À cette limite, s'en ajoute une deuxième plus globalement liée à la posture constructiviste, qui s'est certes largement diffusée, mais n'a pas été complètement poussée à son terme. Des analyses d'auteurs comme Jean-Bernard Racine, Claude Raffestin ou Frank Auriac sont réalisées de manière détaillée pour illustrer et soutenir le propos. Le milieu des années 1980 est présenté comme une rupture marquant la fin de la validité de la lecture kuhnienne du fait de l'apaisement des confits des années 1970 et d'une pluralité scientifique rendant l'usage du terme paradigme « hétérodoxe, voire inapproprié » 441.

Au terme de ce parcours, le modèle kuhnien est validé par la thèse du plain-pied pour la géographie française de 1910 à 1985. Olivier Orain rappelle également dans sa conclusion une piste de recherche que sa thèse lui a permis de construire. Cette piste est plus particulièrement développée dans la section suivante, car elle a joué un rôle majeur dans la constitution de mon travail de doctorat.

[55] J'utilise dans ce manuscrit l'expression « thèse du plain-pied » comme un synonyme de la thèse d'Olivier Orain (2003). Même si un livre a été ensuite tiré de cette thèse, mes analyses se sont concentrées sur la thèse originelle. Les quelques exceptions existantes sont signalées par la référence (Orain, 2009) qui renvoie alors au livre.

c. La « conjecture socio-linguistique »

S'appuyant sur l'idée d'un tournant linguistique chez Thomas Kuhn, Olivier Orain avance ce qu'il appelle une « conjecture socio-linguistique ». Cette dernière peut se résumer ainsi : « on peut supposer qu'un groupe scientifique "change de monde" précisément quand sa sémantique mute » (Orain 2003, 354). L'objectif est par conséquent de repérer cette bascule sémantique pour mieux identifier le changement de paradigme. Olivier Orain met en place une démarche quantitative - qu'il reconnaît être très incomplète - mais qui lui permet tout de même d'expérimenter sa proposition et d'en tirer plusieurs conclusions.

Concernant la phase pré-révolutionnaire (années 1960), Olivier Orain affirme qu'il y a une importation de nouveaux termes (« espace », « modèle », « système »…) dans la géographie française, mais sans véritable changement sémantique. À partir du début de la décennie 1970, les mots-emblèmes de la discipline étaient, selon lui, définis par « un statut notionnel tellement vague que leur usage avait d'abord une fonction d'invocation identitaire (ainsi "espace" connotait un modernisme revendiqué et référait principalement à l'idée d'objet propre – partant légitime – de la géographie) »[56]. Olivier Orain affirme qu'il y a eu alors une équivalence sémantique des termes « espace », « milieu » et « paysage ».

Pendant la période révolutionnaire, le changement sémantique s'affirme : « les praticiens disposaient d'une constellation de mots neufs, susceptible de voir leur sens s'indurer et se préciser dans un contexte paradigmatique nouveau » (*Ibid* 2003, 353). Les exemples donnés sont « espace » et « modèle » dont le sens a changé et s'est précisé dans le cadre du paradigme spatialiste. Une partie du lexique de l'ancien paradigme a été abandonnée : « ici par exemple : "milieu", "agraire", "description" et bien d'autres termes… » (*Ibid* 2003, 354).

Enfin, ce n'est que vers le milieu des années 1980 que finit la phase révolutionnaire dans l'optique d'Olivier Orain. Cette période « correspond, dans un cadre paradigmatique à nouveau stable, à la généralisation d'une sémantique particulière, sur les bases du répertoire adopté depuis déjà quelque temps » (*Ibid* 2003, 354). De grands chantiers, comme une *Géographie Universelle* (Brunet *et al.* 1995) ou un nouveau dictionnaire de référence[57], sont lancés.

Dans la conclusion de sa thèse, Olivier Orain rappelle le caractère inabouti de sa proposition visant l'identification de la bascule sémantique (proposition que ma thèse a prise au sérieux) du fait de l'important travail de numérisation qu'elle implique en amont. Il détaille aussi les problèmes liés à l'idée de l'extension d'une telle étude à la géographie

[56] *Cf.* Bibliographie (Orain 2003, 353).
[57] Ce dictionnaire est intitulé *Les mots de la géographie* (Brunet, Ferras et Théry 1992).

contemporaine en rappelant la multiplicité des courants et les diverses réappropriations sémantiques opérées sur des termes comme « espace » ou « territoire ».

4. Continuation d'une démonstration inaboutie

Sans rentrer ici dans le détail de la réception de la thèse d'Olivier Orain, cette dernière n'a pas permis de mettre fin aux controverses autour de la pertinence ou non d'utiliser le prisme kuhnien pour lire l'histoire de la géographie française. La sortie de son ouvrage en 2009, version remaniée sur la forme, mais non sur le fond de sa thèse, s'est accompagnée de la publication de deux comptes-rendus critiques[58] dans la revue *Géocarrefour.* Le premier de Caroline Leininger-Frézal et le second d'Isabelle Lefort. Dans ce dernier, l'hypothèse de la conjecture sémantique formulée initialement par Olivier Orain est reprise dans une perspective particulièrement critique :

> « O. Orain dessine en fait une phase de malaise et de sortie de "crise" bien davantage marquée par des changements lexicaux que sémantiques – ce point-là est bien démontré – tendant donc à invalider l'existence même d'une "révolution scientifique" *stricto sensu* » (Lefort 2011, 236).

Il y a dans le même numéro la réponse directe d'Olivier Orain à cette remarque avec l'affirmation suivante :

> « Si j'ai conclu mon livre (après ma thèse) sur l'idée qu'une mutation du vocabulaire et de sa signification est l'une des formes tangibles de ce que Thomas Kuhn appelle un "renversement gestalltique", j'admets que la démonstration pour la géographie en est inachevée » (Orain 2011, 240).

Cette reconnaissance d'une démonstration inaboutie, déjà esquissée par Olivier Orain dans sa conclusion de thèse, prend toutefois à travers cette discussion un aspect beaucoup plus problématique. En effet, elle se retrouve au centre de la controverse et a été choisie pour mon projet de thèse dans une mise en œuvre quantitative souhaitée par Olivier Orain lui-même. Du fait de l'implication d'Isabelle Lefort dans la controverse initiale, il me semble utile d'apporter quelques précisions sur le rôle que la co-directrice a joué dans cette réorientation de la recherche. En effet, le lecteur pourrait croire qu'il y a eu une volonté de sa part de mettre un étudiant sur le sujet afin de l'éclairer ou même de trouver des arguments allant dans son sens. Cette réorientation est de mon ressort et correspond à mon souhait de trouver une réponse aux difficultés précédemment développées. Ce sujet a été bien entendu discuté et validé ensemble.

[58] Dans la rubrique : *Lectures et relectures croisées* (Lefort 2011; Leininger-Frézal 2011; Orain 2011).

Du fait de cette orientation, il m'est important de préciser ce qui est entendu par le terme « critique ». Ce dernier vient du grec ancien *kritikos* : « capable de discernement et de jugement ». Il est également lié au verbe *krinein* (« séparer », « choisir », « décider », « passer au tamis »). Il me semble que ces références étymologiques rendent compte de la volonté qui a animé cette recherche. Il est évident que le compte-rendu d'Isabelle Lefort et sa proposition de débat ont joué un rôle important dans cette construction, mais il y a eu de sa part beaucoup d'espace, de temps et de liberté qui m'ont été laissés pour que j'élabore ma réflexion à partir de mes envies de recherche, de mes expérimentations et de mes investigations. Je tiens ici à souligner qu'il m'a fallu un temps particulièrement long pour trouver ma position par rapport à des propositions souvent complexes et subtiles d'Olivier Orain. En effet, s'il est facile d'être dans l'opposition et de critiquer, il est beaucoup plus complexe d'être dans un jugement nuancé qui au-delà du désaccord exprimé apporte une plus-value cognitive. L'objectif a été d'essayer de poursuivre cet idéal.

Sur le fond, le choix de ce nouveau projet de recherche provient avant tout du fait qu'il me permettait de sortir des difficultés précédemment rencontrées. Afin d'expliciter cela, la section suivante reprend les difficultés explicitées sur le précédent projet (*cf.* section Chap2.I) et les développe à l'aune de cette nouvelle orientation de recherche.

5. Les apports du changement effectué

a. Choix du corpus

Ce nouveau projet m'a permis de restreindre le corpus à deux revues en pouvant justifier le choix réalisé. En effet, le paradigme réaliste mis en avant par Olivier Orain, a été historiquement porté par la revue les *Annales de Géographie* (fondée en 1891 par Paul Vidal de la Blache). Sa contestation, par ce qu'Olivier Orain a ensuite appelé « un paradigme "spatialiste" ou "théorico-quantitativiste'" » (2011, 240), a été portée avant tout en France par la revue *L'Espace Géographique* (fondée en 1971 par Roger Brunet). Cette idée n'est pas seulement présente dans la thèse d'Olivier Orain. Elle se retrouve par exemple dans l'ouvrage d'épistémologie de la discipline dirigé par Pascal Clerc :

> « Deux revues majeures de la géographie française se sont développées ainsi, marquant deux ruptures dans l'histoire de la discipline : les *Annales de géographie* qui participent à la naissance de la géographie universitaire en France et, 80 ans plus tard, *L'Espace Géographique* avec le début de la géographie dite "nouvelle" » (Clerc *et al.* 2019, 52).

La partie contenant cette citation intitulée astucieusement « Une revue pour un paradigme ? » montre bien l'importance du prisme kuhnien mais aussi une certaine mise en interrogation qui est sous-entendue sans être développée par Pascal Clerc.

Sur cette question explicite du corpus, en reprenant la thèse d'Olivier Orain, un regret explicite de cet auteur est d'avoir « peu travaillé sur des textes de recherche empirique, mettant surtout l'accent sur la production réflexive ou théorique au sens large, alors que telle n'était pas l'intention initiale » (Orain 2003, 19). Par rapport à cette remarque, il faut souligner que mon travail répond à cette critique en faisant une large place aux textes de recherche empirique : ces derniers sont largement présents dans les deux revues constituant le corpus. Il n'est pas question d'affirmer une image plus complète ou plus précise de la géographie, car le travail d'Olivier Orain est le résultat de nombreuses lectures. Toutefois, la prise en compte de ces articles à forte dimension empirique apporte un éclairage complémentaire au travail de thèse d'Olivier Orain qui est totalement justifié du fait de la reconnaissance par l'auteur lui-même d'une partie des limites de son travail.

Plus globalement, une question se posant par rapport au travail de référence d'Olivier Orain est celle de la sélection des textes qu'il a retenue et mise en avant. Il y a évidemment eu un travail de sa part pour trouver et mentionner les passages les plus pertinents de son point de vue. Toutefois, cette sélection en amont pose le problème de la représentativité des idées avancées. Le corpus utilisé pour cette recherche en ne sélectionnant pas *a priori* les textes pour aller dans le sens d'une thèse, apporte une plus-value heuristique dans le cadre d'un réexamen de la thèse d'Olivier Orain. Plusieurs points légitiment par conséquent le choix de ce corpus permettant de lever les difficultés précédemment rencontrées sur ce plan.

b. Obtention et préparation des données

Concernant l'obtention et de la préparation des données, en limitant le corpus à deux revues, le travail est devenu beaucoup plus abordable. L'UAR *Persée* m'avait déjà fourni à cette époque les données concernant les *Annales de Géographie* pour la période 1892-2006 et celles concernant *L'Espace Géographique* pour la période 1972-2000. Comme la thèse d'Olivier Orain concerne l'intervalle de temps 1910-1985, les données s'inscrivant dans ce cadre de travail étaient déjà acquises. Toutefois, une volonté de ma part a toujours été de ne pas limiter cette thèse à une perspective seulement d'histoire de la géographie, mais de l'inscrire également dans des problématiques contemporaines. Dans cette optique, une demande a été faite à l'équipe du portail *Cairn* pour obtenir les données les plus récentes de ces deux revues. Les numéros étant payants sur ce portail avec une période mobile de trois ans, il a été possible d'obtenir les données jusqu'en 2014. Ce processus a pris du temps, car

il a réclamé l'accord des maisons d'édition de ces deux revues : *Belin* pour *L'Espace Géographique* et *Armand Colin* pour les *Annales de Géographie.*

Lors du processus de préparation des données, plusieurs discussions ont eu lieu avec l'équipe de l'UAR *Persée.* Les connaissances acquises lors de mon travail au sein de cette institution (*cf.* section Chap1.II) ont joué un rôle non négligeable pour préciser mes demandes. L'objectif global était d'obtenir un texte de qualité pour réaliser ensuite des analyses textuelles. Par rapport à cet objectif, des informations par exemple issues de l'OCR qui spécifient la position de chaque mot sur la page, ont été réintégrées dans un format de données très utilisé[59], mais qui habituellement ne les contient pas. Sans rentrer plus en avant dans les détails, il y a eu de la part de l'UAR *Persée* tout un travail de régénération de données adaptées à l'objectif poursuivi. Ce processus m'a permis de pousser cet exercice de préparation des données aussi loin que j'ai pu, ce qui explique également l'importance qu'a prise cette partie par la suite (*cf.* Chap4 et Chap5).

c. Méthodologies d'analyse

Par rapport au projet initial, les échanges et les points de controverses entre Isabelle Lefort et Olivier Orain, justifient l'utilisation d'outils d'analyse textuelle pour objectiver la situation. Logiquement, en prenant la réflexion d'Isabelle Lefort au pied de la lettre, des méthodologies dédiées à la détection du changement sémantique auraient dû être directement privilégiées. Toutefois, à la différence de ma directrice de thèse, je ne pense pas que le changement lexical ait été bien démontré[60] dans le travail d'Olivier Orain. Cette idée explique que ma recherche s'est d'abord tournée vers des outils dédiés à la détection des changements lexicaux. Mes premières investigations ont été influencées par l'article de France Guérin-Pace *et al.* (2012) sur une analyse lexicale des titres et mots-clés de 1972 à 2010 dans la revue *L'Espace Géographique*. La proximité de corpus et d'orientation méthodologique explique que ce travail ait constitué une référence. Ainsi, les méthodologies employées par ces auteurs, c'est-à-dire des analyses de spécificité et des analyses factorielles de correspondances (AFC), ont été privilégiées dans un premier temps.

Une partie de ce travail d'analyse lexicale des articles des deux revues a abouti à une présentation dans un colloque sur les discontinuités à Arras en 2018. Pour les actes de ce colloque, la rédaction d'un article a donné lieu à un long processus de recherche dû à de multiples questionnements sur la méthodologie et l'interprétation des résultats. En effet,

[59] Le XML-TEI : *cf.* section Chap4.I.2
[60] *Cf.* citation section Chap2.II.4

après beaucoup d'expérimentations et de réflexions, il m'a fallu reconnaître qu'il n'était pas pertinent de tirer des conclusions fortes en m'appuyant seulement sur les résultats quantitatifs obtenus. Ces derniers pouvaient être interprétés sous certains angles comme allant dans le sens de la thèse d'Olivier Orain, mais aussi réinterroger cette même thèse profondément sous d'autres angles. Ces différentes interprétations proviennent de la diversité des représentations obtenues avec des paramètres d'entrée variables (échelle de temps, seuils…), mais également à partir d'une même représentation. Ce qui rend difficile l'acte même d'interprétation est qu'il oblige à passer d'observations se jouant à un niveau lexical à des hypothèses relevant de continuités/discontinuités non plus lexicales, mais cognitives (Beligné, Loudcher et Lefort, 2023b).

Après cette étude lexicale, les travaux menés ne se sont pas directement orientés vers une détection du changement sémantique, mais ont pris comme objet d'étude le changement thématique. Pour comprendre cette direction, il faut revenir à une insatisfaction provenant des outils précédemment utilisés. En effet, que ce soit avec les analyses de spécificité ou les AFC, ce qui est toujours mis en valeur correspond à ce qui différencie une ou des partie(s) du corpus par rapport aux autres. Ces techniques privilégient par conséquent la mise en exergue des discontinuités, mais permettent moins bien d'appréhender les continuités.

Par exemple, dans beaucoup de représentations obtenues précédemment, la géographie post-vidalienne est uniquement représentée par des termes de géographie physique qui disparaissent par la suite. Or, s'il est vrai que d'autres formes de géographie ont pris lexicalement le dessus par la suite dans les deux revues étudiées, il est aussi connu que la géographie physique est loin d'avoir disparu. De même, pour la période antérieure, la géographie physique pouvait être lexicalement dominante, mais il existait d'autres formes de géographie que ces représentations tendent à éclipser. Un objectif a été par conséquent d'obtenir une ou des représentations qui permettent d'atténuer ces angles morts et de mieux rendre compte de certaines continuités.

Cette réflexion m'a conduit à m'intéresser à différentes méthodes de saisie de l'évolution de thématiques textuelles : méthode Reinert, *topic model*… Après un état des lieux sur ce sujet, j'ai trouvé pertinent d'adapter un outil *Diachronic'Explorer* pour détecter l'évolution des mondes lexicaux issus de la méthode Reinert. En effet, bien qu'une communauté importante en analyse textuelle utilise la méthode Reinert (notamment via le logiciel *IRaMuTeQ*[61]), il n'existait pas d'outils pratiques de visualisation permettant de répondre à ce désir de suivi dans le temps des thématiques issues de cette méthode. La création de cet

[61] http://www.iramuteq.org/ consulté le 18/09/2023.

outil a donné naissance à une publication (Beligné, Lefort et Loudcher 2020a) dans les actes des Journées internationales d'Analyse statistique des Données Textuelles (JADT).

Il y a eu dans cet article le souci de donner à voir une grande partie du processus de recherche sans toutefois développer l'interrogation épistémologique sous-jacente. Ce choix provient d'une volonté de ne pas écrire toujours sur le même sujet, de tester une forme un peu différente d'écriture et de toucher aussi un public qui n'est pas forcément intéressé par des questions spécifiques d'épistémologie de la géographie. Cette décision doit être également comprise dans le cadre de l'évolution de cette thèse (*cf.* section Chap2.III) qui a conduit non pas à abandonner cette voie du changement thématique, mais à la détacher du contenu final de cette thèse. Toutefois, tout ce travail réalisé sur ce sujet mérite d'être mentionné, car il n'a pas été sans conséquence dans ma manière d'aborder ce qui constitue le cœur de cette thèse, à savoir le changement sémantique.

En effet, parallèlement à cette piste du changement thématique, ce dernier axe, à savoir la détection du changement sémantique, a été abordé. Au niveau des méthodologies employées, il y a eu quelques expérimentations avec les cooccurences, puis un travail important avec les plongements de mots. Ces derniers constituant l'appareillage méthodologique retenu pour la partie quantitative de ce doctorat, ce choix est explicité et détaillé dans un état des lieux réalisé par la suite (*cf.* Chap3). La section suivante aborde enfin le dernier registre de problèmes précédemment mentionné, à savoir ceux qui relevaient des entretiens et de l'articulation quantitatif-qualitatif

d. Les entretiens et l'articulation quantitatif-qualitatif

Le fait d'avoir commencé avec des méthodologies assez faciles à mettre en place - analyses de spécificité et AFC - a eu pour effet l'obtention rapide de résultats. Dans l'optique de mon précédent projet, j'ai alors contacté Olivier Orain pour faire un entretien de commentaires de mes résultats, mais aussi de recherche d'explications dans le fonctionnement des revues. Olivier Orain étant de plus le directeur actuel de *L'Espace Géographique*, il était idéalement placé pour me répondre. Ce dernier a accepté, mais a fortement insisté sur le fait que l'enregistrement pourrait brider la conversation. Après une phase de réflexion avec Isabelle Lefort, nous avons choisi de ne pas enregistrer. L'échange mené a bien permis de commenter les résultats produits par rapport aux connaissances et aux réflexions d'Olivier Orain. La recherche de raisons issues du fonctionnement des revues a également eu lieu. Toutefois, elle a été menée de manière beaucoup plus restreinte et

hypothétique, car la plupart des analyses portaient sur des périodes pendant lesquelles Olivier Orain ne faisait pas partie de *L'Espace Géographique*.

Une autre difficulté, par rapport à ce qui avait été initialement prévu, a été celle de la restitution de ce moment de la recherche. Toujours influencé par l'optique de l'ancien projet, la question de réaliser des entretiens avec d'autres acteurs s'est posée. L'idée d'un résultat hétérogène avec des enregistrements que pour certaines personnes m'a poussé à abandonner cette piste. J'ai alors pu pleinement me concentrer sur le travail d'analyse textuelle.

Ainsi, cette nouvelle orientation de recherche a permis de sortir globalement de l'ensemble de difficultés générées par le précédent projet. Une autre piste évoquée avec mes deux directrices dès le début de la recherche, mais qui avait été mise de côté du fait de tout cet ensemble important de difficultés rencontrées, a pu alors être de nouveau envisagée. Il s'agit de la réalisation d'une application permettant de présenter le travail mené.

6. Réalisation d'une application Web : interface de reproductibilité et d'explorations

Cette idée de réalisation d'une application est due à deux constats personnels :

- Le premier est la difficulté de reprise d'un corpus. En effet, avant d'effectuer des analyses textuelles, de multiples actions sont souvent réalisées pour nettoyer et pour structurer des données. Dans le meilleur des cas, toutes ces actions ont été consignées dans un fichier à part. Réouvrir un tel fichier (quand il existe et qu'on peut y avoir accès) n'est pas une chose aisée. De nombreuses actions ont un sens précis au moment de leur effectuation qui est perdu si de nombreux commentaires d'explication n'ont pas été ajoutés. De plus, il est rarement possible de faire facilement machine arrière, c'est-à-dire de changer une modification précédemment réalisée. Cela demande de revenir à un état antérieur du corpus et de réappliquer toute la chaîne de traitements qui a été réalisée ultérieurement. En général, une telle action demande un travail si important que, dans la pratique, il faut que la modification soit identifiée comme apportant un gain vraiment majeur pour qu'elle soit effectivement réalisée. Ceci limite fortement tout un processus de reprises et d'explorations.
- Le second constat personnel porte sur les résultats. Lors d'un processus de recherche, de nombreuses expériences sont effectuées, générant de multiples représentations. Il y a ensuite pour une publication la sélection d'un nombre restreint de ces représentations de manière assez stratégique suivant ce qui veut être montré. Une fois l'article publié, le lecteur n'a accès dans la plupart des cas qu'à la sélection qui a été réalisée. S'il a un doute ou une envie de générer lui-même ses propres résultats en

changeant les paramètres, ceci implique généralement un effort considérable pour reproduire l'expérience.

Ces deux constats peuvent être inscrits dans un cadre plus général appelé « crise de la reproductibilité » (Pashler et Wagenmakers 2012). Si ce phénomène a été davantage mis en avant dans les sciences dures (Baker 2016), il touche également les SHS (Camerer et *al.* 2018) . Une analyse de Pierre-Carl Langlais rejoint le dernier constat qui vient d'être effectué :

> « c'est que le débat sur la reproductibilité a fait peut-être émerger un problème plus profond : au-delà des enjeux de libre accès, la publication scientifique classique n'a jamais été adaptée à l'utilisation intensive de données quantitatives. La forme "article" ne se prête qu'à l'inclusion de quelques indicateurs imparfaits. Elle dissimule par nature les étapes successives du traitement en n'intervenant qu'en fin de course, alors que tout est déjà joué » (Langlais 2020, 8).

Il conclut que cette crise est une chance pour « faire émerger un nouvel écosystème de production et de circulation des données scientifiques » (*Ibid* 2020, 8).

De nombreuses propositions existent déjà comme les carnets de code (*notebook*) qui permettent de détailler les différentes étapes d'un traitement de données et de rendre compte du fil directeur d'une recherche. Deux études récentes (Adam Rule, Tabard et Hollan 2018; Pimentel *et al.* 2019) montrent cependant que la solution consistant à mettre les codes utilisés en accès libre sur *Github*[62] présente plusieurs limites. L'évolution des langages de programmation explique qu'un code déposé quelques années auparavant ne s'exécute souvent plus correctement. De plus, de mon point de vue, ces dépôts ou ces carnets de code demandent surtout des compétences avancées en informatique pour être explorés. L'enjeu est donc, dans ce contexte, de faire une proposition permettant de comprendre la fabrication des données et de donner la possibilité à l'utilisateur non informaticien d'explorer le processus de fabrication d'un travail et de générer de nouveaux résultats.

Ce projet d'application a engendré des problèmes techniques non négligeables pour être réalisé. Plus globalement, la section suivante rend compte de toutes les difficultés rencontrées par l'ensemble du projet central.

[62] https://github.com/ consulté le 18/09/2023

7. Difficultés rencontrées

Rétrospectivement, trois ensembles de difficultés ont jalonné la réalisation de ce projet central :

- Tout d'abord donc, lors de mon activité d'interprétation des AFC, il m'est rapidement apparu la nécessité de me positionner par rapport à l'important travail interprétatif qu'avait mené Olivier Orain (2003). Ses développements n'ont pas une forme linéaire et classique du type : raisons du choix du modèle / explicitation du modèle initial / ce qui est en adéquation / ce qui est en limite / ce qui est en inadéquation. Il y a de multiples détours et reprises qui, pour certains, sont un cheminement obligatoire d'une pensée complexe et, pour d'autres, peuvent être le moyen de contourner quelques problèmes, mais qui rendent, pour tous, les analyses difficiles. Il faut souvent naviguer entre plusieurs passages en prenant en compte des reformulations parfois fines, mais importantes. Ensuite, le travail d'Olivier Orain mobilise de nombreuses références. Comme Bruno Latour (2005) l'a montré, plus ce réseau de références utilisées est grand, plus un travail critique détaillé est complexe, car il faut aller lire et décrypter des sources multiples. Enfin, le modèle kuhnien utilisé en amont n'est pas exempt d'ambiguïtés. Il y a fallu donc travailler de manière conséquente tout cet héritage[63].
- Conjointement, au fur et à mesure de mes avancées, j'ai ressenti la nécessité d'organiser ma réflexion critique de manière globale et systémique. En effet, pris individuellement, plusieurs points de réflexion que j'avais commencé notamment à développer en interprétant les AFC obtenues, peuvent être considérés comme des détails, mais, rassemblés et organisés, ils forment un ensemble critique plus cohérent et plus fort. Cette réflexion m'a conduit à envisager une nouvelle partie entièrement basée sur des analyses qualitatives, et qui complèterait ma première partie quantitative.
- Parallèlement à ces évolutions, la troisième difficulté est arrivée au moment de la réalisation concrète de l'application. Il m'est apparu que j'avais largement sous-estimé le temps et l'énergie à mobiliser pour réaliser un tel outil. Intégrer tout ce que j'avais réalisé sur la préparation des données, mais aussi le changement lexical, thématique et sémantique était une charge trop lourde à mener en même temps que l'écriture de la thèse. Cette prise de conscience a donné lieu à une nouvelle réorientation de ce projet de thèse formant sa version finale.

[63] La longueur de cette partie en témoigne !

III. Le projet retenu : mises à l'épreuve de la lecture d'Olivier Orain et mises en perspective contemporaines

Pour faire face à la difficulté précédemment mentionnée d'un projet trop ambitieux pour être terminé dans un temps raisonnable, il a fallu faire des choix stratégiques. Une première décision a été de se concentrer sur une prise en charge quantitative seulement du changement sémantique.

1. Focus sur la sémantique et changement de focales

Ce choix d'abandonner le travail effectué sur le changement lexical et thématique peut paraître étrange. En effet, il est possible de penser légitimement que tout ce travail aurait pu être valorisé au-delà des publications réalisées et être réutilisé pour faire partie du corps de ce doctorat. Pourquoi avoir développé ces pistes de recherche et ne pas les mettre en valeur dans le rendu final de ce doctorat ?

Au-delà de la justification précédemment mentionnée – le fait que le changement lexical n'a pas été vraiment démontré par Olivier Orain – une explication plus informelle peut être également avancée pour faire comprendre ce long détour effectué par le changement lexical et thématique. S'attaquer à la problématique du changement sémantique m'a longtemps un peu effrayé. Les méthodologies à mobiliser n'étaient pas simples à mettre en œuvre. L'objet même recherché, à savoir des changements de sens des mots, me paraissait complexe avec un risque que la tâche soit trop difficile et ne parvienne pas à aboutir de manière convaincante. Commencer par le changement lexical et thématique a été aussi une stratégie pour débuter avec des méthodologies qui me paraissaient plus à la portée de mon niveau et qui avaient plus de chance d'aboutir. Il y a eu un besoin de ma part de monter en complexité progressivement.

Il aurait été certes toujours possible de garder l'ensemble de ce travail préalable dans le rendu final de la thèse. Cependant, au terme du chemin effectué, il ne m'a pas semblé que la multiplication des résultats quantitatifs sur les plans lexicaux, thématiques et sémantiques apportait une réelle plus-value heuristique. Il était plus convaincant de prendre en charge uniquement ce qui était le nœud du problème, à savoir le changement sémantique. Il faut ici préciser que ce n'est pas seulement la remarque d'Isabelle Lefort qui met cette dimension au centre de la problématique. Dans les travaux de Thomas Kuhn, c'est bien l'idée d'un changement sémantique qui explique l'incommensurabilité entre deux paradigmes :

> « Étant donné que les mots sur lesquels se cristallisent les difficultés ont été appris en partie par l'application directe à des exemples, les interlocuteurs qui ne se

> comprennent plus ne peuvent pas dire "j'utilise le mot élément (ou mélange, ou planète ou mouvement libre) d'une manière qui est déterminée par les critères suivants" » (Kuhn 2008, 273).

Un changement lexical n'est pas évoqué par Thomas Kuhn. Plusieurs indices laissent même penser que pour lui un changement de paradigme peut avoir lieu sans changement lexical. Thomas Kuhn précise par exemple :

> « Comme la plupart des objets, même dans des ensembles modifiés, continuent à être groupés ensemble, les noms des groupes sont généralement conservés. Néanmoins, le déplacement d'un ensemble secondaire entraîne ordinairement un changement critique dans le réseau de rapports qui les relie » (*Ibid* 2008, 272).

Ce qui en jeu ici n'est pas un changement lexical puisque « les noms des groupes sont généralement conservés », mais s'apparente plus à un changement sémantique du fait de la création d'un nouveau « réseau de rapports » entre les termes. Le changement lexical est donc une hypothèse spécifique à Olivier Orain alors que le changement sémantique peut être inscrit dans une perspective plus large dérivant de la pensée originelle de Thomas Kuhn. Il y a donc une légitimité plus forte à considérer le changement sémantique comme l'élément problématique central par rapport au projet de démonstration du changement de paradigme.

En allant plus loin, il est possible de penser que le changement sémantique est un élément de structure du changement de paradigme. Sur ce sujet, il faut mentionner qu'avant même l'élaboration de sa conjecture socio-sémantique, Olivier Orain revendique dans son introduction l'utilisation d'un outillage issu de la critique littéraire structuraliste, appelé « poétique ». Même si cette méthodologie n'est pas utilisée avec systématisme[64], ce qui est avancé est que ces outils ont permis de « "fracturer", ce qui demeurait latent, sous-jacent, plus généralement "en creux" dans la littérature disciplinaire » (Orain 2003, 14) . L'argument repose sur une image explicite, géologique, amenant le lecteur à penser que ce procédé permet de mettre à jour des structures sous-jacentes à l'ensemble des textes. Il faut ici souligner qu'il existe une cohérence indéniable dans la volonté d'Olivier Orain de vouloir (dé)montrer l'existence d'un changement de structures[65] pour affirmer une révolution structuraliste.

La conjecture socio-sémantique permet de renforcer cette perspective. Ce qui est visé est la possibilité d'une démonstration étayée par des chiffres. Même si Olivier Orain reconnaît les limites de sa partie quantitative, il y a dans cet exercice l'entretien d'un imaginaire faisant valoir une possible preuve allant au-delà des figures de styles ou d'interprétations historiques toujours discutables. L'emploi des termes « démonstration » et « conjecture », très reliés au

[64] Il n'y a pas par exemple d'essai de repérage de toutes les figures de style dans les textes.
[65] Structures littéraires sur la forme, mais cognitives sur le fond.

domaine mathématique, sert incontestablement cet imaginaire. L'aspect cohérent de la démarche ne peut être que souligné puisqu'il s'agit d'apporter une analyse quantifiée du changement de paradigme dans une perspective qui valorise une « nouvelle géographie » qui a été elle-même explicitement quantitative. Il faut ici souligner que cette stratégie est menée en arrière-plan. En effet, Olivier Orain et sa thèse sont très loin d'un profil et d'une revendication strictement quantitativiste. Il est évident que le premier plan est dominé par une recherche plus littéraire. Toutefois, il n'est pas anodin que toute une partie de la recherche d'Olivier Orain ait pris une forme plus quantifiée. Il y a là un essai pour renforcer, asseoir, et peut-être même essayer de conclure la démonstration.

Cette présentation permet de comprendre pourquoi ma thèse s'est concentrée dans un premier temps sur une prise en charge quantitative de la question du changement sémantique. De plus, ce changement de focale a permis de libérer une place et un temps non négligeables pour développer d'autres parties que j'ai jugées avec le temps plus stratégiques qu'une succession d'analyses quantitatives sur les plans lexicaux / thématiques / sémantiques. Parmi celles-ci, une mise à l'épreuve "davantage" qualitative de la lecture épistémologique d'Olivier Orain.

2. Mise à l'épreuve plus « qualitative » de la lecture d'Olivier Orain

Cette mise à l'épreuve est qualifiée ici de plus « qualitative », car elle ne mobilise, contrairement à la partie précédente, aucune méthode quantitative. Elle s'appuie sur des analyses que j'ai construites tout d'abord en explorant dans le détail plusieurs développements réalisés par Olivier Orain pour interpréter mes résultats quantitatifs. Un apport important a été ensuite réalisé par l'approfondissement des réflexions de Jean-Claude Passeron (2006, *1991*). Ces dernières m'ont ainsi permis de mieux comprendre l'insatisfaction éprouvée à l'égard de mes propres résultats quantitatifs, mais ont aussi permis un approfondissement de ma critique.

L'organisation retenue pour cette partie s'appuie sur trois axes que j'ai identifiés comme légitimant et structurant la thèse d'Olivier Orain (au-delà de la proposition de l'hypothèse socio-sémantique précédemment détaillée) :

- Le premier est l'adaptation des différentes étapes du schéma kuhnien. Le terme « étapes » renvoie ici à la succession : paradigme 1 – anomalie – crise – fin de la révolution – paradigme 2. Le fait que pour chaque étape, Olivier Orain détaille et analyse longuement de nombreux auteurs, donne une assise et une force à son propos.

- Le deuxième axe réside dans l'essai réalisé dans la thèse du plain-pied pour retrouver dans les géographies post-vidalienne les quatre éléments identifiés par Thomas Kuhn caractérisant une matrice disciplinaire : généralisations symboliques, métaphysique, valeurs et exemples communs. Ici aussi, le fait qu'Olivier Orain développe longuement chaque élément participe sans conteste à son effet de dé-monstration.
- Enfin, le troisième axe se consacre aux ancrages réflexifs proposés par la thèse du plain-pied au-delà de la géographie. Ce point est très visible dès l'introduction avec la mobilisation de travaux d'Hilary Putnam ou encore de Ian Hacking. Cet aspect est très cohérent avec le contenu de la thèse d'Olivier Orain puisque l'idée de plain-pied est directement en relation avec la problématique de la saisie du réel, thème qui est évidemment transversal à d'autres disciplines que la géographie. Cette mobilisation d'auteurs, extérieurs au champ initial, apporte un poids non négligeable aux propos d'Olivier Orain et mérite d'être analysée en détail.

3. Mises en perspectives contemporaines

Pour finir, lors de ma troisième année de thèse (2019), j'ai eu la chance d'être invité par Olivier Orain à un séminaire organisé par l'équipe *Épistémologie et Histoire de la GéOgraphie* (EHGO) pour présenter mes travaux. Ce que permet de comprendre la partie précédente, c'est qu'entre notre première rencontre (*cf.* section Chap2.II.5.d) et celle-ci, il y a eu une évolution de ma recherche vers une analyse plus outillée et documentée dans la critique de ses travaux. Après l'avoir averti en lui donnant le titre[66] de ma conférence, j'ai profité de cette occasion pour mettre à l'épreuve mes résultats d'alors. Cette présentation d'une heure et demie a été suivie de plus d'une heure de discussion relativement tendue. Les échanges oraux avec Olivier Orain ont été très limités dans la mesure où il a préféré donner avant tout la parole à la salle, ce qui est tout à fait compréhensible. Je le remercie conjointement d'avoir complété cet échange par une réponse écrite envoyée un peu plus tard.

Un reproche important formulé par Olivier Orain concerne le contenu de la présentation réalisée, trop centrée selon lui sur une critique de ses travaux. Sur ce point, je ne peux que lui donner raison. Ayant décidé de tester mon argumentation devant le public, sûrement le plus en désaccord épistémologique, j'ai été aussi loin que possible dans ma critique. Je regrette, avec le recul, de n'avoir pas alors su développer pourquoi ces développements étaient intéressants au-delà d'un débat épistémologique sur l'histoire de la géographie. Le

[66] Le titre en question était le suivant : « Évolutions sémantiques dans les *Annales de Géographie* et *L'Espace Géographique* et remise en cause d'une lecture kuhnienne de l'histoire de la géographie ».

travail conduit ultérieurement m'a permis de combler cette lacune avec des mises en perspectives contemporaines s'inscrivant dans les champs des humanités numériques et de l'analyse de données massives (*big data*). Pour comprendre ces choix de périmètres de réflexions, il faut préciser que l'idée d'une ou de plusieurs nouvelle(s) « révolution(s) » liée(s) aux humanités numériques et/ou aux données massives a été développée par de nombreux auteurs[67]. Il est alors légitime de se demander si la relecture critique effectuée pour la géographie française des années 1960-1970 (qui s'appuyait déjà sur des revendications de méthodologies quantitatives et des discours révolutionnaires) peut être transférée pour essayer de mieux comprendre les situations présentes. Et inversement, est-ce que des réflexions contemporaines permettent de relire différemment le positionnement épistémologique d'Olivier Orain, 20 ans après sa production ?

Ce jeu de miroir permet la réflexion par le développement d'analyses épistémologiques entre plusieurs périodes distinctes : la géographie classique, la « nouvelle géographie » et la situation contemporaine. Les deux dernières périodes ont ce point commun d'être marquées par la présence de discours révolutionnaires. L'examen détaillé de l'argumentation raffinée d'Olivier Orain qui vise à affirmer l'existence d'une révolution kuhnienne dans la géographie française est dans ce contexte certes un point d'entrée restreint, mais stratégique pour des analyses contemporaines. En effet, il s'agit d'analyser un discours qui est allé aussi loin que possible pour adapter le modèle kuhnien de la « révolution scientifique » à une SHS. Ce modèle kuhnien étant lui-même une des recherches les plus poussées et les plus connues pour caractériser ce qu'est une « révolution scientifique », il constitue un archétype et la lecture d'Olivier Orain en est une des déclinaisons les plus travaillées pour les SHS.

L'intérêt de ce point d'entrée peut être développé au-delà de cet argument. Pour cela, une objection faite par Olivier Orain dans sa réponse écrite, suite à ma présentation, mérite d'être analysée. Son reproche porte sur une approche, selon lui, « dichotomiste à l'excès » entre sciences de la nature et SHS : « À vouloir établir une ligne de démarcation absolue entre deux types de sciences qui auraient chacune leur ontologie propre, on s'interdit d'analyser tout ce qui est pensable comme hybride, mixte, etc. »[68]. En effet, ma présentation effectuée pour ce séminaire revendiquait une nette ligne de démarcation en s'appuyant sur les réflexions de Jean-Claude Passeron (2006, *1991*). Or, pour appuyer la critique d'Olivier

[67] Un petit article sur le sujet qui a été beaucoup cité : *The End of Theory* (Anderson 2008) . Pour les sciences de la nature, voir par exemple : *The Fourth Paradigm* (Hey 2012). Pour les sciences sociales, voir par exemple *Digital Humanities : First, Second and Third Wave* (Berry 2011) même si lexicalement l'expression « computational turn » est utilisée par David Berry, le changement qui est décrit relève du registre révolutionnaire (Guichard 2014).

[68] Écrit par Olivier Orain en 2019, mais non publié.

Orain, il est possible de mettre en avant la dynamique des données massives pour revendiquer, *a contrario*, un rapprochement SHS / sciences expérimentales. Ce dernier peut même être présenté sous la forme d'un changement de paradigme comme le réalise Pierre Mounier :

> « Tout d'abord, en arrière-plan de la vague du big data dans la recherche, se profile un véritable changement de paradigme au sein des différentes disciplines scientifiques : la méthode hypothético-déductive caractéristique des sciences expérimentales serait remise en cause et remplacée par une observation des "patterns" que présente la masse de données. Le lent travail de construction d'un corpus de données et de son interprétation laisserait la place à une observation des tendances émergeant "naturellement" du corpus, selon le journaliste Chris Anderson » (Mounier 2018, 17).

Le mouvement de fond de la vague des données massives et de l'apprentissage profond se caractérise donc par le développement des démarches inductives alors qu'une partie de la géographie française des années 1970 a revendiqué sa légitimité scientifique en s'appuyant sur des approches hypothético-déductives. La formule employée par Pierre Mounier à la fin de la citation précédente – « son interprétation laisserait la place à une observation des tendances émergeant "naturellement" du corpus » – montre un recul critique de cet auteur avec l'emploi du conditionnel « laisserait » et la mise entre guillemets de « naturellement », mais sous-entend aussi l'existence d'une nouvelle posture de plain-pied dans les données massives. Il n'est pas question de confondre cette dernière avec la forme de « plain-pied » développée par Olivier Orain pour la géographie post-vidalienne, mais la mise en analogie des deux, avec cette croyance commune plus ou moins développée que l'on donne à voir le réel, mérite d'être réfléchie. Cet axe peut être approfondi, tout autant pour le champ des données massives (champ auquel n'appartient pas cette thèse, mais qui a un rôle moteur dans cette interrogation) que pour celui des humanités numériques (champ auquel appartient cette thèse et, bien évidemment, concerné par cette interrogation).

Par rapport au reproche d'Olivier Orain – une approche dichotomique SHS / sciences dures – je tiens à préciser qu'il n'y a pas eu, de ma part, de postulat de départ sur ce sujet. J'ai réalisé un chemin de recherche sur une problématique relevant des humanités en utilisant de manière importante l'outil informatique, dont une partie de la matrice intellectuelle, par la modélisation et la logique, est proche des sciences dures. Par conséquent, le champ des humanités numériques dans lequel je me situe est plutôt à l'opposé d'une approche dichotomique et a pour idéal une recherche hybride aussi bien au niveau des méthodologies que des cadres théoriques. Mon positionnement n'est donc pas lié à un postulat de départ, mais à mes réflexions quant à ce chemin de recherche effectué dans les humanités numériques.

La double mise en perspective contemporaine, humanités numériques et domaine des données massives, a engendré un important travail réflexif. Comme l'analyse quantitative de la partie contemporaine des deux revues n'était plus au centre de ce projet, les données obtenues à partir du portail *Cairn* ont finalement été exclues du corpus analysé. Cette décision n'a été prise que pour alléger la charge de travail en supprimant toute la problématique d'homogénéisation des données provenant de deux portails différents. Ceci a eu pour effet de réduire également le travail sur l'application qui présente cette préparation des données.

Par rapport au projet précédent, il y a ainsi eu une réarticulation assez différente des parties quantitatives et qualitatives. En effet, la partie qualitative avait été définie dans le projet précédent comme la phase d'interprétation des résultats quantitatifs obtenus. Ce processus devait avoir lieu sur trois axes : le changement lexical, thématique et sémantique. Dans ce nouveau projet, seule la partie sémantique est conservée, mais le travail qualitatif est fortement renforcé puisque les deux dernières parties de ce mémoire relèvent d'approches non computationnelles. Au-delà d'une conscience accrue des limites de ces méthodes quantitatives, il y a surtout eu une réorientation pour aborder ce qui a été jugé comme étant les problématiques les plus stratégiques à traiter au terme de tout ce parcours :

- En premier lieu, une prise en charge quantitative du changement sémantique sur la période étudiée par Olivier Orain.
- Ensuite, une relecture qualitative de son travail sur les autres points forts qui soutiennent sa lecture kuhnienne.
- Enfin une double mise en perspective de ce travail dans le champ des humanités numériques et de l'analyse des données massives (*big data*).

IV. Réflexions conclusives

Cette rentrée dans le processus de fabrication du sujet de cette thèse constitue sans nul doute un départ quelque peu long, dense et complexe. Toutefois, ces trois qualificatifs sont aussi parfaitement adaptés pour caractériser cette phase de défrichement effectuée au début de ce doctorat. Nul doute que certains considéreront que ce développement des projets intermédiaires n'était pas vraiment nécessaire, au sens où il aurait suffi de présenter le dernier projet. Peu importe les dessous de la construction, seul compterait le résultat final.

Mes autres missions effectuées sur divers projets de recherche avant ce doctorat m'ont amené à penser qu'une telle optique conduit à écrire souvent *a posteriori* une grande partie de l'état des lieux pour justifier logiquement les choix réalisés alors que ceux-ci relèvent en

réalité de raisons occultées. Par exemple, le stagiaire recruté sur tel projet ne parle qu'anglais, ce qui conduit à construire un corpus dans cette langue. Dans un autre, l'enseignant-chercheur encadrant un projet, maitrise plutôt telle méthodologie, ce qui explique que le projet s'oriente dans cette direction, mais ce choix est ensuite justifié en arguant d'une ou deux références appropriées. Il aurait été possible de faire une telle présentation pour ce doctorat. Le choix réalisé est ici tout autre et rend sûrement ce travail plus fragile aux yeux de certains.

Il est ici intéressant de souligner qu'Olivier Orain use d'un jeu de mots à la fin de sa thèse pour suggérer aux géographes une injonction forte : « soyez irréaliste, demandez le constructible ! » (Orain 2003, 297). Or, le constructible n'est pas seulement le résultat final. En reprenant la distinction de Bruno Latour entre une « science en train de se faire » : « vivante », « incertaine », « informelle » et « changeante » (Latour 2005, 29) par rapport à une « science toute faite » : « austère », « sûre d'elle-même », « formaliste », « réglée » (*Ibid* 2005, 29), il est possible de considérer les développements précédents comme une tentative pour ne pas occulter ce versant de la science en train de se faire. Nul doute que la distinction que fait Bruno Latour est très dichotomique pour les besoins de son exposé. Nul doute également que ce prologue n'expose pas vraiment au sens propre la science en train de faire. Il y a eu évidemment une reconstruction. Toutefois, tout l'effort réalisé dans cette partie pour donner à voir la construction de la recherche est loin d'être secondaire.

Cette proposition est d'autant plus forte, me semble-t-il, qu'elle touche une recherche qui a un versant quantitatif. La tentation de faire disparaître l'ensemble des éléments contingents et contextuels sous une forme de logique souvent reconstruite, et même parfois fictive, est à mon avis d'autant plus grande que les domaines sont fortement marqués par des approches quantitatives. La compétition existante entre scientifiques pousse à évacuer fortement toute considération humaine qui aurait pu jouer sur la recherche. Dans ce contexte, c'est sûrement une proposition assez marginale qui est ici effectuée, mais j'ai trouvé cet exercice particulièrement intéressant et stimulant à réaliser. Il me semble assez stratégique de l'avoir réalisé dans le cadre d'une thèse. En effet, c'est un temps privilégié pour mettre en œuvre un chemin de recherche conséquent et la taille du rendu permet de tels développements.

Il s'agit maintenant après ce prologue de traiter le projet scientifique ainsi défini. Dans un premier temps, les choix méthodologiques et la préparation des données sont détaillés.

Première partie :

Choix méthodologiques et préparation des données

Cette partie, ainsi que la suivante, s'appuie sur une application Web. Deux des objectifs de cette application − l'exploration et la reproductibilité de la recherche − ont été précédemment développés dans le prologue (*cf.* section Chap2.II.6). Après avoir écrit les premiers scripts *Python* pour préparer et traiter les données, un travail a été réalisé avec l'aide d'un ingénieur d'études, Pierre Prablanc, pour améliorer des parties du code. Cette étape a permis de clarifier et de mieux structurer plusieurs objets et fonctions. Ensuite, le choix d'une application Web s'est progressivement imposé[69] pour simplifier l'accès à tous en évitant les problèmes liés à une procédure d'installation. Le framework *Django*[70] a permis de facilement structurer les différentes parties du code. Une partie de ce travail a fait l'objet d'une présentation au séminaire *CoCoPySHS*[71] (Beligné 2023).

Au niveau de la navigation dans l'application, cette dernière a été testée sur un écran d'ordinateur avec comme navigateur *Firefox* ou *Chrome.* Les navigations sur tablette ou smartphone n'ont pas fait l'objet de test ni d'optimisation. Il est donc conseillé pour naviguer dans cette application d'utiliser un ordinateur avec les navigateurs sus-cités.

Plusieurs utilisations sont ensuite possibles :

- La première consiste, à partir de la lecture de ce manuscrit, à ouvrir les liens permettant d'accéder directement à des parties précises de l'application. Ces liens commencent tous par un **@** et sont écrits en gras, ce qui permet de facilement les identifier. Le lecteur peut ensuite revenir au manuscrit et continuer en ouvrant ainsi au fur et à mesure les liens qui l'intéressent.

- Une deuxième utilisation possible consiste, à partir de la fin de l'introduction de cette partie, à poursuivre la lecture directement sur l'application en suivant ce lien : https://analytics.huma-num.fr/EtPistezMots/

  La page d'accueil qui s'ouvre est alors la suivante :

[69] Au départ, le code était sous la forme d'une application hébergée dans une machine virtuelle, mais non accessible par le Web.
[70] https://www.djangoproject.com/ consulté le 18/09/2023.
[71] Je remercie ici vivement Emilien Schultz pour cette opportunité de présentation et la mise en ligne.

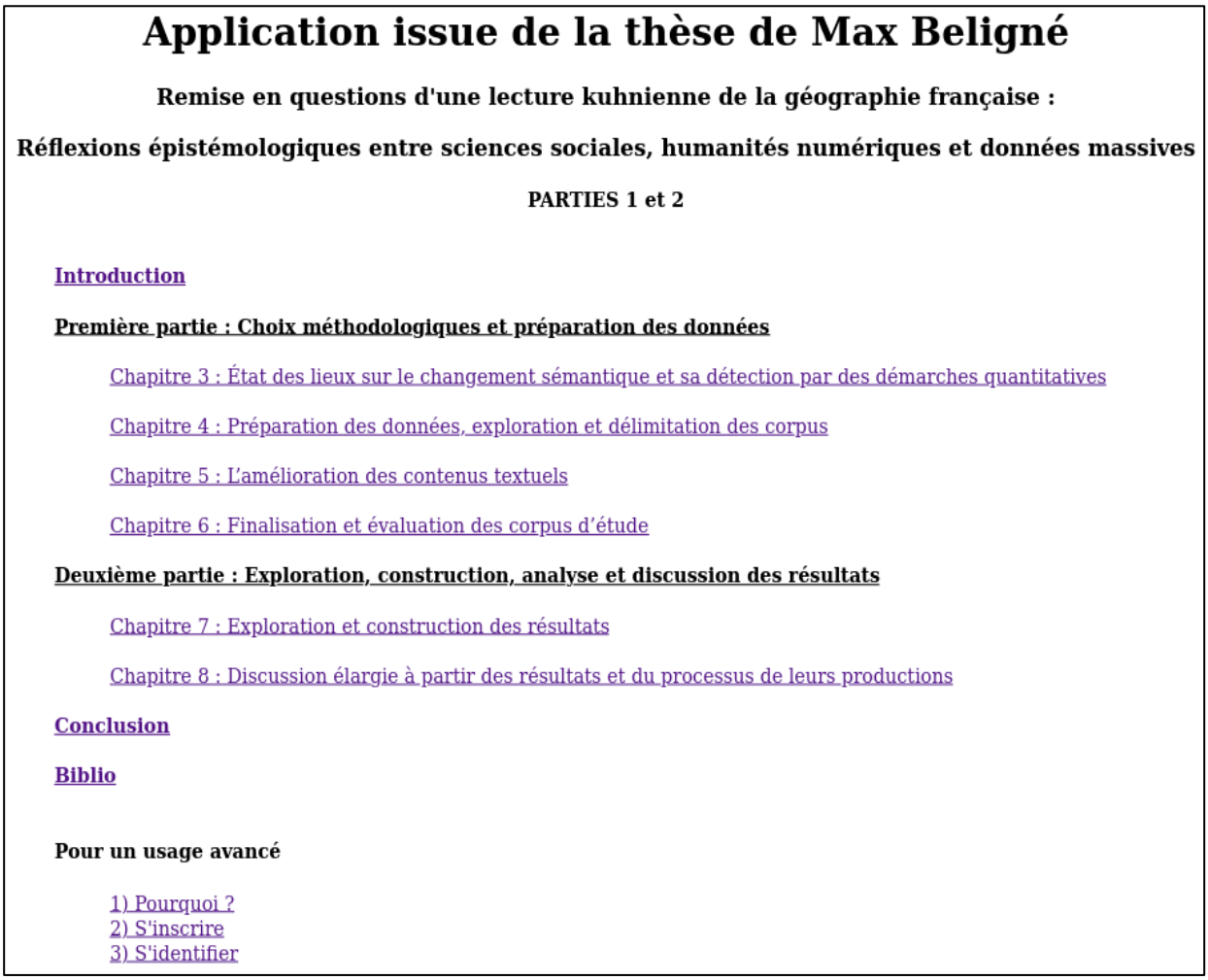

Figure n°1 : La page d'accueil de l'application Web.

La lecture dans l'application s'effectue en cliquant sur les liens ou en utilisant les raccourcis « Ctrl + S » (page suivante) et « Ctrl + R » (page précédente). La lecture directement sur ce support permet d'éviter de multiples allers-retours (notamment lors des étapes de préparation du corpus et de productions des résultats où les liens vers l'application sont plus nombreux).

Si vous avez lu précédemment le prologue de ce manuscrit, il est inutile de lire l'introduction de l'application qui résume à grand trait la problématique de la recherche. Dans le cas, donc, où vous avez lu le prologue de ce manuscrit et vous souhaitez passer sur l'application pour la lecture des parties 1 et 2, la démarche conseillée est d'effectuer ce basculement à la fin de cette introduction en commençant directement par le chapitre 3 (pour éviter des confusions, les chapitres ont été numérotés de la même manière dans le manuscrit et sur l'application Web).

En cas de lecture complète de la thèse, il est nécessaire de revenir au manuscrit après la conclusion de l'application. En effet, les dernières parties, 3 et 4, n'ont pas été intégrées à ce support, car cela aurait demandé un travail non négligeable pour une plus-value heuristique minime.

- Une troisième manière d'utiliser l'application passe par l'annexe « Pour un usage avancé », visible en bas de la Figure n°1, permettant de s'inscrire et de s'identifier. Cette inscription permet à des utilisateurs de produire quelques résultats en faisant des choix différents que ceux présentés dans cette thèse. Toutefois, certains traitements prennent un temps important, ce qui peut ralentir l'application[72]. De plus, en multipliant les possibilités laissées à l'utilisateur, la probabilité de problèmes techniques s'accroît aussi. Pour ces raisons, il m'a semblé préférable de garder le contrôle sur certains traitements. Par conséquent, même en étant inscrit, plusieurs traitements demandent l'envoi d'un courriel pour être réalisés. Le système d'enregistrement et d'identification me permet d'affecter des données et des résultats à un ou des utilisateur(s) spécifique(s). Cette organisation offre une stabilité de l'application de base tout en laissant la possibilité effective d'autres constructions.

- Une quatrième manière d'utiliser l'application est de la générer en local à partir du code disponible à l'adresse suivante : https://github.com/etpistezmots/AppliThese. Pour cela, il est nécessaire de maitriser des opérations telles que la création d'un environnement virtuel et le lancement d'un projet *Django*. Toutefois, la production de résultats par une personne extérieure reste limitée, car il manque les sources utilisées. À cet effet, j'ai effectué quelques démarches pour rendre les sources disponibles, mais cela nécessite l'accord des comités de direction des deux revues, de leurs maisons d'édition ainsi que de l'UAR *Persée*. Il s'agit donc de démarches assez lourdes qui n'ont pas eu le temps d'aboutir. Si la mise en accès du code ne permet pas de générer directement[73] de nouveaux résultats, elle demeure néanmoins à mon sens un acte significatif en termes de reproductibilité. En effet, toute personne intéressée peut rentrer dans tous les détails du fonctionnement de l'application. Il reste vrai que par rapport à l'idéal d'ouverture initial, quelques réductions et ajustements ont été nécessaires.

[72] Toute l'application est synchrone. Ce n'est qu'à la fin de ma thèse que j'ai découvert les possibilités de développement Web avec de l'asynchrone permettant d'éviter ce problème de surcharge par certains traitements.

[73] Indirectement, il est possible de le faire en demandant à l'UAR *Persée* un accès aux sources utilisées.

Ces différentes utilisations ne sont évidemment en rien exclusives l'une de l'autre et peuvent être combinées. Au niveau du contenu, le chapitre suivant vise tout d'abord à présenter un état des lieux sur la notion de changement sémantique et la problématique de sa détection quantitative. L'objectif est d'examiner les travaux existants afin de mieux définir les choix théoriques et méthodologiques effectués.

Chapitre 3 :
État des lieux sur le changement sémantique et sa détection par des démarches quantitatives

Il s'agit pour commencer de préciser ce qui est cherché sous cette appellation générique de « changement sémantique ». Sémantique, comme adjectif, est défini dans le dictionnaire du Trésor de la Langue Française (TLF) comme ce « qui a rapport à la signification d'un mot ou d'une structure linguistique ». D'un point de vue formel, un changement sémantique peut donc être défini comme un changement de signification concernant un ou plusieurs mot(s) ou structure(s) linguistique(s). Toutefois, cette définition ne fait que reporter la difficulté sur la question suivante : qu'est-ce que la signification d'un mot ou d'une structure linguistique ? Par rapport à cette question, deux grandes approches existent : d'un côté les sémantiques référentielles, de l'autre les sémantiques différentielles.

I. Approches théoriques du changement sémantique

1. Les sémantiques référentielles

François Rastier (1995) montre que l'origine des sémantiques référentielles remonte à Aristote avec une conception tripartite : parole / états de l'âme / choses. Il met également en évidence une évolution de ce modèle (Boèce, Thomas d'Aquin, Ogden et Richards) jusqu'à son durcissement « par le positivisme logique qui exprime un idéal de correspondance si l'on peut dire terme à terme entre un mot, un concept et un objet » (Rastier 1995). Une telle conception s'applique difficilement au langage naturel. En effet, il est courant qu'un mot renvoie à plusieurs concepts ou objets. Dans ce cas, la relation référentielle peut être pensée d'une manière plus souple que celle du positivisme logique. Schématiquement, deux approches peuvent être distinguées. La première renvoie à un système préexistant d'objets et d'états dans le monde physique alors que dans la seconde, les références sont établies par rapport à un monde mental. Dans la première, le changement sémantique désigne un changement de la chose désignée alors que dans la seconde, il renvoie plus à une modification des concepts permettant d'appréhender une chose. Il est certain que ces deux conceptions ne sont pas antinomiques et qu'il existe une multiplicité d'articulations possibles dans la conception tripartite d'Aristote précédemment citée : parole / états de l'âme / choses. L'objectif n'est pas ici de développer l'ensemble de ce groupe des sémantiques référentielles au-delà de ces grandes perspectives.

Ce groupe s'oppose à celui des sémantiques différentielles.

2. Les sémantiques différentielles

Ces dernières ne reposent sur aucun système d'objet ou d'état préconçu. Le sens d'un mot n'est pas défini de manière positive par son contenu, mais uniquement par ses relations avec les autres termes. Cette conception est présente par exemple chez Saussure où la langue est un système dont « tous les termes sont solidaires et où la valeur de l'un ne résulte que de la présence simultanée de l'autre » (Saussure 1995, 159). La représentation de l'ensemble de ce système pour une langue est particulièrement ardue, car il s'agit d'un système complexe et ouvert aux évolutions. Dans cette perspective, il existe une première sémantique différentielle qui peut être qualifiée de componentielle (Sabah 2000). Le sens des mots est divisé en composants[74], appelés sèmes qui sont des éléments primitifs de sens. Ainsi, le terme 'couteau' peut être représenté de manière simplifiée par deux sèmes /couvert/ et /pour couper/. Ces éléments primitifs peuvent s'agencer de multiples manières pour créer de nouvelles significations.

Par rapport à cette approche, il existe un autre type de sémantique différentielle qui prend en considération de manière plus effective l'impossibilité de décrire de manière complète et directe le système sémantique d'une langue en s'orientant avant tout vers l'analyse locale de ses manifestations. En effet, en étudiant les distributions d'occurrences des entités linguistiques sur plusieurs textes, il est possible de caractériser des différences existantes et ainsi, dans certains cas, de mettre en avant des emplois sémantiques différenciés. Cette optique est par exemple celle de la sémantique interprétative de François Rastier. Bien qu'elle utilise le même vocable de « sème » que la sémantique componentielle précédemment vue, la perspective est inverse comme le souligne Bénédicte Pincemin : « Ce sont les sèmes qui sont décrits par un ensemble d'occurrences (/couvert/ = {'fourchette', 'couteau', 'cuillère'})[75] ; et ce n'est qu'indirectement qu'un ensemble de sèmes peut être attribué à un mot » (Pincemin 1999b, 72). Cette attribution des sèmes à un mot n'a rien d'automatique. Elle dépend de l'interprétation d'un chercheur. Dans ce cas, il n'y a jamais sur un texte de clôture du processus interprétatif au sens où diverses lectures sont toujours possibles. Le sémantisme n'a plus alors de base établie.

[74] « Components » en anglais, d'où le terme d'analyse componentielle.

[75] Cette formulation synthétique veut dire que le sème /couvert/ est par exemple décrit dans certains textes par les mots : « fourchette », « couteau » et « cuillère ».

3. Le positionnement adopté pour cette recherche

Du fait de son objectif d'étude du changement sémantique à partir des textes et non d'un système d'objets ou d'états préétablis, cette thèse s'inscrit dans le champ des sémantiques différentielles. Les réflexions de François Rastier ont été particulièrement utiles pour comprendre *a posteriori* le positionnement de cette recherche, mais il n'y a pas eu de tentative en amont pour être au plus proche de cette théorie. La démarche suivie a été surtout inspirée initialement par les réflexions de Bénédicte Pincemin (1999b) qui se demande de manière générale comment trouver des sèmes à partir de l'outillage de l'analyse textuelle. Sa réponse est loin de spécifier une méthode unique puisqu'elle présente une pluralité d'outils (cooccurrence, modèle de l'espace vectoriel…) pour atteindre cet objectif. Ainsi, le choix de ce cadre théorique général n'a pas permis d'en déduire logiquement une méthode. En revanche, ce choix théorique a permis de stimuler plusieurs réflexions, notamment en considérant de possibles attractions vis-à-vis d'autres positionnements.

En effet, dans le cadre d'une démonstration d'un changement sémantique, il est évident qu'une approche des sèmes ouverte et relative comme celle proposée par François Rastier n'est pas l'idéal. Une approche componentielle serait indiscutablement plus démonstrative. Par exemple, par rapport à l'hypothèse d'Olivier Orain sur l'équivalence sémantique des termes « espace » et « milieu » puis leur différenciation, l'idéal serait de montrer qu'ils ont été constitués par les mêmes composants sémantiques de base puis que cette constitution a ensuite changé. *A contrario*, dans le cas d'une lecture des sèmes qui repose sur des différences locales d'utilisation, mais qui dépend *in fine* de l'interprétation d'un chercheur, l'objectivité dans la démonstration est moindre. Il y a donc une tension à prévoir dans cette recherche par rapport à une approche plus componentielle qui fournirait plus de poids à la démonstration.

De plus, il est ici utile de rappeler que l'objectif derrière la détermination d'un changement sémantique est bien de caractériser un changement cognitif dans la manière de faire de la géographie. Le changement sémantique est ici utilisé comme un proxy[76] : ce qui est visé est au fond la mise en avant d'un changement majeur dans les structures conceptuelles des géographes. Cette situation amène à penser qu'au cours de la recherche une tension va aussi probablement se manifester par rapport à une sémantique référentielle poussant à interpréter des changements sémantiques comme des modifications plus ou moins importantes de structures de pensée.

[76] Au sens d'une variable qui en remplace une autre moins facilement mesurable.

La section suivante complète cette approche très globale par une perspective plus spécifique à cette recherche. En effet, la problématique de cette recherche, au-delà de la détection d'un changement sémantique, est celle d'un changement de paradigme. Or, il n'y a rien d'évident dans ce passage qui vise à déduire du changement sémantique un changement de paradigme. Cette réflexion m'a conduit à m'intéresser à un point théorique précis dont la formalisation imagée m'a permis de mieux comprendre la recherche en cours.

II. La métaphore de la plasticité de la matière

Le travail de thèse de Caroline Reutenauer (2012) apporte une première réponse à cette interrogation sur le lien changement sémantique / changement de paradigme par le biais d'une analogie. L'évolution sémantique y est comparée à la plasticité de la matière. Ce qui permet la déformation, ce sont les « potentiels de sens, c'est-à-dire des facettes sémantiques susceptibles de s'actualiser » (Reutenauer 2012, 40) pour chaque mot. Cette analogie permet d'introduire les concepts de transformation réversible et irréversible. Dans le premier cas, il y a « une déformation temporaire du sens, qui reste tributaire des contraintes de l'environnement, c'est-à-dire du contexte, mais qui n'affecte pas le sens lexical lorsque les contraintes contextuelles se relâchent » (*Ibid* 2012, 40). Dans le second cas, à l'inverse, il y a une déformation durable et une restructuration du sens littéral qui peut « acquérir un nouveau contenu d'un point de vue cognitif » (*Ibid* 2012, 40). Ce que permet de penser cette analogie, c'est qu'au-delà d'une élasticité de la langue avec un processus de création de sens polysémiques, il peut exister des « seuils d'élasticité » impliquant des possibilités ou impossibilités de retour à l'état initial en fonction du franchissement ou non de ces derniers. L'idée d'un changement irréversible tel qu'il existe dans la théorie kuhnienne peut être appréhendée à travers cette image.

La détection du changement sémantique ici menée s'oriente par conséquent dans la mise en évidence d'un changement irréversible, sans retour possible à l'état initial. Il faut toutefois souligner que ce qui est recherché va au-delà de l'irréversibilité, car il faut qu'il y ait en plus, selon la théorie kuhnienne, « incommensurabilité ». Il est particulièrement difficile de fonder une étude quantitative sur un concept aussi flou, car il ne relève pas d'un critère directement mesurable. Dans une définition stricte d'ailleurs, le terme « incommensurable » renvoie à « des grandeurs qui n'ont pas de communes mesures » (Dictionnaire *Le Robert* 2002). Il y a incontestablement un certain paradoxe à vouloir mesurer une incommensurabilité. Par rapport à cette difficulté, un point intéressant dans la reprise du travail d'Olivier Orain, est qu'il permet de s'appuyer sur une prise assez concrète pour étudier là où se joue, selon lui, l'incommensurabilité d'un point de vue sémantique : sa conclusion affirme des dynamiques

qui partent d'une équivalence sémantique des termes « espace » et « milieu » avec au cours de la « révolution » une différenciation conceptuelle aboutissant à « un cadre paradigmatique à nouveau stable » (Orain 2003, 354) et à la « généralisation d'une sémantique particulière » (*Ibid* 2003, 354) au milieu des années 1980.

Cette affirmation s'appuie en partie sur une étude quantitative. Cette dernière pose question quand elle est examinée en détail tant au niveau du corpus (14 livres dont un seul après 1972) que de la méthode (des calculs d'occurrences et de rapports sans vraiment d'interrogation sur l'origine et la portée de cet outillage). Il ressort par conséquent d'un examen plus attentif que ce qui est avancé par cet auteur relève avant tout d'appréciations qualitatives. De plus, les interprétations réalisées apparaissent assez problématiques quand elles sont analysées en détail. En effet, Olivier Orain déclare une certaine équivalence des termes « espace », « milieu » et « région » tout en leur assignant tout de même des cadres sémantiques distincts[77]. Il affirme ensuite l'apparition de sens émergents du terme « espace », « plus ou moins en rupture avec la doxa » (*Ibid* 2003, 223), mais finit par dénier lui assigner une sémantique sous les arguments qu'il est « tout et dans tout », qu'il désignerait l'« objet légitime de la géographie », qu'il serait un « déictique de l'identité disciplinaire » (*Ibid* 2003, 223). Pour finir, il y a l'affirmation de sens émergents, mais sans que le changement sémantique soit vraiment précisé.

Nul doute que l'image de Caroline Reutaneuer de l'existence d'une polysémie avant le franchissement d'un seuil de rupture peut ici aider à la compréhension du processus que souhaite mettre en avant Olivier Orain. Il y aurait eu un foisonnement polysémique autour du terme « espace » qui aurait abouti à des changements irréversibles dans la signification de ce terme et d'autres qui lui étaient associés. Il n'en reste pas moins que ce processus est surtout intuité dans le travail de thèse d'Olivier Orain plus que véritablement montré. Il s'agit par conséquent de reprendre cette idée en étudiant les dynamiques sémantiques des termes « milieu » et « espace » afin de tester plus en profondeur la thèse d'Olivier Orain. Avant de réfléchir concrètement à la méthode utilisée, l'état des lieux a été poursuivi pour savoir s'il existait déjà des travaux étudiant des changements sémantiques dans une double optique, à la fois quantitative et kuhnienne, comme celle envisagée dans ce travail.

[77] « Quand on examine dans les détails les acceptions du terme inférables des usages qui en sont faits, on peut aussi bien rabattre « espace » sur les grandes notions classiques, « paysage » (qui est aussi donné comme la manifestation visible du spatial), « milieu » (« assiette » d'une certaine relation homme/nature ou globalité régie par un ensemble d'interactions), « région » (notamment conçue comme une unité organique observable) » (Orain 2003, 223).

III. D'un premier état des lieux : approches sémantiques, quantitatives et kuhniennes

La recherche bibliographique d'approches sémantiques, quantitatives et kuhniennes a abouti à un nombre très réduit de références. Dans un premier temps, il est possible de penser que ce résultat est la conséquence directe d'une requête finalement assez spécifique. Un examen approfondi des références trouvées apporte des éléments d'explication plus complexes et intéressants par rapport à la problématique de ce travail. Le fait d'avoir trouvé peu de références a permis aussi de rentrer dans les détails de celles-ci, et notamment de pouvoir réfléchir la recherche ici envisagée par rapport aux travaux existants. La section suivante détaille un ensemble de références trouvées relevant d'une méthodologie spécifique, celle de l'analyse des mots-associés (*co-word analysis*).

1. L'analyse des mots-associés

D'un point de vue technique, l'analyse des mots-associés consiste à étudier le réseau de cooccurrences de mots-clés d'un corpus donné. Tout un outillage spécifique (indice d'inclusion, indice d'association, diagramme stratégique…) existe et s'est perfectionné avec le temps (Callon *et al.* 1983; Callon 1986; Callon, Courtial, et Laville 1991). S'il existe quelques références qui s'inscrivent dans une perspective kuhnienne (Steinberg 1994; Liu *et al.* 2014), elles sont incontestablement peu nombreuses par rapport à l'ensemble des recherches se revendiquant de cette méthode. Cette observation s'explique assez facilement par le fait que l'analyse des mots-associés a été historiquement liée à la théorie de l'acteur-réseau (Callon 1986) du fait d'une création et d'une mise en avant par les mêmes auteurs. Ces derniers revendiquent un fort lien entre leur approche théorique et cet outillage méthodologique.

Il est d'autant plus intéressant de se pencher sur ce lien qu'il est construit dans un positionnement explicitement contre la théorie kuhnienne. Cette théorie y est critiquée pour son cadre fixe qui ne permettrait pas d'appréhender de manière satisfaisante toute la complexité des dynamiques de construction des connaissances. L'approche kuhnienne est assimilée à un outillage méthodologique de détection de regroupements (*clustering*) dans les réseaux de co-citations. Or, en reprenant l'hypothèse sémantique précédemment détaillée (*cf.* section Chap2.II.3.c) il est possible de penser, qu'au contraire, l'analyse de l'évolution des mots-associés peut être utilisée dans une optique kuhnienne. Il y a dans cette opposition forgée initialement par Callon *et al.* (1983) un certain coup de force pour affirmer et outiller leur théorie.

Le travail de Jean-Philippe Cointet permet une approche différente quand il affirme :

> « En dépit de la circulation incessante des entités décrites par la thèse de l'acteur-réseau, la méthode des mots-associés vise bien à identifier des structures stables qui émergent de la répétition des occurrences d'un mot dans une série de contextes différents ; quelles configurations d'équilibres émergent de la façon dont les acteurs associent les mots ou posent les problèmes. » (Cointet 2017, 16).

Ces « structures stables » ou « configurations d'équilibres » pourraient théoriquement correspondre à des paradigmes. Cette perspective permet de réhabiliter l'utilisation de l'analyse des mots-associés dans un cadre kuhnien.

De plus, par rapport à cette méthode des mots-associés, il faut préciser qu'elle est, de nos jours, très peu employée telle qu'elle a été développée historiquement par les partisans de la théorie de l'acteur-réseau. Les évolutions qui ont eu lieu se sont jouées sur plusieurs plans. Tout d'abord, de nombreuses recherches privilégient en entrée[78] le texte intégral (avec ensuite parfois une sélection à partir des termes les plus fréquents et/ou discriminants) plutôt qu'une présélection de mots-clés comme cela se faisait antérieurement. De plus, tout l'outillage complexe précédemment cité (indice d'inclusion…) est souvent remplacé par la détection de clusters (Chen *et al.* 2016). Enfin, les analyses bibliométriques ne sont plus opposées, mais utilisées en complément de l'analyse des mots cooccurrents (Braam et Moed 1991).

Toutefois, il ne faut pas négliger l'importance de cet héritage qui a établi une bipartition entre théorie kuhnienne et analyse de co-citations, d'un côté, et théorie de l'acteur-réseau et analyses des mots-associés, de l'autre. Par exemple, l'ensemble d'articles publiés sous la houlette de Chaomei Chen (2003) dans l'ouvrage *Visualizing scientific paradigms* est totalement orienté vers des analyses de co-citations. Même si cet auteur met par ailleurs en avant l'importance des analyses par mots-associés (Chen *et al.* 2002), ces dernières restent concrètement rares dans les études adoptant le modèle kuhnien comme cadre théorique.

D'autres éléments que cette bipartition peuvent être développés pour tenter d'expliquer cette rareté. Une hypothèse est que les réseaux de co-citations donnent à voir des ruptures plus nettes que celle des réseaux de mot-associés. De plus, un recul de l'épistémologie kuhnienne[79] doit sûrement être considéré aussi comme un facteur permettant de comprendre pourquoi l'emploi des méthodes de clustering sur les réseaux de cooccurrences se fait souvent sans utiliser ce cadre théorique.

---

[78] Cette expression « en entrée » renvoie ici et quand elle est utilisée par la suite au début de la chaîne de traitement des données. Elle correspond à l'expression anglaise d'« input ».

[79] Recul développé précédemment pour la géographie française (*cf.* section Chap2.II.2.b), mais qui est un phénomène plus global dépassant cette discipline.

La spécificité de ma recherche est de bénéficier grâce aux travaux d'Olivier Orain d'une lecture kuhnienne approfondie de mes corpus. Ceci explique la formalisation initiale d'hypothèses plus spécifiques et plus avancées. Il ne s'agit pas de se contenter d'affirmer que le réseau de cooccurrence global a évolué pour affirmer qu'il y a eu un changement de paradigme. L'hypothèse porte sur des dynamiques sémantiques précises et la réflexion prolongée à l'aide du travail de Caroline Reutaneuer est un apport non négligeable pour penser théoriquement cette évolution. D'après les recherches bibliographiques menées, il n'existe pas d'études présentant de telles caractéristiques. Toutefois, il faut souligner la présence d'un travail se basant sur un outillage un peu différent de celui de l'analyse des mots-associées avec un cadre théorique kuhnien bien affirmé. Il s'agit de la thèse de Carmela Chateau-Smith (2012) sur les changements sémantiques liés à l'affirmation de la théorie de la tectonique des plaques et se basant sur une méthode appelée prosodie sémantique. L'étude de cette référence a donné lieu à un ensemble de réflexions par rapport à l'étude ici envisagée.

2. La thèse de Carmela Chateau-Smith

La méthode employée par Carmela Chateau-Smith dans sa thèse relève de la prosodie sémantique. Cette dernière consiste à examiner la polarité, positive ou négative, des cooccurrents d'un mot. L'idée sous-jacente de cette auteure est qu'un tel changement de polarité peut être un bon marqueur d'un changement de paradigme. Cette idée s'appuie sur le fait qu'une inversion de polarité peut en effet renvoyer à un changement profond de valeurs. Par rapport au travail spécifique d'Olivier Orain, il est tout à fait possible qu'en étudiant certains auteurs qu'il cite (Pierre George, Jacqueline Beaujeu-Garnier…), un changement de polarité autour du terme « espace » puisse être identifié. En effet, certains auteurs ont développé des critiques du spatialisme alors que d'autres ont, au contraire, valorisé fortement cette approche. Toutefois, la perspective d'Olivier Orain dépasse ces positionnements individuels. Il promeut en effet une dynamique sémantique plus globale. Pour étudier une telle dynamique, une méthode axée seulement sur un changement de polarité des cooccurrents apparaît comme assez réductrice. En effet, si seuls les cooccurents marqués fortement par une polarité positive ou négative sont gardés, un risque encouru est de réaliser des analyses liées surtout à quelques positionnements individuels et de passer à côté d'une grande partie des dynamiques.

Sur un autre plan, l'étude de Carmela Chateau-Smith peut faire penser qu'il serait intéressant de réaliser une recherche comparative, notamment par rapport à des cas de changement de paradigmes plus avérés relevant des sciences physiques (tectonique des

plaques, théorie de la relativité…). Au-delà du problème de production de corpus comparables, il ne serait pas pertinent à mon avis d'essayer d'établir ainsi un seuil commun de changement sémantique à partir duquel il est légitime de parler d'un changement de paradigme. Il n'existe pas de seuil universel déterminant en la matière. Il ne s'agit pas ici de dénier tout intérêt pour une perspective comparative, car il serait évidemment instructif d'étudier dans une même recherche des changements sémantiques dans des disciplines relevant de modes différents de scientificité. Toutefois, par rapport à l'optique quantitative ici privilégiée, il ne me semble pas qu'une étude comparative pourrait régler le problème de détection de cette thèse au moyen de l'établissement d'un seuil universel.

Enfin, un dernier élément par rapport à la méthodologie du travail de Carmela Chateau-Smith mérite d'être souligné. Il me semble que l'optique de recherche adoptée par cette auteure est moins « critique » que celle qui a animé ma recherche. Cette différence de positionnement peut tout d'abord s'illustrer dans le statut du cadre kuhnien qui est considéré comme une donnée par Carmela Chateau-Smith, alors qu'il est une interrogation centrale dans cette thèse. Cette différence de positionnement peut aussi se lire à travers un point spécifique de l'approche méthodologique : Carmela Chateau-Smith s'appuie sur une référence au travail de Robert Daley (2004) pour justifier d'une fenêtre optimale de cooccurrence de cinq à gauche et quatre à droite[80]. Par rapport à cette affirmation d'un empan qui serait universellement plus valide que les autres, le positionnement de ce travail est tout autre. L'alternative a bien été résumée par Bénédicte Pincemin sur cette question :

> « soit on argumente pour démontrer que, parmi toutes les définitions de contexte que l'on pourrait envisager, l'une est plus pertinente que les autres ; soit on considère que la définition du contexte peut varier suivant les types de textes et les applications visées, et que c'est un paramètre à ajuster, souvent sur des considérations heuristiques (tel choix "marche mieux" que tel autre dans tel cas de figure) » (Pincemin 1999b).

D'une manière assez intuitive[81], la seconde voie d'une fenêtre définie comme un paramètre à ajuster a ici été privilégiée.

Si cet état des lieux sur les travaux articulant changement sémantique, perspective kuhnienne et approche quantitative, a permis de préciser et de développer plusieurs points, il n'a pas réglé la question du choix méthodologique. Un état des lieux plus élargi a eu lieu par rapport à cette problématique. Pour cela, j'ai effectué une recherche en m'intéressant

[80] La fenêtre de cooccurrence permet de définir la zone contextuelle dans laquelle si deux mots sont présents, alors ils sont définis comme cooccurrents. Si un mot est proche d'un autre, mais hors de cette fenêtre de cooccurrence, les deux mots ne sont pas cooccurrents.

[81] Intuitif au sens où ce n'est pas passé par un travail critique sur la référence originelle employée par Carméla Chateau-Smith (Daley *et al*. 2004) mais bien plutôt par un positionnement semblant *a priori* plus ouvert et pertinent.

toujours aux travaux sur le changement sémantique avec une approche quantitative, mais en abandonnant la contrainte d'un rattachement à une perspective kuhnienne.

IV. à un second état des lieux : approches sémantiques et quantitatives sans perspective kuhnienne

1. De l'axe syntagmatique aux plongements de mots

Le nombre important de références trouvées confirme que le constat effectué par Bénedicte Pincemin en 2009 d'une grande diversité d'outils de la statistique textuelle pour travailler sur les sèmes (et donc potentiellement le changement sémantique) est toujours d'actualité. Par exemple, Julien Longhi et André Salem mobilisent la technique des segments répétés et l'analyse factorielle des correspondances pour étudier le changement sémantique autour du terme « ennemi » dans la série chronologique du Père Duchesne (Longhi et Salem 2018) ; Armelle Boussidan mobilise le concept de clique sur le réseau de cooccurrences pour étudier l'évolution sémantique du terme de mondialisation entre 1998 et 2001 (Boussidan 2013) ; de nombreux travaux proposent des adaptations du modèle bayésien LDA pour prendre en compte l'évolution temporelle et s'intéresser plus spécifiquement à la problématique du sens des mots (Frermann et Lapata 2016). D'une manière globale, il est intéressant de noter que tous les travaux s'attaquant au problème complexe du sens des mots s'appuient plus ou moins explicitement sur l'affirmation de John Rupert Firth (1957) : « you shall know a word by the company it keeps »[82]. En effet, il n'existe pas de sens pour un mot isolé. C'est le contexte qui permet d'étudier cette dimension sémantique d'un terme.

Dans sa thèse, Coralie Reutenauer (2012) étudie l'évolution des cooccurrents du terme « tsunami ». Alors qu'initialement ce mot est accompagné d'un vocabulaire relevant assez exclusivement des sciences de la nature (« mer », « vague », « terre », « côte »…), il y a une diversification des cooccurrents avec notamment l'apparition de termes marqués par le monde de la finance (« capitalisme », « économie réelle », « zone euros », « dollars »…). Cette méthode de la coocurrence s'appuie sur les relations syntagmatiques (la proximité des termes), mais elle ignore les relations paradigmatiques (l'utilisation de termes plus ou moins substituables dans une même structure). Pour illustrer ce phénomène, imaginons un texte contenant plusieurs occurrences de ces deux syntagmes : « l'espace économique régional » et « le milieu économique régional ». Dans cet exemple, après avoir défini une fenêtre de cooccurrence au moins supérieure à 1 à droite, le terme « espace » cooccurre avec les mots

[82] Traduction personnelle : « Vous pouvez connaître un mot à partir de la compagnie qu'il possède ».

« économique » et « régional ». De la même manière, « milieu » cooccurre alors avec « économique » et « régional ». Cependant, « espace » et « milieu » ne vont pas forcément cooccurrer[83]. Le fait que ces deux termes soient pris dans une structure lexicale similaire n'est pas pris en compte en utilisant le calcul de cooccurrence.

Or, une partie de ce que soutient Olivier Orain repose dans un premier temps sur une équivalence sémantique entre les termes « espace », « milieu » et « région » (*cf.* section Chap2.II.3.c). Il y a donc une certaine légitimité à essayer de prendre en compte cet axe paradigmatique. Ce sont les recherches de Zellig S. Harris (1969) avec des analyses dites distributionnelles qui sont le plus souvent citées comme la référence historique sur ce sujet. Il faut ici préciser qu'à la base, le critère utilisé était surtout la grammaticalité. Ainsi, dans l'exemple « le chat est noir », il est possible de remplacer « blanc » par « noir ». Ces deux termes font alors partie du même paradigme[84]. Cependant, grammaticalement, il est correct d'écrire « le chat est connecté ». « connecté » appartient donc aussi au paradigme contenant « noir » et « blanc ». Dans ce cas, c'est une catégorie grammaticale que permet d'approcher cet axe paradigmatique. Cependant, en ne partant plus d'hypothèse grammaticale fictive, mais de corpus existants, il a été montré que la reprise de ce principe de test de commutation au sein d'une phrase permet de mettre en avant des classes de mots ayant une véritable proximité sémantique (Rubenstein et Goodenough 1965).

Des mesures spécialisées dans le calcul d'une proximité sur l'axe paradigmatique existent. Elles consistent à calculer un vecteur pour chaque mot en fonction des termes qui composent ses contextes d'apparition. Plusieurs métriques existent ensuite pour évaluer la proximité entre deux vecteurs (Ferret 2010). Une limite de ces approches est qu'elles ne prennent pas en compte l'axe syntagmatique. S'appuyant toujours sur le principe de ces analyses distributionnelles, une autre approche s'est développée ces dernières années. Elle a donné lieu à un champ de recherche particulièrement dynamique en utilisant des approches basées sur les réseaux neuronaux. Ces méthodes appelées « plongements de mots » ont été implémentées initialement par Bengio *et al.* (2003), mais le véritable développement de cette voie de recherche ne s'est réalisé qu'après les travaux de Mikolov *et al.* (2013) avec la diffusion de la méthode *Word2Vec*.

[83] Cela dépend de la fenêtre de cooccurrence définie et de la proximité des deux syntagmes « l'espace économique régional » et « le milieu économique régional ».

[84] Ici, paradigme est pris dans son sens linguistique désignant une classe d'éléments homogènes qui peuvent être substitués les uns aux autres en laissant l'énoncé grammaticalement correct.

L'intérêt de ces méthodes est de « condenser » l'information sémantique[85]. Il faut souligner que ces méthodes sont généralement appliquées sur de très gros corpus, mais des expérimentations montrent qu'ils peuvent aussi être efficaces sur des corpus beaucoup plus réduits. Par exemple, Alix Rule *et al.* (2015) ont appliqué un tel outillage sur les discours de l'Union des États-Unis de 1790 à 2014. Les résultats obtenus sont convaincants. Ces auteurs montrent ainsi que les voisins sémantiques du terme « land » renvoient à plusieurs univers thématiques relevant des frontières (river, territory, shore, possession, etc.), de la propriété (ownership, private, public domain, title, etc.) et des ressources (soil, mine, natural ressource, etc.).

Le choix des plongements de mots correspond aussi, comme l'introduction de cette thèse, l'a mentionné à la conséquence d'une certaine acculturation (*cf.* section Chap1.III.1). Ce n'est qu'après avoir expérimenté et trouvé les résultats assez convaincants pour être valorisés que j'ai adopté ces techniques.

2. Principales méthodes de plongement de mots

Un point commun des méthodes de plongement de mots détaillées dans cette section est d'obtenir pour chaque terme du vocabulaire du texte donné en entrée un vecteur de nombres. Les termes qui ont des contextes similaires d'apparition vont obtenir pour résultat des vecteurs correspondants proches[86]. En reprenant le principe des analyses dites distributionnelles, à savoir que deux termes avec des contextes similaires d'apparition ont de grande chance d'avoir un sens proche, il est facile de comprendre pourquoi ces méthodes de plongement de mots sont utilisées pour calculer des proximités sémantiques. De plus, il a été montré expérimentalement que les vecteurs produits par ces algorithmes permettent de résoudre un grand nombre d'analogies. Ainsi, si un plongement de mots fonctionne bien, en prenant le vecteur correspondant au mot « roi », puis en lui soustrayant le vecteur

[85] Notamment par rapport aux méthodes précédentes qui représentent chaque mot par un vecteur composé à partir des termes de ses différents contextes. Ces méthodes se caractérisent le plus souvent par des vecteurs dits creux, c'est-à-dire composés par beaucoup de composantes nulles. En effet, pour être comparable à d'autres vecteurs, il faut représenter dans le vecteur d'un terme les mots composant ses contextes d'apparition, mais aussi ceux qui ne le composent pas (mais qui vont composer les contextes d'apparitions d'autres termes). La conséquence est alors en général la création des longs vecteurs composés de beaucoup de 0. L'information contenue pas ces vecteurs est considérée comme peu « dense ». D'où l'expression précédente de « condensation » de l'information sémantique avec la production par les méthodes de plongements de vecteurs avec moins de 0 et composé à partir de la distribution (syntagmatique et paradigmatique) des termes.

[86] Il existe plusieurs manières pour calculer la distance entre deux vecteurs. Une des plus classiques est la similarité cosinus.

correspondant au mot « homme » et enfin en lui additionnant le mot « femme », le terme le plus proche du vecteur ainsi créé devrait être « reine ». Ce qui peut être résumé par l'équation suivante : roi – homme + femme = reine.

Les analogies qui peuvent être résolues sont multiples. Un test connu (Gladkova, Drozd, et Matsuoka 2016) permettant d'évaluer les plongements de mots donne un aperçu de cette diversité en distinguant les analogies d'inflexion (singulier/pluriel…), de dérivation (nom + suffixe[87]…), de lexicographie (hyperonyme[88]…) et d'encyclopédie (géographie[89]…). Les résultats obtenus à partir d'un corpus massif[90] avec un algorithme classique (*Word2Vec Skip-Gram*) sont loin d'être parfaits : autour de 61 % des analogies d'inflexions, 11 % des analogies de dérivation, 9 % des analogies lexicographiques et 26 % des analogies encyclopédiques sont effectivement trouvées (Drozd, Gladkova, et Matsuoka 2016). Les autres méthodes classiques de plongement de mots permettent des gains mineurs (quelques points de pourcentage) sur certaines catégories, mais restent de cet ordre de grandeur.

Cependant, ces résultats sont suffisamment meilleurs que ceux obtenus par les outils antérieurs, expliquant pourquoi les plongements de mots ont été un champ de recherche très dynamique[91] pendant toute la période d'élaboration de cette thèse. Par rapport aux développements théoriques précédemment effectués, il faut préciser que ce ne sont pas des sèmes qui sont trouvés par les plongements de mots. Toutefois, comme il n'existe pas de méthode permettant de détecter automatiquement les sèmes, il faut trouver des outils permettant non pas de déterminer, mais plutôt d'approcher la sémantique des termes. Dans ce cadre-là, les plongements de mots sont des techniques imparfaites, mais qu'il est légitime d'utiliser. Mes expérimentations tendent à me faire penser que plus les détections des analogies lexicographiques et encyclopédiques sont bonnes par une méthode de plongement de mots, plus les plus proches voisins d'un terme obtenus sont sémantiquement intéressants.

Les sections suivantes détaillent le fonctionnement des algorithmes les plus classiques : la méthode *Word2Vec* avec ses deux modèles *Continuous Bag of Word* (*CBOW*) et *Skip-gram*, la méthode *GloVe* et enfin la méthode *FastText* avec les deux mêmes modèles : *CBOW* et *SkipGram*. La présentation commence par *Word2Vec* car sa date de création est antérieure aux deux autres.

---

[87] Par exemple, accomplishment - accomplish + achieve = achievement

[88] Comme cat – feline + dog = canine

[89] Par exemple, Paris – France + Grèce = Athènes

[90] Corpus constitué du contenu anglais de Wikipedia en 2015 (1,8B tokens), d'Araneum Anglicum Maius (1.2B) et de l'ukWaC (2B) avec une fenêtre de contexte de 8, une taille de vecteur demandé de 300.

[91] Il faut ici souligner pour comprendre cette dynamique que les plongements de mots ont été appliqués avec succès à un nombre important de domaines : étiquetage morpho-syntaxique, reconnaissance d'entités nommées, analyse de sentiments…

a. Fonctionnement de la méthode *Word2Vec*

La méthode *Word2Vec* a été développée par (Mikolov *et al.* 2013) . Cette sous-section présente dans un premier temps de manière simplifiée le modèle *CBOW*.

Ce modèle demande en entrée un corpus et une fenêtre de contexte. En guise d'exemple, la phrase suivante sera considérée comme un corpus : « L'armature urbaine joue un rôle fondamental dans l'organisation régionale » et la fenêtre de contexte sera fixée à 4 (2 à gauche et 2 à droite). Si nous considérons par exemple le mot cible « joue », le contexte retenu va être {« armature », « urbaine »} à gauche et {« un », « rôle »} à droite. Le modèle *CBOW* vise à prédire le mot cible à partir de son contexte.

Pour cela, chaque mot du contexte est représenté par un vecteur correspondant à la taille du vocabulaire (c'est-à-dire le nombre de mots dans le corpus) avec des 0 partout sauf à l'emplacement du mot en question. Dans notre exemple, le vocabulaire est l'ensemble : {« l' », « armature », « urbaine », « joue », « un », « rôle », « fondamental », « dans », « organisation », « régional »}. En suivant la règle précédente, « armature », c'est-à-dire le premier mot du contexte de « joue » qui est noté dans le schéma suivant w(t-2) du fait de sa position de deuxième mot à gauche, est encodé [0,1,0,0,0,0,0,0,0,0]. Sur le même principe, il est possible d'encoder l'ensemble des autres mots, notamment w(t-1), w(t), w(t+1), w(t+2). Le fonctionnement du modèle *CBOW* peut alors être modélisé ainsi :

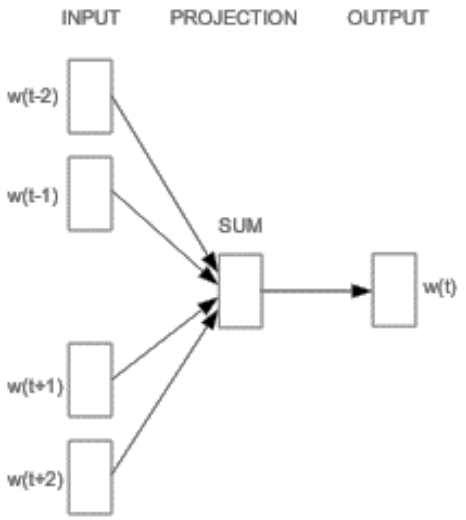

Figure n°2 : Fonctionnement du modèle CB*OW* (Mikolov *et al*. 2013).

À partir des entrées (INPUT) : w(t-2), w(t-1), w(t+1), w(t+2), l'objectif est que le modèle arrive à prédire en sotie (OUTPUT) : w(t). Le modèle a une couche intermédiaire ici appelée

PROJECTION. Ce graphique extrait de la publication originelle de (Mikolov *et al.* 2013) est trop synthétique pour comprendre même de manière simplifiée ce qui est réalisé par l'algorithme. À cette fin, un nouveau schéma est ici présenté :

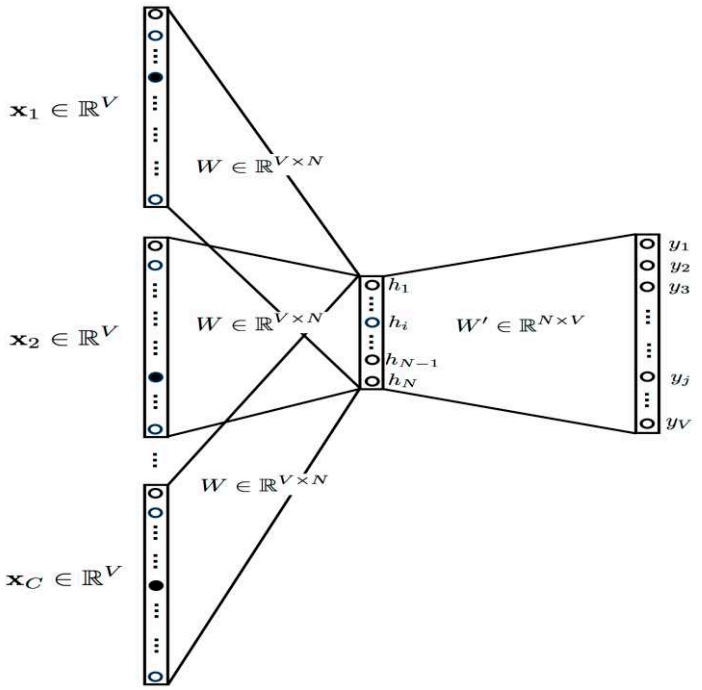

Figure n°3 : Fonctionnement plus détaillé du modèle *CBOW*.
Source : https://fr.wikipedia.org/wiki/Word_embedding, consulté le 25 juin 2020.

Par rapport à la Figure n°2, l'entrée dépend ici de la taille du contexte choisie (noté C dans Xc). Dans l'exemple précédent, le contexte était de 4, mais il peut être différent (d'où l'existence de cette variable dans le schéma ci-dessus). Le codage des mots sous forme de 0 et de 1 est représenté ici par des points blancs et noirs dans la suite du vocabulaire. Un point blanc correspond à un 0 et un point noir à un 1 par rapport à l'encodage précédemment explicité. La taille du vocabulaire est notée V. Ce que permet d'appréhender ce nouveau schéma, c'est la présence d'un paramètre N qui doit être fixé par l'utilisateur qui correspond à la taille de la couche intermédiaire qui va être un vecteur à N dimensions. Ce qui était représenté dans le schéma précédent par des flèches simples de transformation est en fait concrètement 2 matrices : une notée W de taille (VxN) et une notée W' de taille (NxV).

L'algorithme correspondant au modèle *CBOW* cherche à optimiser ces deux matrices pour trouver un vecteur final (y1, y2…, yv) le plus proche possible du résultat attendu. Cette optimisation est réalisée en passant sur tous les mots du corpus en déplaçant à chaque fois la fenêtre contextuelle précédemment définie. Elle est réalisée de manière itérative avec

comme paramètre fixé par l'utilisateur le nombre d'itérations, appelé « epochs », à réaliser. Une fois cette optimisation effectuée, ce n'est pas le vecteur final (y1, y2…, yv) obtenu qui est le plus intéressant, mais le contenu de la matrice W. En effet, cette dernière contient autant de lignes que de mots du vocabulaire. La ligne i de cette matrice est un vecteur de taille N qui correspond au résultat du plongement pour le mot i. Pour plus de détails techniques sur l'ensemble du processus d'optimisation qui est aussi appelé phase d'apprentissage ou d'entraînement, nous renvoyons le lecteur intéressé à l'article original (Mikolov *et al*. 2013).

Le second modèle constituant la méthode *Word2Vec* est *Skip-gram*. À l'inverse de *CBOW*, l'objectif est de prédire les éléments de contexte à partir du mot cible. Par rapport à l'exemple précédent, il s'agirait d'inférer à partir du terme « joue » le contexte {« armature », « urbaine », « un », « rôle »} si la fenêtre contextuelle choisie est toujours de 4. Schématiquement, le modèle *Skip-gram* peut être représenté ainsi :

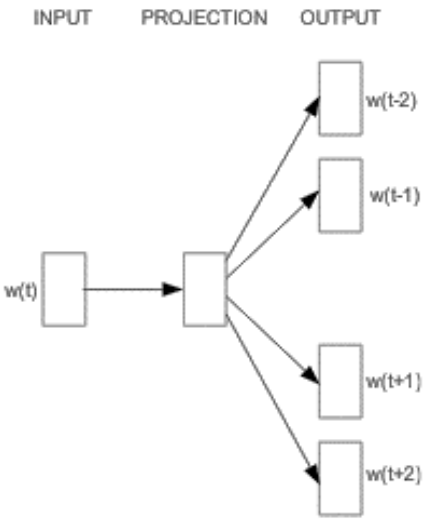

Figure n°4 : Fonctionnement du modèle *Skip-Gram* (Mikolov *et al*., 2013).

Comme pour le modèle *CBOW*, ce schéma peut être développé avec des matrices W et W'. De la même manière, les résultats sont pour chaque mot un vecteur de dimension N issu de la matrice W. N étant toujours un paramètre qui doit être fixé par l'utilisateur en amont. Par rapport au modèle *CBOW*, il est reconnu que le modèle *Skip-gram* permet d'obtenir de meilleurs résultats pour les mots rares, mais il est moins précis pour les mots fréquents. Toutefois, le choix ne se réduit pas à une alternative puisque deux autres algorithmes qui

font aussi maintenant partie des méthodes classiques de plongement de mots peuvent également être utilisés : *GloVe* et *FastText.*

b. Deux autres méthodes classiques de plongements de mots : *GloVe* et *FastText*

La méthode *GloVe* (Pennington, Socher, et Manning 2014) demande comme les modèles de *Word2Vec* en entrée un corpus, une fenêtre de contexte, une taille pour les vecteurs constituant les résultats et un nombre d'itérations. Cette ressemblance dans les paramètres initiaux cache des processus de calcul totalement différents. Dans le cas de *GloVe*, une matrice de cooccurrence globale des mots est créée à partir de la fenêtre contextuelle définie. La méthode peut être décrite comme une décomposition de la matrice Mots-Contextes (MC) en deux matrices : une Mots-Variables (MV) et une Variables-Contextes (VC).

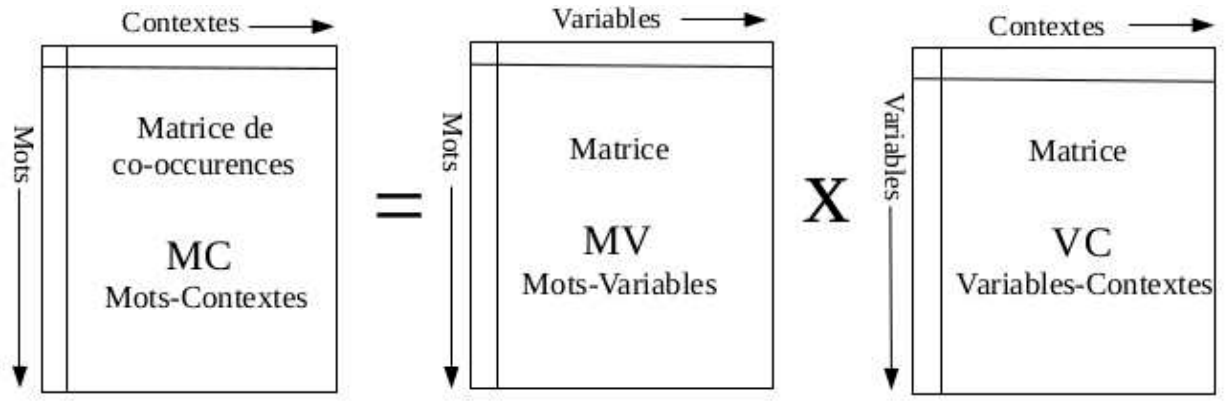

Figure n°5 : Principe de fonctionnement de *GloVe.*
Source : (Dipanjan, 2018)

Le nombre de variables correspond à la taille des vecteurs de plongement obtenus. Cette taille est un paramètre choisi par l'utilisateur. Les matrices MV et VC sont initialisées au départ avec des poids aléatoires et l'algorithme utilise la méthode de descente de gradient stochastique pour minimiser les erreurs dans l'approximation de la matrice MC au fur et à mesure des itérations. Le vecteur de chaque mot est contenu dans les lignes de la matrice MV. Ainsi, des vecteurs ayant des propriétés similaires que ceux précédemment présentés pour les modèles de la méthode *Word2Vec* sont obtenus. Par rapport au test précédent (Gladkova, Drozd, et Matsuoka 2016), des gains pour les analogies lexicographiques et encyclopédiques peuvent être observés, passant respectivement à 11 % et 31,5 %. Pour autant, *GloVe* est loin d'avoir supplanté *Word2Vec*. Il y a une coexistence des différentes

méthodes avec, suivant les auteurs et les tâches à réaliser, une préférence pour l'une ou l'autre. Il faut souligner que *GloVe* demande une ressource en mémoire vive beaucoup plus importante pour travailler sur la matrice de cooccurrences. La forte popularité de *Word2Vec* s'explique aussi par de nombreuses extensions réalisées à partir du même fonctionnement.

Parmi ces dernières, une très connue est *FastText* (Bojanowski *et al.* 2017). Cet algorithme ne travaille plus au niveau du mot comme *Word2Ve*c mais des n-grams. En choisissant par exemple le terme « young » et des n-grams de n=4 à 6, l'ensemble suivant de n-gram va être créé {<you, <youn, <young, youn, young, young>, oung, oung>, ung>}, car *FastText* ajoute deux signes « < » et « > » comme marqueur de début et de fin de mot. En réalisant cette opération pour tous les mots d'un corpus, des vecteurs correspondants aux plongements de tous ces n-grams sont obtenus en utilisant les modèles de la méthode *Word2Vec* (*CBOW* ou *Skip-gram*). L'intérêt de ce changement d'échelle, des mots aux n-grams, est qu'il y a une amélioration de la représentation sémantique de termes moins fréquemment employés, mais qui sont construits sur la même base que d'autres plus employées par ajout de préfixe et de suffixe. Par exemple, il est possible d'imaginer que dans un corpus, le terme « young » est sémantiquement proche d'« adolescent ». Il est possible que d'autres formes comme « preadolescent » soient plus rares et par conséquent moins bien représentées en utilisant classiquement la méthode *Word2Vec*. *FastText* permet d'obtenir que les vecteurs des n-grams de « young » soient globalement proches de plusieurs vecteurs de n-grams d'« adolescent », mais aussi du même coup proche de ceux du terme « preadolescent », ce qui contribue à améliorer sa représentation sémantique.

Ce mécanisme explique des gains importants notamment dans la détection des analogies dites d'inflexion et de dérivation dans le test précédent avec une baisse mineure sur les analogies dites encyclopédiques (Lakmal *et al.* 2020). Dans la perspective de cette recherche, il est possible d'imaginer par exemple un rapprochement d'« espace » et de « spatial » par l'intermédiaire du tri-gram « spa ». Toutefois, ces formes sont très employées dans mon corpus et il n'est pas donc certain qu'il soit utile d'utiliser *FastText*. Il est possible aussi que cette méthode engendre un certain bruit, du fait notamment de termes dont la syntaxe est proche, mais non la sémantique. Seule l'expérimentation est susceptible d'apporter des réponses à cette question, car il n'y a pas de règle générale. La spécificité de chaque corpus et de chaque questionnement joue un rôle important en la matière. Dans le cadre de notre problématique, tout un autre champ de recherche se doit également d'être exploré : celui des plongements de mots diachroniques.

c. Plongements de mots diachroniques

La première étude inaugurant cette voie de recherche est celle de Kim *et al.* (2014) qui consiste à considérer les plongements de mots appris à la période t comme initialisation pour apprendre ceux de la période t+1. La variation des distances existantes entre plusieurs termes (par l'intermédiaire des vecteurs calculés) peut ensuite être visualisée dans le temps. Par exemple, l'évolution des similarités cosinus entre le vecteur du terme « gay » et les vecteurs de ses plus proches voisins en 1900 (« cheerful », « pleasant ») et en 2009 (« lesbian », « bisexual ») est représentée sur cette figure.

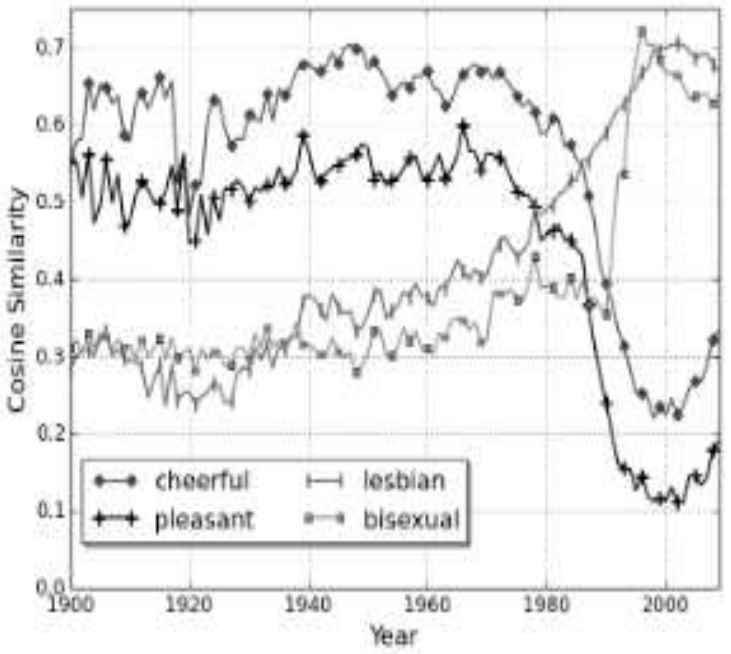

Figure n°6 : Évolution de la similarité cosinus des plus proches voisins du terme « gay » en 1900 et 2009 d'après les travaux de Kim *et al.* (2014).

Cette représentation permet de figurer de façon remarquable un changement sémantique. Des méthodes plus sophistiquées ont été ensuite proposées consistant à calculer les plongements à différentes périodes et à essayer d'aligner les espaces vectoriels obtenus à l'aide de la meilleure transformation linéaire possible (Kulkarni *et al*. 2015; Hamilton, Leskovec, et Jurafsky 2016; Dubossarsky, Weinshall, et Grossman 2017; Szymanski 2017).

Quelques études peuvent être considérées comme particulièrement intéressantes pour ce travail, car elles revendiquent une approche de l'évolution non plus des mots, mais des concepts. Une étude plus poussée de ces publications amène à être prudent par rapport à une

telle affirmation. Par exemple, Wevers *et al.* (2015) proposent un système qui peut prendre en entrée non pas un mot, mais un groupe de mots. L'évolution dans le temps peut être suivie grâce à deux méthodes : une « non-adaptative » où l'apprentissage est fait de manière indépendante pour chaque époque ; une « adaptative » où les résultats obtenus à une époque t servent à constituer le nouveau groupe de mots pris comme référence à l'époque t+1. La dimension conceptuelle repose ainsi surtout sur cette échelle du groupe de mots ce qui est discutable. Recchia *et al.* (2017) proposent une méthode de clustering pour mieux séparer des groupes de mots qui relèvent a priori de concepts différents. Leur conclusion montre bien les limites d'un tel travail : « Toutes ces considérations soulignent la complexité et la promesse des nouvelles approches pour suivre le vocabulaire associé à des concepts spécifiques au fil du temps »[92]. Cette citation montre en creux une recherche balbutiante attirée par les lumières d'une approche cognitiviste du sémantisme, mais qui doit reconnaître que le passage des mots aux concepts est loin d'être résolu.

Pour finir cet état des lieux avant de mieux situer ce travail, deux évolutions importantes qui ont eu lieu au cours de cette thèse, doivent être abordées.

d. Plongements de graphes et plongements de mots contextualisés

Les méthodes de plongement de graphes permettent de ne pas réduire une phrase à une suite continue de mots, mais au contraire de pouvoir prendre en compte sa structure syntaxique. Dans la pratique, l'étude de Zucherman *et al.* (2019) montre que les résultats ainsi obtenus par exemple sur le test de résolution d'analogies (Gladkova, Drozd, et Matsuoka 2016) sont moins bons qu'avec des algorithmes classiques de plongement de mots. La raison invoquée est la création d'un certain « bruit » provenant du fait que des phrases peuvent avoir des structures similaires alors que leurs contenus sémantiques sont très éloignés. D'autres approches doivent néanmoins être soulignées comme celle de *Word-Node2Vec* (Sen, Ganguly, et Jones 2019) qui permet de prendre mieux en compte pour les cooccurrences le niveau du document. Les auteurs affirment une amélioration de la tâche de résolution d'analogies, mais les gains restent mineurs par rapport à *GloVe* ou *FastText*.

Les plongements de mots contextualisés – *ELMo* (Peters *et al.* 2018), *BERT* (Devlin *et al*. 2018), GPT… – permettent des gains nettement plus importants. La nouveauté de ces méthodes est de pouvoir produire des vecteurs différents pour un même terme suivant ses

[92] Traduction personnelle de : « All of these considerations underscore the complexity and promise of novel approaches to tracking the vocabulary associated with particular concepts over time » (Recchia *et al.,* 2016).

contextes d'apparition. Sans rentrer dans le détail de ces méthodes complexes puisqu'elles n'ont pas été mobilisées dans cette thèse, il est nécessaire de reconnaître qu'il y aurait eu une pertinence à les utiliser. Ce regret m'amène à justifier plus précisément le choix de la méthode retenue pour cette recherche.

3. Explicitation de la méthode utilisée et raisons de ce choix

Il y a dans une thèse plusieurs étapes. Il existe notamment un temps pour l'expérimentation et un autre pour la finalisation. Le format d'une thèse implique un temps long de finalisation. Pendant celui-ci, la recherche continue d'avancer. À un moment donné, j'ai choisi de ne plus modifier la méthodologie malgré les innovations existantes. Cela nécessite de reconnaître le caractère daté et situé de la recherche. Le choix effectué, à savoir les plongements de mots non contextualisés en utilisant les modèles de base (*Word2Vec*, *GloVe*, *FastText*) peut et se doit d'être ainsi compris. Ce choix correspond à l'aboutissement d'une phase de trois ans de réflexions et d'expérimentations aussi bien au niveau de la problématique que des méthodologies (*cf.* Chap2). Toutefois, le chemin effectué ne me semble pas la seule raison expliquant ce choix.

Une deuxième raison est que ce socle (*Word2Vec*, *GloVe*, *FastText*) est généralement bien identifié alors que les techniques qui en dérivent forment un champ plus profus et moins stabilisé. Ce constat est partagé par Syrielle Montariol et Alexandre Allauzen sur la question de la diachronie :

> « Le domaine en plein essor qu'est l'apprentissage de plongements de mots dynamiques manque encore de la cohésion que possèdent les tâches plus anciennes du traitement automatique des langues […]. Notons qu'un cadre d'évaluation est difficile à définir dans le cas qui nous intéresse, tant les attentes applicatives vis-à-vis d'un modèle diachronique peuvent varier » (Montariol et Allauzen 2019, 12)

Au niveau de cette étude, l'objectif a été évidemment d'utiliser les technologies qui semblaient adaptées pour pouvoir répondre au problème posé, mais sans forcément faire de la dernière technologie disponible une nécessité absolue. Il y a eu un certain compromis réalisé entre l'avancée technologique, le besoin de terminer ce travail et la nécessité d'avoir un certain recul par rapport aux outils utilisés.

Un troisième ordre de raison est plus technique et pragmatique. Tout d'abord, il faut souligner que les algorithmes de plongement de mots contextualisés n'étaient pas si faciles à implémenter dans les premiers temps de leurs développements (2018-2020). Ensuite, les phrases étant une unité de référence importante dans ces outils les plus avancés, il faudrait à mon avis retravailler la qualité des textes utilisés. En effet, certaines erreurs d'OCR

perturbent la reconnaissance de ces structures de base dans le corpus utilisé pour cette recherche. Des améliorations sont possibles par rapport à ce problème, mais elles demandent un travail supplémentaire important.

Pour finir, la technique utilisée doit être détaillée pour saisir un aspect pragmatique du choix réalisé. La démarche retenue peut être décomposée en quatre étapes :

1) Calculer des plongements de mots pour chaque époque différente en essayant les modèles de plongements de mots les plus classiques (*Word2Vec*, *GloVe*, *FastText*).
2) Après avoir choisi la méthode de plongement semblant la plus efficace, retenir les p-voisins les plus proches d'un terme choisi[93] (p étant un paramètre fixé par l'utilisateur).
3) Sur les vecteurs de ces p-voisins, appliquer des méthodes de clustering pour essayer d'organiser des regroupements sémantiques de termes voisins du terme choisi.
4) Comparer dans le temps l'évolution des clusters (groupes) obtenus pour chaque époque.

Le facteur décisif pour comprendre cette méthode est le travail antérieurement effectué sur l'évolution thématique. En effet, une recherche importante avait été réalisée en reprenant et développant une méthode pour suivre l'évolution de clusters dans le temps (Dugué, Lamirel, et Cuxac 2016a). Cette partie est détaillée par la suite (*cf.* section Chap7.III.3). Ce qu'il faut ici comprendre, c'est que cet outil sur lequel tout un travail avait déjà été effectué, a été logiquement testé sur cette problématique de l'évolution sémantique. Les expériences menées étant assez concluantes et le temps des expérimentations ayant été déjà bien consommé, il y a eu une adoption de cette démarche méthodologique. Pour finir, il faut mentionner une évolution décisive dans les réflexions tirées de ces expérimentations : elles m'ont conduit à penser que les conclusions de fond réalisées ne seraient pas modifiées par les améliorations techniques alors disponibles. Tous ces facteurs expliquent le choix méthodologique retenu.

Avant d'appliquer ces méthodes et outils, il est nécessaire de présenter les données et de délimiter plus finement le corpus, ce qui est l'objet du chapitre suivant.

[93] Dans le cas de cette recherche, les termes « espace » et « milieu ».

Chapitre 4 :
Présentation des données, exploration et délimitation des corpus

Les données utilisées proviennent de l'Unité d'Appui et de Recherche (UAR) *Persée* qui a pour principal objectif de numériser et de mettre à disposition sur le Web des documents produits historiquement par les SHS.

I. Présentation des données

1. Au-delà du PDF et des données ouvertes

Le portail créé en 2005 par cette UAR permet ainsi d'accéder aux deux revues utilisées pour cette recherche :

- Les *Annales de Géographie* : https://www.persee.fr/collection/geo
- *L'Espace Géographique* : https://www.persee.fr/collection/spgeo

Ce portail permet de naviguer dans chaque revue par année, de consulter librement les articles et de les télécharger en fac-similé (document PDF).

L'UAR *Persée* ne peut pas fournir directement tous les documents PDF d'une revue, car cet organisme ne stocke pas ces fichiers, mais les régénère à chaque demande. Concrètement, l'UAR *Persée* ne fait que réassembler des images et des données textuelles pour former ses documents PDF, mais ce détail est symboliquement intéressant dans l'optique de ce travail. En effet, ce qui peut apparaître pour l'utilisateur habituel du portail *Persée* comme le document de base, celui qui rend le mieux compte du document physique initial, n'est en fait qu'une reconstruction permanente. Ce fait permet de rappeler que le document numérique n'est pas un simple décalque d'un document « réel ». Il est le résultat d'une construction qu'il est utile de comprendre pour mieux utiliser les données acquises.

La chaîne de documentation de l'UAR *Persée* peut être décomposée en 5 étapes représentées par le graphique ci-dessous :

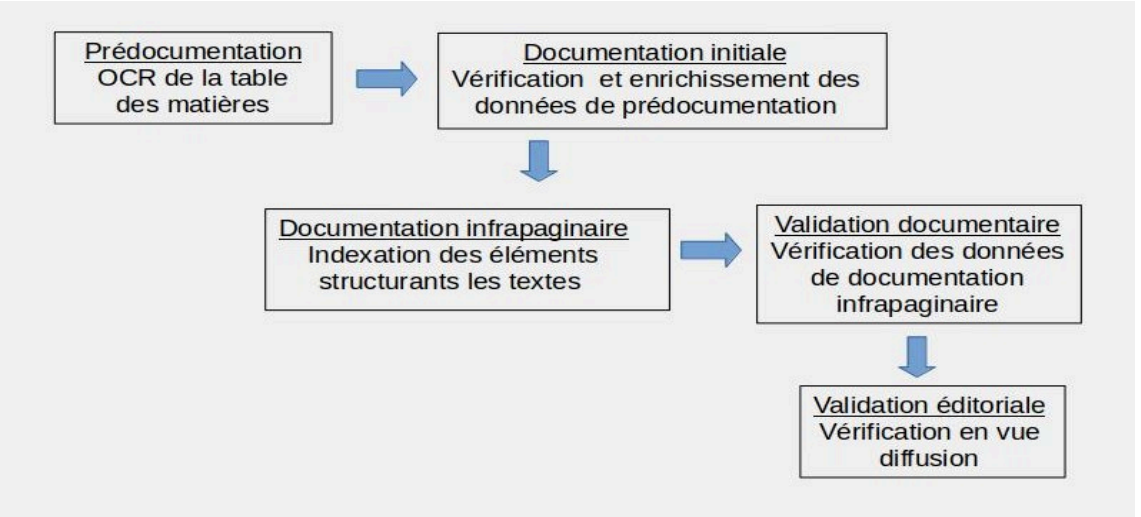

Figure n°7 : La chaîne de documentation de l'UAR *Persée.*

L'objectif n'est pas ici de rentrer dans les détails de chaque étape. Celle qui mérite le plus d'attention dans notre perspective est la documentation infrapaginaire. Ce travail consiste à passer sur chaque page d'un document pour annoter tous ses éléments structurants (titres, figures, notes, bibliographies…). Pour chaque élément, une fenêtre graphique est tracée sur la zone concernée et des informations sont précisées. Par exemple, si le titre a été mal reconnu par la Reconnaissance Optique de Caractères (OCR), une version corrigée est notée. Ces informations servent ensuite à l'indexation du document sur le Web. L'objectif est qu'un utilisateur trouve en fonction de sa recherche sur internet les données les plus pertinentes. L'idée est ici d'utiliser ces données de documentation infrapaginaire dans un autre objectif à savoir la préparation des données textuelles plus pertinentes à analyser.

Il est en effet possible, à partir de ces informations, d'enlever certains éléments comme les notes de bas de page, les bibliographies… Le but de ces opérations est ainsi d'éviter le bruit statistique que peuvent engendrer ces éléments qui relèvent de genres textuels spécifiques. Toutefois, les informations de documentation infrapaginaire ne sont pas présentes dans les documents PDF. Il est par conséquent nécessaire de s'appuyer sur un autre format produit par l'UAR *Persée* : le XML-TEI. Ce format n'est pas disponible en accès libre. La section suivante détaille les données acquises sous ce format.

2. Les données en XML-TEI

L'abréviation XML provient de l'anglais « eXtensible Markup Language »[94]. L'abréviation TEI fait référence à la « Text Encoding Initiative » qui est une communauté qui fixe des recommandations pour encoder des documents textuels. L'application Web permet d'explorer un tel fichier en cliquant sur ce lien : **@Afficher exemple XML-TEI**[95]. Les différentes balises peuvent être ensuite ouvertes et fermées à volonté en cliquant sur les petits items à gauche de chaque balise. Deux grandes parties structurent le fichier :

- La première, l'en-tête (balise <tei:teiHeader>), contient les métadonnées de l'article : son titre, son ou ses auteurs…
- La deuxième, le corps (balise <tei:teibody>) contient le détail de l'OCR localisé avec les coordonnées graphiques de chaque mot (balise <word> avec les attributs « left », « top », « right » et « bottom »)

```
−<tei:body>
   <tei:pb xml:id="geo_0003-4010_1966_num_75_409_T1_0268_0000" n="268"/>
  −<tei:p>
     <word left="415" top="842" right="650" bottom="913" order="1">Après</word>
     <word left="709" top="841" right="778" bottom="912" order="2">la</word>
     <word left="834" top="865" right="1157" bottom="912" order="3">réforme</word>
     <word left="1214" top="868" right="1492" bottom="913" order="4">agraire</word>
     <word left="1547" top="841" right="1958" bottom="878" order="5">iranienne*</word>
   </tei:p>
  −<tei:p>
     <word left="1563" top="1031" right="1649" bottom="1060" order="6" bold="true">par</word>
     <word left="1681" top="1018" right="1893" bottom="1060" order="7" bold="true">Hossein</word>
     <word left="1926" top="1018" right="2010" bottom="1061" order="8" bold="true">Ma</word>
```

Figure n°8 : Exemple du fichier XML-TEI présenté correspondant au début du corps avec l'OCR localisé.

Les informations correspondant aux coordonnées graphiques de chaque mot proviennent d'un temps d'échanges en amont qui a permis à l'UAR *Persée* de générer des données leur semblant les plus appropriées par rapport aux objectifs poursuivis par ma recherche. D'habitude, les fichiers XML-TEI de cet organisme contiennent seulement le contenu textuel de l'OCR. Les coordonnées graphiques des mots sont stockées dans des fichiers à part.

[94] Ce qui peut être traduit en français par : « langage de balisage extensible ».
[95] Ce fichier correspond à l'article accessible à l'adresse suivante : https://www.persee.fr/doc/geo_0003-4010_1966_num_75_409_17238 consulté le 18/09/2023.

Par rapport au travail d'amélioration des contenus textuels réalisé ensuite (*cf.* Chap5), un point important à souligner est que les informations de documentation infrapaginaire sont présentes à la fin de chaque page (balise <tei:pb>). Les corrections réalisées ne sont pas directement intégrées dans le texte issu de l'OCR. Par conséquent, pour bénéficier d'un texte avec les corrections réalisées, il faut réintégrer ces corrections à leurs places respectives. Si ces données en XML-TEI ont été celles qui ont été principalement utilisées pour cette recherche, il faut mentionner l'emploi plus ponctuel de trois autres types de données.

3. Trois autres types de données utilisées

Un premier format de données utilisé en complément de celles en XML-TEI provient de l'équipe *Érudit* du Centre d'édition de l'Université de Montréal[96]. L'utilisation de ce format s'explique par le fait que plusieurs informations, comme des coordonnées graphiques de fenêtres tracées lors de la phase de documentation infrapaginaire, existent dans le format Érudit alors qu'elles sont absentes du format XML-TEI obtenu. Ces informations, étant utiles pour le travail d'amélioration des contenus textuels (*cf.* Chap5), ont été importées à partir de ces fichiers. Le lien suivant permet d'accéder au même article que précédemment dans ce format Érudit : **@Afficher exemple Erudit**

Une seule information, les coordonnées graphiques correspondant aux mots-clés, a été identifiée comme étant absente aussi bien du format XML-TEI que du format Érudit. Une demande a par conséquent été effectuée à l'UAR *Persée* pour récupérer ces données. Ces dernières ont été obtenues sous la forme de deux tableaux (un correspondant à *L'Espace Géographique*, un correspondant aux *Annales de Géographie*) qui proviennent d'une extraction des bases de données de l'UAR *Persée*. Ces deux tableaux sont téléchargeables au format CSV en cliquant sur les deux liens en note de bas de page[97, 98]. Chaque ligne correspond à une zone de mot-clé. Ce sont les colonnes D, E, F et G qui ont été utilisées, car elles référencent les coordonnées graphiques de toutes ces zones.

Enfin, une dernière source de données a été utilisée en récupérant ce qui a été mis en ligne par l'UAR *Persée* sur son triplestore[99]. Il s'agit des métadonnées de tous les articles et les

[96] https://apropos.erudit.org/fr/ consulté le 18/09/2023.
[97] **@Lien téléchargement données mots-clés Annales de Géographie en format CSV**
[98] **@Lien téléchargement données mots-clés Espace Géographique en format CSV**
[99] http://data.persee.fr/ressources/le-triplestore-de-persee/ consulté le 18/09/2023.

auteurs des deux revues concernées par cette recherche sous forme de triplet RDF[100]. Ces fichiers sont téléchargeables en cliquant sur les quatre liens en note de bas de page[101,102,103,104]. Dans notre cas, les informations issues de ces fichiers ont servi à remplir une base de données utilisée pour améliorer la qualité du texte des articles (*cf.* Chap5). Il était en effet assez logique de vouloir relier dans cette base de données les documents à leurs auteurs. Sur ce sujet, un travail spécifique a été réalisé par l'UAR *Persée* avec l'*Agence Bibliographique de l'Enseignement Supérieur (ABES)* pour mieux identifier les auteurs[105]. Il aurait été dommage de ne pas inclure cette amélioration dans l'identification des auteurs au sein de cette recherche. Ceci explique l'utilisation de ces fichiers.

Le processus de réduction du corpus qui suit provient de la constatation que les revues ne sont pas qu'une succession bien ordonnée d'articles. Il existe des rubriques (bibliographies, tables des documents publiés, nécrologies…) qu'il n'est pas pertinent d'intégrer aux analyses envisagées. Il y a eu par conséquent toute une réflexion menée autour d'une délimitation plus précise du corpus.

II. Principes et première modélisation pour délimiter les corpus

Cette opération de délimitation s'inscrit dans un processus connu en analyse textuelle. Bénédicte Pincemin parle à ce propos d'« emboîtements » entre plusieurs corpus : « le corpus existant », « le corpus de référence », « le corpus d'étude » et « les corpus distingués » ci-dessous . Par exemple, dans cette recherche, le corpus existant peut être considéré comme composé de toutes les revues de géographie française disponibles sous forme numérique. Le corpus de référence peut renvoyer aux deux revues, les *Annales de Géographie* et *L'Espace Géographique*, choisies en fonction de la problématique retenue. Le corpus d'étude est composé par l'ensemble résultant de cette opération de délimitation plus précise, mais aussi du processus ultérieur visant à améliorer la qualité du texte des articles (*cf.* Chap5). Enfin, les corpus distingués dépendront des analyses menées ensuite avec la délimitation de sous-corpus à comparer dans le temps.

---

[100] Un triplet RDF est une association (sujet, prédicat, objet). Un triplestore est une base de données composée d'un ensemble de triplets RDF.

[101] **@Lien téléchargement données Documents Annales de Géographie en format RDF**

[102] **@Lien téléchargement données Documents Espace Géographique en format RDF**

[103] **@Lien téléchargement données Auteurs Annales de Géographie en format RDF**

[104] **@Lien téléchargement données Auteurs Espace Géographique en format RDF**

[105] Un identifiant unique peut être récupéré facilement pour chaque auteur grâce au travail réalisé par l'UAR *Persée* avec l'Agence Bibliographique de l'Enseignement Supérieur (ABES) : http://info.persee.fr/lalignement-des-autorites-persee-au-referentiel-idref/ consulté le 18/09/2023.

La problématique de cette étape de travail est donc d'identifier les critères les plus pertinents pour construire un ou plusieurs corpus d'étude. Dans cette optique, il est possible d'énoncer les critères retenus avec quelques justifications rapides, mais la démarche ici adoptée est toute autre. Cette dernière vise à accompagner le lecteur dans l'exploration du corpus de référence pour déterminer ces critères et à rentrer dans les questionnements liés à ces choix. De plus, une complexité a été ajoutée en voulant laisser la possibilité de construire des corpus d'étude différents à ceux qui manifesteraient des envies d'expérimentation avec d'autres choix que les critères ici retenus. Cette idée a motivé la création d'une base de données en amont qui recense, organise et stocke tous les documents du corpus de référence. Un corpus d'étude n'est alors qu'une sous-partie de cet ensemble. La base de données pour gérer le corpus de référence et les corpus d'études est modélisée d'un point de vue conceptuel selon le diagramme UML[106]de classes suivant :

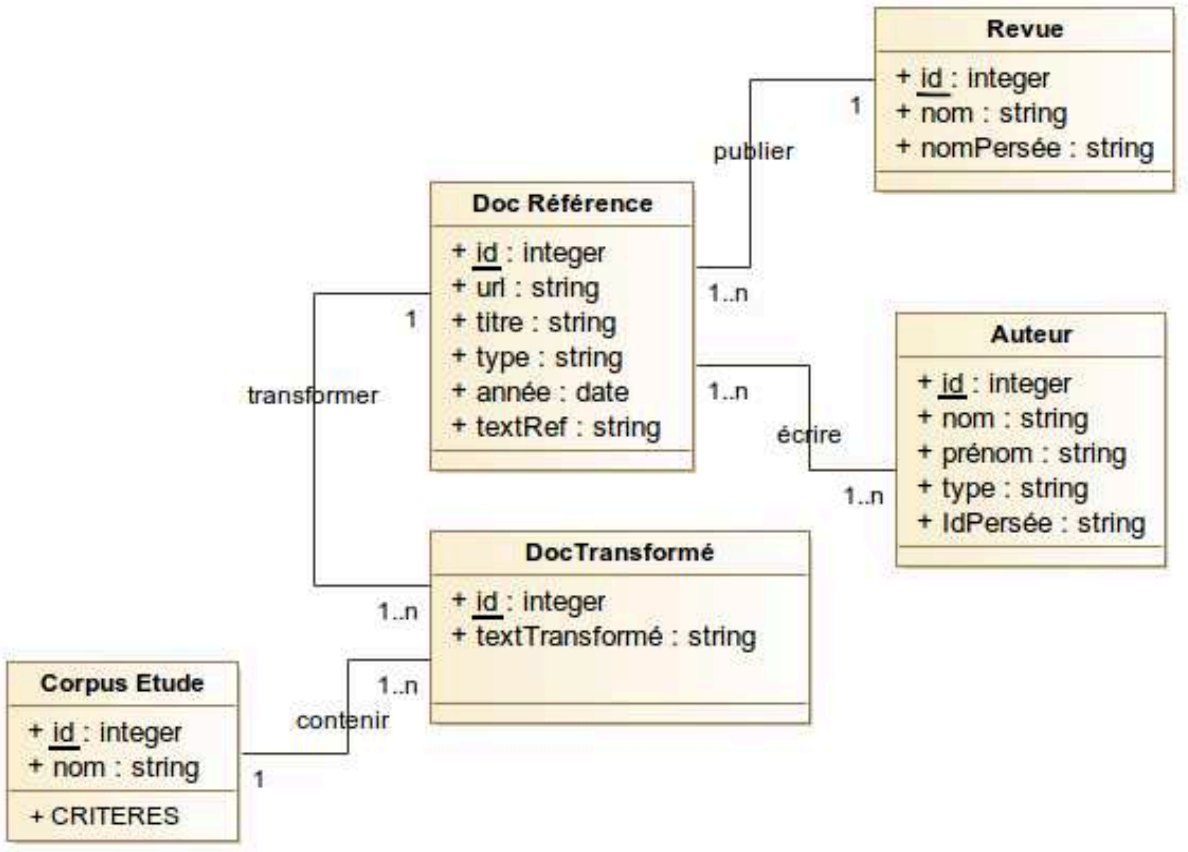

Figure n°9 : Première modélisation conceptuelle de la base de données pour délimiter les corpus d'étude.

[106] *Unified Modeling Language* (Langage de Modélisation Unifié en français).

Cette figure comporte 5 classes : « Doc Référence », « Revue », « Auteur », « Doc Transformé » et « Corpus Etude ». La classe « Doc Référence » modélise tous les documents du corpus de référence où chaque document est décrit par un identifiant (« id »), une url, un titre, un type (article, compte-rendu…), l'année de publication du document et son texte de référence (l'OCR et les informations de documentation de l'UAR *Persée*). Les indications en face de chacun de ces attributs correspondent au format attendu : « integer » pour un entier, « string » pour une chaîne de caractères. Les textes de références, comme d'ailleurs les textes transformés pouvant être assez longs, ils ne sont pas stockés directement dans la base de données. Ce sont les noms des fichiers correspondants qui sont enregistrés. L'attribut « nomPersée » de la classe « Revue » renvoie aux deux abréviations[107] : « geo » pour les *Annales de géographie* et « spgeo » pour *L'Espace Géographique.* L'attribut « type » de la classe « Auteur » provient de la distinction entre auteur individuel et auteur collectif. L'attribut « IdPersée » renvoie au travail précédemment mentionné réalisé par l'UAR *Persée* et l'ABES afin de mieux identifier les auteurs (*cf.* section Chap4.I.3).

Pour la classe « Corpus Etude », la mention « CRITERES » ne correspond pas directement à un attribut, mais à un ensemble d'attributs que le travail qui suit doit définir. Par exemple, si un utilisateur veut créer un corpus d'étude visant à analyser tous les documents du corpus de référence de 1910 à 1980, la création de deux nouveaux attributs, « DateMin » et « DateMax », dans la classe « Corpus Etude » est nécessaire en amont. Pour le corpus d'étude de cet utilisateur, les deux critères « DateMin » et « DateMax » prennent alors respectivement comme valeur : 1910 et 1980. Au-delà de cet exemple basique, ce système offre l'avantage de pouvoir retrouver facilement[108] l'ensemble des critères utilisés pour construire chaque corpus d'étude. De plus, il laisse la possibilité de construire de nouveaux corpus en choisissant des valeurs différentes pour les critères définis.

L'objectif est maintenant d'explorer le corpus de référence pour déterminer les critères retenus et expliquer les choix réalisés.

III. Explorations du corpus de référence pour l'établissement de critères de délimitation des corpus d'étude

L'application Web permet d'accéder à plusieurs modules d'exploration. Chaque module a un objectif spécifique résumé par son titre et correspondant à une fonction facilement accessible dans le code. Si le résultat d'une exploration a déjà été créé, il est immédiatement

[107] Abréviations utilisées à la base par l'UAR *Persée* et reprises dans le cadre de cette thèse.
[108] Il suffit de requêter la base de données où ces critères sont stockés pour chaque corpus d'étude.

affiché. Si ce résultat n'a pas été créé, alors le code pour le construire en question est exécuté. Ce système permet à l'utilisateur d'afficher rapidement les résultats obtenus tout en ayant la possibilité de reprendre le code qui a permis de les créer[109]. La première exploration proposée réalise un affichage dans l'ordre chronologique de tous les numéros de la revue sélectionnée.

1. Afficher les numéros par ordre chronologique

L'UAR *Persée* classant ses documents à l'intérieur de chaque revue par numéro, il s'agit de la première information à laquelle est confrontée une personne consultant ces données. La liste des numéros obtenus n'est pas si longue : 632 pour les *Annales de Géographie* (**@AnnNumeros**) et 113 pour *L'Espace Géographique* (**@EspNumeros**). Ces résultats bruts permettent de détecter facilement une syntaxe commune. Elle est ici décryptée à partir du premier numéro de *L'Espace Géographique* : « spgeo_0003-4010_1972_num_1_1 ». L'abréviation « spgeo » renvoie à la revue[110]. L'indication «0003-4010 » est un numéro de version interne à l'UAR *Persée*. Cette indication peut paraître un détail sans importance, mais cela renvoie au fait qu'il a existé auparavant d'autres versions. Il en existera également dans le futur de nouvelles, du fait des modifications et des améliorations à venir apportées par l'UAR *Persée*. Ces données ne sont pas donc immuables, mais en évolution. Le travail de cette thèse a été effectué sur cette version «0003-4010 » des données.

Ensuite, « 1972 » est la date de ce premier numéro de *L'Espace Géographique*. Puis, « num » est un code servant à spécifier que c'est un numéro « normal ». Un examen rapide des résultats obtenus permet de remarquer qu'il existe une exception en 1993 avec un numéro « hos ». Enfin, la mention « 1_1 » renvoie au fait que c'est la première année de publication et le premier numéro de cette année. Il est facile de détecter à partir de l'ensemble des résultats des exceptions et de comprendre à quoi elles correspondent en explorant les numéros concernés.

Ainsi, le numéro « hos » pour *L'Espace Géographique* renvoie à un hors-série tout en anglais publié en 1993. Pour les *Annales de Géographie*, il faut souligner que de nombreux numéros surtout dans les premières années sont marqués « bib ». Cette situation s'explique par le fait qu'historiquement la *Bibliographie Géographique Internationale* était une publication intégrée dans les *Annales de Géographie*. Pour cette recherche, il a été décidé

[109] Les fonctions de chaque module sont situées dans le fichier « view.py » du dossier « DelimitCorpus » faisant partie de : « AllApps/PreTraitement/Persee».

[110] « geo » pour les *Annales de Géographie*, « spgeo » pour *L'Espace Géographique* dans la codification de l'UAR *Persée*, qui est également reprise dans l'application créée.

d'exclure les numéros « hos » et « bib » des corpus d'étude. En effet, pour le numéro « hos », cela pose la question du multilinguisme qui n'est pas une problématique prise en charge par cette étude. Pour les numéros « bib », les bibliographies ne sont pas des éléments les plus pertinents à prendre en compte pour faire des analyses textuelles et mettre à jour un changement sémantique.

De plus, une différence importante peut être observée entre les deux revues : les numéros des *Annales de Géographie* vont jusqu'en 2006 alors que les numéros de *L'Espace Géographique* s'arrêtent à la fin des années 2000. Cette différence provient du fait que *L'Espace Géographique* a été numérisée après les *Annales de Géographie* et que ces numéros sont disponibles sur le portail *Cairn* depuis 2001. En raison de cette asymétrie dans les données, la partie 2001-2006 des *Annales de Géographie* n'a pas été prise en compte dans les corpus d'étude. Cette première fonction aboutit donc à mettre à jour deux critères importants dans la délimitation des corpus d'étude : le premier s'appuie sur la composition de la dénomination des numéros et le deuxième sur la date maximale prise en compte.

Le deuxième module proposé réalise un affichage non plus des numéros, mais de tous les documents du corpus de référence.

2. Afficher tous les documents par ordre chronologique

Les listes de documents obtenues sont relativement conséquentes : 14 422 lignes pour les *Annales de Géographie* (**@AnnDocuments**) et 2217 lignes pour *L'Espace Géographique* (**@EspDocuments**). Comme dans le cas précédent, une syntaxe commune se détache nettement, mais aussi des exceptions. L'idée est d'examiner ces exceptions sans toutefois être perdu face à la longueur de ces listes. Ceci est l'objectif de la troisième fonction appelée « Afficher les noms des fichiers hors normes ».

3. Afficher les documents ayant un nom « hors norme »

La stratégie retenue pour cette fonction a été d'écrire une expression régulière résumant la forme commune[111] et permettant d'extraire seulement les exceptions. Tous les fichiers qui ont un nom ayant une syntaxe ne respectant pas l'expression régulière sont affichés. L'ouverture des résultats créés par cette troisième fonction peut laisser croire qu'il y a encore

[111] Voici l'expression régulière utilisée :"article_" + revue + "_[0-9]{4}-[0-9]{4}_[0-9]{4}_num_[0-9]{1,4}_[0-9]{1,4}_[0-9]{1,8}_tei\.xml". Revue est une variable : soit "geo" pour les *Annales de Géographie*, soit "spgeo" pour *L'Espace Géographique*. Pour les non-connaisseurs des expressions régulières, une syntaxe du type [0-9]{4} signifie qu'une suite de quatre chiffres entre 0 et 9 est attendue.

de trop nombreuses exceptions pour être traitées facilement par un être humain : 3226 pour les *Annales de Géographie* (**@AnnDocHorsNorme**) et 172 pour *L'Espace Géographique* (**@EspDocHorsNorme**). Or, les structures très répétitives permettent de faire de premières analyses.

Il y a deux types d'exceptions : des documents qui commencent par "article" et d'autres qui commencent par "corpus". Ces derniers correspondent à des recensions correspondant à chaque fois au contenu d'un numéro. Ils ne sont pas intéressants à traiter dans le cadre des analyses ici réalisées. Les exceptions "article" correspondent quant à elles, à des documents dont la forme ou le contenu est particulier. Par exemple, des tables décennales ou encore des comptes-rendus commençant ou finissant au milieu d'une page. Il est possible de vouloir faire un tri à ce niveau en voulant par exemple ne retenir que ce qui correspond vraiment à des articles dans ces fichiers. En effet, ils commencent tous par la dénomination « article », mais cela recouvre des situations très différentes. Ceci est l'objet d'une quatrième exploration (« Afficher les différents types des documents nommés "article" ») qui vise à connaître le contenu réel de ces fichiers.

4. Afficher les différents types des documents nommés "article"

La réalisation de cette fonction est possible grâce aux informations ajoutées par les annotateurs de l'UAR *Persée* dans les métadonnées associées à chaque document. Les résultats obtenus[112] montrent la diversité des documents : article, illustration, compte-rendu, ouvrage reçu, note critique, note biblio, liminaire, table, édito, données, note biographique… Pour des raisons d'homogénéité, une première réflexion réalisée pour créer un corpus d'étude dans la perspective de cette recherche a été de sélectionner que les articles. Ceci a conduit à créer un critère pour sélectionner les documents composant les futurs corpus d'étude en fonction de cette métadonnée renseignée par l'UAR *Persée*.

Si précédemment le hors-série de *L'Espace Géographique* en anglais a été retiré, il est possible que des articles en langue étrangère existent, sans que cela ait donné naissance à un hors-série. La cinquième exploration permet de donner un aperçu des langues présentes pour chaque revue au niveau des articles.

[112] Les résultats obtenus peuvent être consultés avec les liens suivants (**@AnnTypeDocArt** et **@EspTypeDocArt**), mais ils n'apportent pas vraiment d'information plus intéressante que ce qui vient d'être décrit.

5. Afficher les langues des articles

Ce qui est récupéré par cette fonction est l'information donnée sur ce sujet par l'UAR *Persée* dans les métadonnées des articles. Les résultats obtenus sont les mêmes pour *L'Espace Géographique* et les *Annales de Géographie* : français, anglais et allemand[113]. Dans le cas des corpus d'étude créés pour cette thèse, seuls les articles en français ont été retenus pour éviter le problème de gestion du multilinguisme. Cela a nécessité de mettre en place un critère basé sur cette métadonnée « Langue » pour délimiter les corpus d'étude.

Par ailleurs, le travail réalisé a permis aussi de constater que plusieurs articles n'avaient pas de contenu textuel. Une sixième exploration a été menée pour mieux préciser ce phénomène.

6. Afficher les articles vides

Il s'agit d'un nombre très réduit de documents : 4 pour les *Annales de Géographie* (**@AnnArtVides**) et 13 pour *L'Espace Géographique* (**@EspArtVides**). Ce phénomène s'explique par deux raisons différentes. Pour *L'Espace Géographique*, il s'agit d'un problème de droits d'auteurs. Suite à la constatation que parmi ces documents vides, il existait des articles écrits par des géographes importants dans le système interprétatif d'Olivier Orain (comme Jean-Bernard Racine et Antoine S. Bailly), j'ai décidé de numériser, d'océriser et d'importer ces données. Pour les *Annales de Géographie,* les articles manquants concernent un petit document de 1903 (une page) et plusieurs documents de 2002. Sur ces derniers, il s'agit en fait, en amont, d'une modification de l'UAR *Persée* dans la structure de la TEI[114] qui explique que l'algorithme construit pour trouver les mots ne fonctionne pas. Comme les articles postérieurs aux années 2000 n'ont pas été retenus, il n'y a pas eu de traitement de ces articles manquants.

En explorant les documents, il s'est avéré que la catégorie "article" définie par l'UAR *Persée* est moins homogène que prévu. Ceci s'explique par plusieurs raisons. Tout d'abord, il existe des formats d'articles très différents entre les petits essais d'une page et de grands développements sur plusieurs dizaines de pages. Ensuite, les limites entre catégories peuvent parfois être floues. Il n'y a pas par exemple de critères définitifs pour différencier un article, un compte-rendu et une note critique. Enfin, il peut y avoir des erreurs de classification. Pour

[113] Comme précédemment, les résultats obtenus peuvent être consultés avec les liens suivants (**@AnnLangArt** et **@EspLangArt**), mais ils n'apportent pas vraiment d'information plus intéressante que ce qui vient d'être décrit.

[114] Pour une raison inconnue et qui relève très probablement d'une erreur.

mieux cerner ce problème, une exploration a été menée par l'intermédiaire d'une fonction capable d'afficher les catégories dans lesquelles les revues ont, elles-mêmes, classé les articles.

7. Afficher les catégories des articles

Les catégories récupérées par le module d'exploration sont issues des sommaires qui proposent souvent des regroupements d'articles sous des thématiques. L'UAR *Persée* ayant consigné cette information dans les métadonnées, il est possible d'y avoir facilement accès. Un tableau a été créé avec pour chaque catégorie le nombre d'articles concernés, le nombre moyen de pages ainsi que le nombre médian. Les premiers résultats obtenus (**@AnnCatArtBrut** pour les *Annales de géographie* et **@EspCatArtBrut** pour *L'Espace Géographique)* montrent quelques catégories avec de multiples variantes orthographiques[115]. C'est pourquoi les catégories dont les titres étaient proches ont été agrégées[116]. Les résultats nouvellement obtenus peuvent être facilement consultés pour les *Annales de géographie* (**@AnnCatArtAgreg**) et *L'Espace Géographique* (**@EspCatArtAgreg**).

Il faut souligner que les catégories qui peuvent être ainsi interprétées sont les catégories mentionnées par les revues et non pas les catégories réelles[117]. Il y a une grande dissymétrie entre les deux revues. Pour les *Annales de Géographie,* le classement a longtemps été dominé par la bipartition « Géographie générale / Géographie régionale » et s'est ensuite transformé avec une grande catégorie « Article ». Pour L*'Espace Géographique*, il y a eu une multiplication des thématiques. Cette opposition n'est pas développée ici, car elle est déjà connue dans l'histoire de la géographie (Claval, 1998). Dans l'optique de cette thèse, cette exploration a plutôt été utilisée pour améliorer le périmètre des documents retenus pour le corpus d'étude. En effet, on peut remarquer que des documents classés par les revues en « compte-rendu » ont été *a posteriori* rangés par l'UAR *Persée* dans la catégorie « article ». De plus, certaines catégories comme « Dans l'air du temps » laissent davantage penser à des essais qu'à des articles. Un critère de sélection a par conséquent été ajouté de telle sorte

[115] Par exemple, « I – Géographie générale » et « II – Géographie générale » sont considérées comme deux catégories différentes. Il suffit d'un caractère, parfois difficilement visible comme un espace, pour casser l'identité entre deux chaînes de caractères. Cela explique la présence de catégories dans les résultats qui semblent des doublons, mais qui diffèrent d'un point de vue d'une stricte égalité informatique.

[116] L'agrégation s'est faite en utilisant la distance de Racliff/Obershelp (https://en.wikipedia.org/wiki/Gestalt_Pattern_Matching consulté le 18/09/2023) et le seuil de 0,9.

[117] Par exemple, pour les *Annales de Géographie*, la catégorie « Amérique » ne contient qu'un article alors qu'il y a évidemment plus d'un article sur les Amériques dans cette revue pour la période 1892-2006.

qu'un utilisateur puisse exclure certaines catégories de son choix dans la construction d'un corpus d'étude.

Conjointement, le choix a été fait dans un premier temps de garder seulement les articles de plus de trois pages. En effet, beaucoup de catégories relevant plutôt de l'essai que de l'article présentaient des moyennes et des médianes inférieures ou égales à trois pages. Après la première rencontre avec Olivier Orain, ce seuil a été quelque peu remis en question. Il m'a en effet fait remarquer qu'il y avait plusieurs numéros de *L'Espace Géographique* correspondant à l'émergence de thèmes importants (par exemple, celui de la justice spatiale avec David Harvey, vol 7, n°4, 1978) qui étaient en fait des compilations de textes de deux ou trois pages. Par ailleurs, une page ne signifie pas exactement la même chose pour *L'Espace Géographique* et les *Annales de Géographie.* Suite à ces remarques, ce seuil des trois pages n'a pas été gardé. Une possibilité de choisir un seuil par nombre de mots et non plus par nombre de pages, a été implémentée.

8. Synthèse et choix des critères

Ces différentes explorations permettent de préciser et de comprendre plusieurs attributs de la classe « Corpus Etude » précédemment résumés sous la mention « CRITERES » (*cf.* Figure n°9) :

- « Stop word dossier » renvoie aux conclusions de la première exploration en permettant d'exclure du corpus d'étude tous les documents d'un dossier contenant certains mots spécifiés par l'utilisateur (comme « bib », « hos »…).
- « Type doc » permet de sélectionner les documents en fonction de leur nature en se référant à la métadonnée créée par l'UAR *Persée* à ce sujet (« article », « compte-rendu »…).
- « Date min » et « Date max » définissent l'intervalle temporel retenu.
- « Langue » définit les langues des documents formant le corpus d'étude.
- « Mot min » et « Page min » définissent des seuils minimaux pour le nombre de pages et le nombre de mots.
- « Type catégorie exclude » permet d'exclure certaines catégories d'articles en s'appuyant sur les données obtenues lors de la dernière exploration.

Cinq autres attributs ont été ajoutés pour les raisons suivantes :

- « User restrict » permet d'affecter un corpus d'étude particulier à un utilisateur préalablement inscrit. Ce système permet d'imaginer la création de nouveaux corpus sans perturber une base stable nécessaire à la présentation de ce travail.

- « Revue » est un paramètre qui anticipe le fait que des utilisateurs puissent ne vouloir travailler que sur une seule revue.
- « Date » n'est pas un critère de sélection en soi, mais c'est une information pour garder en mémoire la date de créations des différents corpus d'étude.
- « Comment » n'est pas plus un critère de sélection à proprement parler, mais cet attribut permet d'ajouter un commentaire de description pour chaque corpus d'étude.
- « Extract Mot » permet de définir la méthode pour extraire les mots des textes. Dans ce que j'ai réalisé, une seule option est disponible : « FromTEIDocMot ». Cette option signifie que les mots sont extraits des fichiers *Persée* XML-TEI précédemment présentés (*cf.* section Chap4.I.2). Même si aucune autre option n'est disponible, j'ai ajouté cet attribut pour signifier que d'autres méthodes d'extraction des mots sont techniquement possibles. En utilisant par exemple directement les fichiers d'OCR de l'UAR *Persée*, quelques problèmes rencontrés notamment au moment du traitement des bibliographies (*cf.* section Chap5.IV.8) auraient pu être évités. Pour des raisons de temps et de priorités, cette solution n'a pas été implémentée, mais cet attribut permet de ne pas oublier cette réflexion et cette piste de travail.

Grâce à cette modélisation, il est possible maintenant de délimiter concrètement un ou plusieurs corpus d'étude.

IV. Délimitation des corpus d'étude

1. Du modèle conceptuel à la délimitation effective

Le modèle conceptuel précédemment exposé a été décliné en un modèle logique de type relationnel et le modèle physique a été implémenté dans le code du projet *Django*[118]. Une page dans l'application (**@ImportDonnees**) permet de remplir les tables « Auteurs », « Revues » et « Doc Référence ». Si l'utilisateur clique sur ce bouton, une indication « déjà fait » apparaît. En effet, le corpus de référence étant toujours le même, cette étape a déjà été réalisée en amont. S'il est légitime de s'interroger alors sur l'utilité d'un tel bouton, il faut souligner qu'à partir de ce dernier, un lecteur intéressé peut facilement remonter à la

[118] Un projet *Django* permet de créer directement une base de données. Le code correspondant à cette création est situé dans le fichier AllApps/PreTraitement/Persee/AmelioText/models.py. La base de données générée à partir de ce fichier est hébergée par l'infrastructure de recherche *Huma-Num* dans *PostgreSQL*.

fonction[119] qui a permis de constituer le corpus de référence dans la base de données. Il s'agit par conséquent d'un élément, parmi d'autres, visant à améliorer la reproductibilité de cette recherche.

Un formulaire a été ensuite créé à partir de la table « Corpus Etude » pour permettre de créer et d'enregistrer de nouveaux items dans cette table. Son accès dans l'application n'est pas ouvert à tous pour les raisons précédemment explicitées (*cf.* section IntroPart1). Le formulaire se présente ainsi :

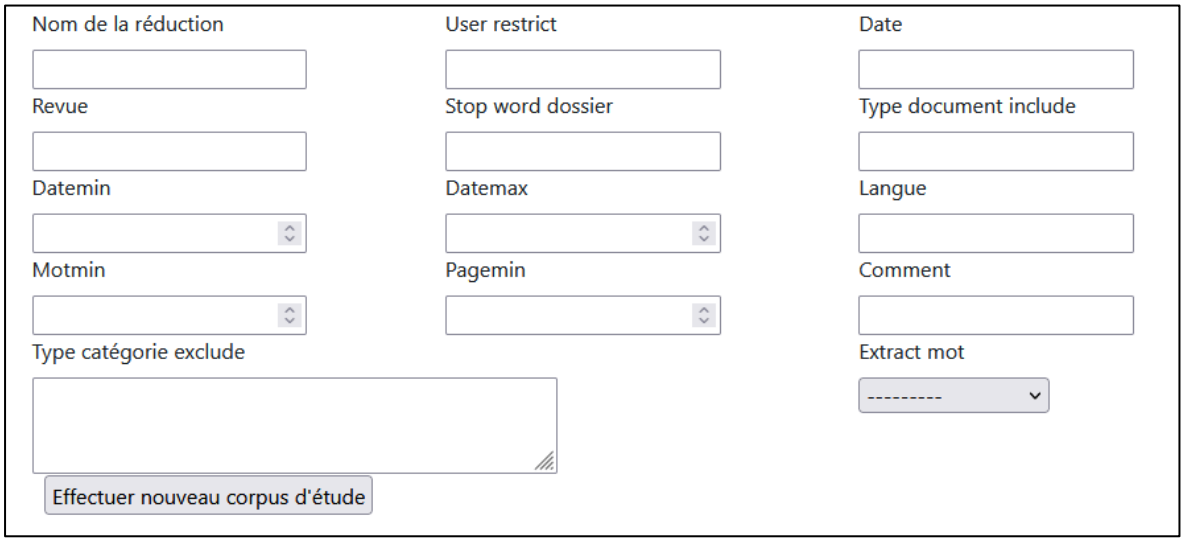

Figure n°10 : Formulaire de l'application pour créer et enregistrer de nouveaux corpus d'étude dans la base de données.

L'enregistrement d'un nouveau corpus d'étude dans la base de données à partir de ce formulaire s'accompagne de l'effectuation d'une fonction[120] qui détermine quels documents du corpus de référence sont concernés en fonction des critères choisis. Ce dispositif technique étant en place, les corpus d'étude utilisés pour cette thèse doivent être plus précisément définis.

[119] La fonction « InsertMesDonneesPersee » dans le fichier « views.py » du dossier « AllApps/Pretraitement/Persee/DelimitCorpus ». D'un point de vue technique, la fonction en amont est construite ainsi : « Si les deux revues utilisées pour cette recherche ne sont pas dans la base de données, alors la fonction qui réalise toute l'insertion des données initiales est effectuée. Sinon, l'indication 'Insertion déjà réalisée' est marquée ».

[120] Voir la fonction « reduction » dans le fichier « views.py » du dossier « DelimitCorpus » faisant partie de « AllApps/PreTraitement/Persee ».

2. Vers la création de deux corpus d'étude

Lors de la création des premiers corpus, plusieurs décisions ont découlé de la problématique et des explorations précédemment menées : revues concernées, mots excluant certains dossiers, date minimum, date maximum… Un critère a suscité davantage de questionnements : le nombre de mots minimal. J'ai choisi au fil du temps un seuil de 1000 mots minimum pour ces premiers corpus, mais cette limite n'a pas été strictement déterminée par des calculs. Ce seuil résulte plutôt d'une reconnaissance qu'il n'existe pas de seuil parfait en la matière. Il faut reconnaître une certaine fragilité et une certaine contingence dans ce découpage. Il aurait été possible d'abandonner ce critère, de choisir tous les documents et de passer sous silence ces questionnements. Il n'en demeure pas moins qu'il existe une certaine hétérogénéité entre des petits documents qui sont rarement des articles et des documents plus longs. Dans le même temps, il est possible de penser que des variations de cette limite ne provoquent que l'ajout ou le retrait de petits textes sans vraiment d'impact sur les résultats du fait même de la masse de données textuelles. Ceci est bien sûr une hypothèse qu'il conviendrait de tester. Toutefois, il m'est apparu plus important de mettre en avant une autre décision qui est plus impactante au niveau de la constitution des corpus d'étude.

En effet, la réflexion initiale avait précédemment conduit à privilégier plutôt les articles et à exclure les comptes-rendus pour des raisons d'homogénéité. Il est évident qu'il est possible pour soutenir un tel choix d'arguer que le compte-rendu est un genre textuel à part. Cependant, la délimitation n'est parfois pas si évidente. De plus, les comptes-rendus peuvent permettre d'avoir une petite image de la production qui est passée « par les livres » et qui est extrêmement importante en SHS. Cette décision est loin d'être secondaire, car elle ajoute une masse textuelle importante[121]. Il peut y avoir certes des comptes-rendus qui ne reflètent pas le contenu scientifique d'une revue, mais globalement le choix de « ce dont une revue rend compte » peut être aussi considéré comme intimement lié à un positionnement intellectuel.

Pour cette raison, il a été décidé finalement de construire deux corpus d'étude : un premier ne prenant en compte que les articles en français avec le seuil minimum de 1000 mots, un second qui inclut aussi les comptes-rendus en français sans limite pour le nombre minimal de mots[122]. Le premier corpus nommé « Article » est considéré comme le corpus d'étude principal. C'est sur celui-ci que seront présentés les premiers résultats. Le deuxième corpus

[121] Il existe souvent plusieurs comptes-rendus par numéros.

[122] Le choix de ne pas retenir de nombre minimal de mots, ni pour les articles, ni pour les comptes-rendus dans ce second corpus, provient du fait qu'en incluant les comptes-rendus, il y a une recherche d'homogénéité moindre. Il devient moins pertinent de rentrer dans un tel détail.

nommé « ArticlePlusCr » est destiné à réaliser des tests pour voir si cet ajout important des comptes-rendus modifie (ou non) les résultats obtenus.

La section suivante présente la visualisation de ces deux corpus à partir de l'application.

3. Visualisation des corpus dans l'application

Toutes les valeurs entrées dans le formulaire pour la construction des deux corpus peuvent facilement être visualisées grâce à la page suivante : **@FaireEtVisualiserCorpus**

Pour ces deux corpus, le critère « User restrict » est un peu particulier, car ce sont des corpus ouverts au niveau de la lecture des résultats déjà réalisés, mais protégés quant à la production de nouveaux résultats. Ce statut spécifique a été codé par le chiffre « 0 », mais ne renvoie pas à un utilisateur particulier.

Le critère « Langue » est mentionné en anglais, car la syntaxe utilisée par l'UAR *Persée* a été reprise[123].

Les choix réalisés pour le critère (« Type catégorie exclude ») s'expliquent à partir des premiers résultats (non-agrégés) de la septième exploration[124]. Au niveau de la syntaxe utilisée, certaines catégories étant constituées de virgule, le séparateur utilisé a été « * ».

Ces deux corpus d'étude étant constitués, l'étape suivante est celle de l'amélioration de leurs contenus textuels.

[123] Disponible à partir des résultats de la cinquième exploration.
[124] **@AnnCatArtBrut** pour les *Annales de géographie* et **@EspCatArtBrut** pour *L'Espace Géographique.*

Chapitre 5 :
Amélioration des contenus textuels

L'objectif de cette étape est d'améliorer la qualité des textes qui seront ensuite traités quantitativement. Cette étape est fondamentale : si la qualité des données est insuffisante, les résultats obtenus ont de grande chance d'être insatisfaisants quel que soit l'algorithme utilisé. Néanmoins, la notion de qualité appliquée à un corpus textuel n'est pas si facile à définir, car elle revêt de multiples dimensions.

I. Un objectif multi-dimensionnel et problématique

Cette notion peut tout d'abord être conçue dans le sens d'une fidélité par rapport au texte initial. Si l'OCR a mal fonctionné, le texte obtenu comporte des écarts par rapport à la version originale. L'objectif est alors logiquement de corriger ces erreurs qui modifient le texte initial.

Une autre dimension importante de la qualité des contenus textuels est celle de la pertinence des données. Cet aspect a déjà été mentionné lors de l'étape précédente (la délimitation du corpus) pour le choix des textes, mais il joue également un rôle à une échelle plus fine, à l'intérieur même des textes choisis. Par exemple, dans la revue *L'Espace Géographique*, certaines publicités à la fin des articles ont été numérisées et n'ont pas été délimitées par l'UAR *Persée* comme des éléments à part. Elles font alors partie des articles. Comme ces éléments ne relèvent pas du contenu scientifique des articles, il est évident qu'il n'est pas pertinent de les intégrer dans le corpus à analyser. Cet exemple montre que la définition de la qualité dépend intrinsèquement du contexte et de l'objectif poursuivi. En effet, si l'objectif avait été de rendre compte des revues dans leur entièreté, ces publicités auraient eu toute leur place dans le corpus.

Enfin, une autre dimension importante de la qualité, dans la perspective des analyses prévues, concerne le volume de données. D'une manière générale, l'augmentation du nombre de données (si elles sont de qualité en se basant sur les deux points précédents) tend à accroître la significativité des résultats. À travers ces remarques, l'objectif n'est pas de recenser de manière exhaustive toutes les dimensions constitutives de la qualité d'un corpus de textes[125], mais de faire saisir que cette notion générale recouvre une pluralité d'acceptions et d'objectifs.

[125] Nous renvoyons le lecteur intéressé par un panorama plus complet sur ce point à la thèse de Bénédicte Pincemin (1999a), notamment au chapitre VII : « Caractérisation d'un texte dans un corpus : du quantitatif vers le qualitatif », §A « Définir un corpus », p. 415-427. Ce passage est publié à l'adresse suivante : http://www.revue-texto.net/1996-2007/Corpus/Publications/pincemin_ad_1999.pdf consulté le 18/09/2023.

De manière globale, il n'existe pas de seuil à partir duquel il est possible de dire qu'un niveau satisfaisant de qualité a été atteint pour telle analyse. Dans le cas présent, c'est-à-dire une étude d'un changement sémantique, il est vrai que le déplacement d'une virgule peut, par exemple, changer le sens d'une phrase. Toutefois, les méthodes utilisées et les analyses menées n'ont pas la prétention d'atteindre un tel niveau de finesse. Il s'agit d'une étude quantitative qui peut sûrement supporter un certain degré de bruit dans les données initiales. Un travail en profondeur a été réalisé pour améliorer, autant que possible, dans le temps imparti, la qualité des textes utilisés dans cette recherche. Cependant, il reste un nombre d'erreurs non négligeable (*cf.* section Chap6.II). Il y a donc un certain pari dans le fait de penser que ces résultats sont tout de même pertinents.

Ce pari ne repose pas sur une pure spéculation. Une méthode très utilisée en analyse textuelle est le retour dans les textes pour mieux comprendre et interpréter les quantifications, les agrégations et les visualisations que fournissent les différents algorithmes. Dans le cadre de cette thèse, les résultats obtenus peuvent aussi être comparés avec ce qui est connu qualitativement de l'évolution de la géographie française. Toutefois, toutes ces méthodes sont essentiellement qualitatives. Il est certes possible de réaliser des évaluations quantitatives des plongements de mots, mais, comme cela a été précédemment développé (*cf.* section Chap3.IV.2), ce n'est pas alors la capture des traits sémantiques et de leurs évolutions qui sont vraiment évaluées. Enfin, comme il n'y a pas de seuil permettant d'affirmer que les résultats sont valides ou/et pertinents, il existe par conséquent toujours *in fine* une appréciation qualitative qui est susceptible d'être discutée et interrogée.

Dans la pratique, l'amélioration de données textuelles peut être réalisée de diverses façons et la section suivante essaye de mieux caractériser le travail qui a été réalisé sur ce sujet.

II. Un processus d'amélioration semi-automatisé

Quand un corpus est petit, il est la plupart du temps « nettoyé »[126] manuellement. Le chercheur lit alors l'ensemble du corpus et corrige chaque faute au fur et à mesure de sa lecture. Ici, la taille du corpus ne permettait pas une telle opération. Il a donc été nécessaire de définir et d'automatiser différentes opérations. Par rapport à une telle entreprise, il est toujours intéressant de chercher les cas qui ne sont pas traités correctement par les algorithmes de traitement mis en place. Quand un critère pouvait être trouvé pour repérer ces cas traités de manière erronée par l'automatisation et que ces derniers pouvaient être

[126] Les guillemets sont ici utilisés pour signifier un renvoi à l'expression commune de « nettoyage des données » et à une certaine réticence quant à la pertinence de cette expression comme l'explique cette partie.

traités dans un temps raisonnable, des corrections « manuelles » ont été réalisées. L'amélioration des contenus textuels réalisée peut par conséquent être qualifiée de « semi-automatisée », au sens où elle combine des tâches de corrections effectuées en masse à l'aide d'algorithmes et d'autres modifications effectuées au cas par cas.

Il est nécessaire de préciser plus finement cette articulation, car différentes formes de « nettoyage » semi-automatique peuvent exister. Par exemple, il est possible dans un tel cadre de faire une grande partie de l'amélioration à l'aide de procédures automatisées et ensuite de corriger manuellement les erreurs identifiées au fur et à mesure de leur découverte. Il m'importe de préciser ici que ce n'est pas cette articulation entre la partie automatisée et manuelle qui a été adoptée pour cette thèse. Dans une optique de reproductibilité, l'importance des corrections manuelles liées à des consultations erratiques a été réduite au maximum.

Il est bien sûr arrivé qu'après une consultation du corpus, des erreurs soient repérées. Dans ce cas-là, la procédure mise en place est la suivante : il s'agit tout d'abord d'essayer de détecter globalement ce type d'erreur dans l'ensemble du corpus. Si cette détection est réalisable dans le temps disponible et que le gain de qualité induit par rapport à d'autres actions possibles est jugé intéressant, alors l'essai d'une procédure d'automatisation pour repérer et traiter toutes les erreurs de ce type est réalisé. Sinon l'erreur est laissée. Les corrections manuelles n'interviennent qu'en dernière instance. Le plus souvent, c'est après la mise en place d'une procédure pour détecter les cas sur lesquels les traitements automatisés ont échoué, mais aussi également après le constat que l'amélioration de l'automatisation est plus coûteuse en temps que l'effectuation de corrections manuelles.

Dans ce processus, la perception du temps disponible joue un rôle fondamental. En effet, en ayant beaucoup de temps comme au départ d'une thèse, il est possible de rentrer dans les détails et de viser une grande qualité de texte. Il y a eu, au fur et à mesure de l'avancée de mes travaux, l'adoption d'une approche de plus en plus pragmatique pour arriver à terminer cette préparation du corpus tout en conservant du temps pour la suite de la recherche.

Au niveau de la rédaction, l'accumulation des taches réalisées est particulièrement difficile à présenter. Il y a une multiplication des parties qui correspond, en amont, à la division en sous-tâches. Au sein de chaque sous-tâche, il y a eu la volonté d'accompagner le lecteur pour expliquer ce qui a été fait et pourquoi cela a été réalisé. Le prix à payer a été, sans aucun doute, celui d'une longue rédaction. Malgré la conscience de ce désagrément, autant pour la réalisation de ce travail que pour les futurs lecteurs, il y a eu un maintien de cette forme par conviction de rigueur scientifique. Lors du colloque international *Histoire,*

langues et textométrie[127], j'ai ainsi plaidé pour que les travaux d'analyse textuelle présentent beaucoup plus leur travail de préparation des données par rapport à ce qui est fait actuellement dans la majorité des cas (Beligné, Loudcher et Lefort 2023a). Alors que cette étape représente un temps de travail important[128] et qu'elle est souvent loin d'être neutre, elle se trouve souvent résumée au mieux en un paragraphe, au pire en deux ou trois lignes pour laisser place à l'explicitation de la méthode et aux résultats.

La solution présentée à ce colloque est ici reprise. Il s'agit d'intégrer à cette phase de « nettoyage » des données souvent très technique une dimension de « critique des sources », étape classique du travail de l'historien (Langlois et Seignobos, 2014, *1897*). Pour cela, il est nécessaire d'accepter de présenter les limites des améliorations effectuées. C'est en effet à l'endroit où cette opération rencontre des difficultés qu'il est souvent possible de développer un discours critique sur les données utilisées. Cette perspective implique de ne pas éluder les difficultés et d'accepter de développer certains points de faiblesse. L'utilisation de cette méthode permet de déconstruire les données acquises. D'une certaine façon, cette optique remet en cause cette métaphore du « nettoyage des données », car il ne s'agit pas seulement d'éliminer les erreurs, mais de prendre aussi en considération ce qu'elles nous apprennent.

Cette orientation méthodologique s'inscrit dans la position défendue par Claire Lemercier pour « une deuxième vague quantitative avec une critique des sources au centre et l'acceptation de données non "nettoyées" » (Lemercier, 2020). Si son refus de l'expression de « nettoyage des données » est ici partagé, sa proposition d'utiliser plutôt « simplification des données » n'a pas été reprise. D'un certain point de vue, un positionnement inverse pourrait être revendiqué puisqu'il y a dans la recherche menée aussi une part d'enrichissement des données avec la création de nouvelles informations (*cf.* sections Chap5.IV.5, 6 7 et 12).

Une autre référence permettant de penser cette orientation méthodologique est une présentation de Pierre-Carl Langlais qui montre comment la « modélisation éditoriale » (au sens de détection des différents éléments constitutifs d'une revue) peut être utilisée pour « historiciser par algorithme » (Langlais, 2019). Ce verbe « historiciser » renvoie à la compréhension des contextes de production des données et de leurs changements. Toutefois, il est vrai que Pierre-Carl Langlais applique surtout cette perspective pour créer et analyser

[127] Colloque organisé par le *Pôle Informatique de Recherche et d'Enseignement en Histoire* (PIREH) en 2019 à la Sorbonne.

[128] Un intervalle entre 50 et 80 % du temps d'une recherche passé à préparer les données est souvent annoncé de manière générale dans le domaine des données massives (Lohr, 2014). Il me semble aussi bien correspondre à la réalité de la plupart des études d'analyses textuelles même quand les ensembles de données travaillés ne sont pas massifs.

directement des résultats, et non comme une phase de prétraitement permettant de construire conjointement une critique des sources. C'est par conséquent une voie complémentaire qui est ici proposée en mobilisant ce processus réflexif dès la préparation des données[129].

L'expression d'« amélioration des contenus textuels » ici privilégiée ne rend malheureusement pas vraiment compte de tout l'effort sous-jacent de critique des sources. Cependant, il est particulièrement délicat de trouver la bonne expression qui permettrait de rendre compte dans le même temps de la dimension très technique et basique (corrections d'erreurs) et de la dimension heuristique et critique (compréhension et recul critique par rapport aux documents). Il ne m'a pas semblé pertinent de nommer le travail réalisé directement comme « critique des sources », car ça a été plus un effet induit et une prise de conscience tardive qui n'a pas forcément respecté les codes de cet exercice.

La section suivante présente la technique générale utilisée pour réaliser ce travail d'amélioration.

III. Des améliorations stockées dans la base de données

Le modèle conceptuel présenté précédemment a été complété en ajoutant des attributs à l'association « Transformer ». Cette association peut être modélisée ainsi :

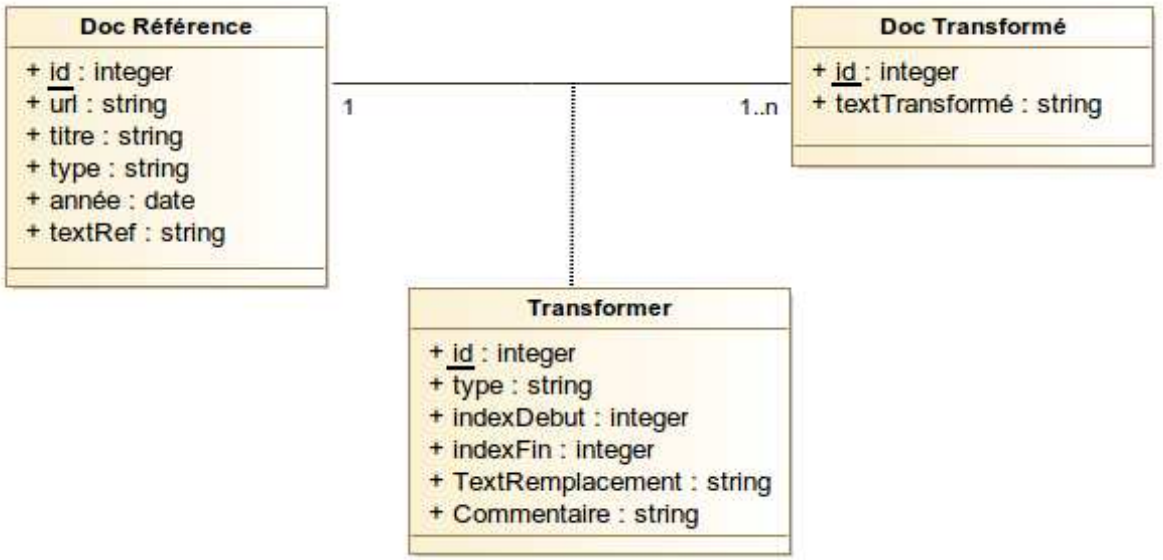

Figure n°11 : Ajout d'une classe-association
pour stocker les différentes améliorations des contenus textuels.

[129] L'historicisation peut en effet découler des traitements statistiques ultérieurs.

Un exemple concret permet de comprendre facilement le fonctionnement physique de la table « Transformer ». Si, par exemple, un document a une faute dans son titre issu de l'OCR, si, également, le titre exact est connu grâce aux corrections de l'UAR *Persée*, et si, enfin, la localisation du titre erroné est connue, alors, un nouvel élément mérite d'être créé dans cette table « Transformer ». Dans ce cas, la correction est de « type » : « titre ». Si la position du titre dans l'OCR brut est sur les quatre premiers mots du document, « indexDebut » est égal à 0 et « indexFin » à 3. Le titre corrigé est noté dans « TextRemplacement »[130]. Une possibilité d'ajouter un commentaire existe, mais n'est pas forcément nécessaire. Ce système permet de stocker toutes les transformations pour chaque document de chaque corpus d'étude. À la fin du processus, les textes issus de l'OCR sont améliorés en prenant en compte toutes les transformations enregistrées. De plus, un utilisateur ayant créé un nouveau corpus d'étude, peut enregistrer ses propres améliorations à apporter aux textes grâce à ce système. La section suivante détaille la création de ces éléments d'amélioration sur les deux corpus d'étude précédemment construits en examinant successivement toutes les parties d'un document.

IV. Améliorations réalisées grâce à l'examen des composantes des documents

1. Organisation générale et remarques introductives

Pour chaque partie (titre, résumé, mot-clé…), trois étapes sont proposées, intitulées respectivement « Explorations » / « Traitements automatisés » / « Traitements manuels ». L'étape « Explorations » vise, au-delà d'une découverte des données, à faire comprendre l'origine des méthodes et des paramètres utilisés pour les traitements automatisés. Les traitements manuels permettent d'insérer des corrections spécifiques si nécessaire dans la démarche précédemment explicitée (*cf.* section Chap5.II). Les parties examinées ont été nombreuses : titres principaux, résumés, mots-clés, hauts de page, bas de page, notes, bibliographies, annexes, titres secondaires, figures et fins de document. De plus, la volonté d'expliciter ce qui a été fait et ce que cela nous apprend sur les documents utilisés explique une certaine longueur de chaque partie.

Une synthèse a été écrite (*cf.* section Chap5.VII) pour résumer les traitements réalisés. Il serait toutefois dommageable que ce travail de synthèse effectué pousse les lecteurs à ne pas

[130] Dans le cadre d'une correction qui vise seulement à supprimer un élément comme une note de bas de page, il suffit de laisser ce champ « TextRemplacement » vide.

lire les sous-parties suivantes. En effet, il existe dans cette confrontation avec les données et dans les solutions trouvées localement pour résoudre de multiples petits problèmes, le rendu d'un travail qui a permis à la fois de lever un verrou scientifique (en permettant de passer de l'acquisition des données à leur utilisation effective) et de réfléchir épistémologiquement (en s'interrogeant notamment sur le positionnement d'un travail quantitatif suivant que cette partie de préparation des données est ou n'est pas développée).

Les premières parties considérées pour ce travail d'amélioration sont les titres. L'enchaînement des autres parties (résumés, mots-clés, hauts de page…) restitue la chronologie du travail effectué.

2. Titres

a. Explorations

L'UAR *Persée* possède le titre principal de chaque document stocké en métadonnée, mais la localisation exacte de chaque titre dans les textes issus de l'OCR n'est pas connue. Cette situation peut paraître étrange, car il a été précédemment affirmé que pendant la phase de documentation infrapaginaire réalisée par l'UAR *Persée* (*cf.* section Chap4.I.1) les coordonnées graphiques des principaux éléments de chaque document sont définies et stockées. Cette remarque oblige à rentrer plus dans le détail de la chaîne de production de cet organisme. Le titre de chaque document a en fait déjà été référencé en amont de la phase de documentation infrapaginaire. En effet, dans la phase de documentation initiale (*cf.* section Chap4.I.1), chaque numéro de revue est découpé à partir de la table des matières en autant de documents qu'il contient et pour chacun le titre et les pages concernées sont spécifiés. C'est pourquoi l'UAR *Persée* n'a pas inclus la localisation du titre comme un objectif de la documentation infrapaginaire. Le contenu textuel de ces éléments, qui est l'information qui intéresse le plus l'UAR *Persée* en vue d'une indexation sur le Web, a déjà été récupéré lors de l'étape précédente. La conséquence de cette division du travail est que les coordonnées graphiques des titres de chaque document n'existent pas dans les données acquises[131].

Par rapport à l'objectif d'amélioration des contenus textuels, le repérage des titres est important pour deux raisons. La première est évidemment de remplacer le titre possiblement bruité de l'OCR. La deuxième est qu'il y a souvent en amont des titres, des indications

[131] L'UAR *Persée* ayant travaillé à une refonte de leur chaîne de production, il est possible que la problématique qui est ici exposée soit prise en compte par cet organisme si bien qu'elle ne sera plus d'actualité dans quelques années.

mentionnant le nom de la revue et son adresse qui méritent d'être enlevées. En localisant précisément le titre, ces indications peuvent ainsi être facilement ôtées par la suite.

Pour débuter ce travail, une première fonction d'exploration a été créée pour afficher tous les titres référencés par l'UAR *Persée* d'une revue faisant partie d'un corpus d'étude. Un point particulièrement frappant dans le cas du corpus « Article » et des *Annales de Géographie* (**@ArticleAnnTitres**) est la présence de titres, pour la période la plus contemporaine, à la fois en français et en anglais, séparés par une ou deux barres obliques. Par exemple, « La Grande-Motte, Ville permanente, ville saisonnière//La Grande-Motte, a town for all seasons and summertime resort ». Cette syntaxe est une astuce trouvée par l'UAR *Persée* pour faire face à une situation qui n'avait sûrement pas été anticipée. La revue des *Annales de Géographie* décide à partir de 1999 d'adopter ce format de titre avec deux langues. Il aurait été en effet préférable de coder cette information avec une structure appropriée dans le schéma XML-TEI utilisé. Même si une telle structure peut être facilement proposée, cette situation renvoie au fait qu'il est impossible de prévoir un schéma anticipant tous les cas. La découverte au fur et à mesure de certaines caractéristiques qui n'avaient pas été prévues amène soit à modifier le schéma, soit à trouver des astuces[132] pour prendre en compte ces spécificités.

Il faut souligner que ces titres « français/anglais » ne sont pas présents dans le cas de la revue *L'Espace Géographique* (**@ArticleEspTitres**). Ils ne concernent pas non plus les comptes-rendus des *Annales de Géographie* comme le montrent les résultats obtenus pour le corpus « ArticlePlusCR » (**@ArticlePlusCrAnnTitres**).

Le choix réalisé pour ces deux corpus a été d'enlever la partie anglaise de ces titres dans les textes destinés à être analysés. Le traitement imaginé consiste à repérer les titres où il y a une ou plusieurs barres obliques. Ensuite, la première partie du titre située avant la première barre oblique est gardée comme titre de référence. Le problème d'un tel traitement est qu'il peut toucher d'autres titres qui possèdent aussi une ou des barres obliques, mais qui ne relèvent pas de la même construction.

Une deuxième fonction d'exploration prend en charge ce problème. Cette dernière permet d'afficher seulement les titres référencés par l'UAR *Persée* qui sont construits avec au moins une barre oblique. Le résultat obtenu (**@ArticleAnnTitreSlash**) pour la revue des *Annales de Géographie* dans le corpus « Article » permet de confirmer que si ce traitement est réservé aux articles de cette revue depuis 1999, il n'affecte pas de manière erronée d'autres titres.

[132] Ces astuces qui relèvent d'une forme de bricolage se justifient par le fait que si le schéma change, il est nécessaire de régénérer toutes les données produites auparavant pour les adapter au nouveau schéma. Il est parfois moins coûteux d'utiliser localement un tel bricolage.

Une troisième fonction permet d'afficher les résultats d'une méthode construite pour identifier les titres dans les textes. Il s'agit d'utiliser une mesure de similarité entre deux chaînes de caractères pour identifier où se situe le plus probablement le titre principal dans les textes issus de l'OCR. Il existe de nombreux calculs différents de la similarité entre deux chaînes de caractères : distance de Levenshstein, de Hammind, de Jaro Winkler, de Jaccard… Dans cette thèse, la distance de Ratcliff/Obershelp a été choisie après une expérimentation[133] qui a donné des résultats assez convaincants. Cette méthode recherche les deux plus longues séquences communes non redondantes entre deux chaînes de caractères. Le nombre de caractères de ces deux séquences communes non redondantes est multiplié par 2 et divisé par le nombre de caractères des deux chaînes. Le résultat est compris entre 0 et 1 : 0 pour deux chaînes de caractères totalement différentes, 1 pour deux chaînes de caractères identiques.

Une des difficultés est que l'OCR a pu couper ou relier des mots. Ainsi, il n'est pas forcément pertinent de rechercher toujours une séquence de même taille que le titre corrigé par l'UAR *Persée*. Il n'était cependant pas envisageable de tester des séquences de toutes les tailles dans tous les textes pour des raisons de temps de calcul. La stratégie choisie a consisté à réaliser un premier calcul avec le même nombre de mots que le titre corrigé dans les 50 premiers mots du texte. Ce seuil s'explique par le fait que dans la majorité des cas, le titre est au début du document. Si le résultat trouvé n'est pas totalement convaincant (moins de 0,95 avec la méthode de Ratcliff/Obershelp sachant qu'avec deux chaînes strictement identiques, le résultat obtenu est 1), des tailles plus petites et plus grandes que la taille de référence ont été testées[134]. Le meilleur résultat est gardé comme le titre le plus probable. Enfin, avant d'afficher les résultats, ces derniers ont été triés dans l'ordre croissant. Ainsi, les moins bons résultats, ceux qui sont le plus susceptibles de contenir des erreurs apparaissent en premier.

Les résultats obtenus sont disponibles dans l'application : pour le corpus « Article », les *Annales de Géographie* (**@ArticleAnnTitreCherche)** et *L'Espace Géographique* (**@ArticleEspTitreCherche)** ; pour le corpus « ArticlePlusCr », les *Annales de Géographie* (**@ArticlePlusCrAnnTitreCherche)** et *L'Espace Géographique* (**@ArticlePlusCrEspTitreCherche)**.

---

[133] La raison expliquant le choix de cette distance relève de la commodité. Elle a été implémentée dans une librairie Python couramment utilisée (difflib). Elle est par conséquent facilement applicable.

[134] Si la taille du titre référencé par l'UAR *Persée* est de n mots, pour les titres de moins de 10 mots, la recherche s'est effectuée sur l'ensemble des chaînes de caractères comprises entre n-3 et n+3 mots sur les 50 premiers mots du texte ; pour les titres de 10 à 15 mots, entre n-5 et n+5 mots ; pour les titres de + de 15 mots, entre n-7 et n+7 mots. Ces seuils ont été fixés en fonction de ce qui a été jugé comme semblant le plus adapté suivant les situations. L'objectif a été d'essayer d'avoir une bonne efficacité tout en gardant des temps de calcul raisonnables.

Les résultats comportent pour chaque document cinq lignes :

- L'identifiant du document donné par l'UAR *Persée* sous forme d'un lien vers leur portail permettant de consulter directement le document en question.
- Le score de ressemblance obtenu, c'est-à-dire la distance de Ratcliff/Obershelp entre le titre documenté par l'UAR *Persée* et la chaîne de caractères dans l'OCR identifiée comme étant la plus proche du titre selon l'algorithme créé.
- La chaîne de caractères précédemment identifiée.
- Le titre documenté par l'UAR *Persée.*
- L'index de début et de fin de la chaîne de caractères présumée être le titre.

Pour simplifier l'affichage des résultats dans l'application, ces derniers ont été limités aux 100 premiers et aux 100 derniers. Il est toutefois possible sur la page de l'application correspondante à chaque corpus et revue de télécharger à chaque fois un fichier avec l'ensemble des résultats.

Au niveau de l'analyse de ces résultats, il est nécessaire de distinguer les deux corpus. Pour le premier, « Article », il apparaît que le nombre d'erreurs est assez réduit et peut être traité à la main.

Ainsi, une quatrième exploration a été réalisée avec des analyses de tous les cas étant les plus susceptibles de comporter des erreurs. Les liens suivants (**@ArticleAnnAnalyseCas** pour les *Annales de Géographie* et **@ArticleEspAnalyseCas** pour *L'Espace Géographique*) permettent de consulter le travail réalisé. Chaque titre est surligné en rouge quand la chaîne de caractères trouvée par la méthode décrite ne correspond pas du tout au titre ; en orange quand la chaîne de caractères trouvée a un rapport avec le titre réel, mais reste erronée en grande partie ; en jaune quand la chaîne de caractères trouvée correspond presque au titre ; enfin en vert quand le titre a effectivement été trouvé. Cette analyse des résultats a été arrêtée quand les résultats ont été jugés comme commençant à être très majoritairement bons avec une suite d'au moins 7 documents ayant des titres correctement reconnus[135].

Ces travaux permettent de montrer qu'après un score de ressemblance de 0,74, les titres commencent à être très bons pour *L'Espace Géographique*. Les erreurs commencent à être moins nombreuses pour les *Annales de Géographie* à partir de 0,77 et ne deviennent vraiment rares qu'à partir de 0,83. Ces résultats différents pour les deux revues s'expliquent par plusieurs raisons. Tout d'abord, une meilleure océrisation[136] de *L'Espace Géographique*

[135] Ce seuil de 7 documents consécutifs est un choix subjectif permettant d'arrêter les analyses quand les résultats sont globalement très bons.

[136] Cette meilleure qualité s'explique autant par les documents de base (la qualité du papier est parfois moins bonne et la typographie moins stable quand on remonte dans le temps) et par les progrès des logiciels d'OCR. Les *Annales de Géographie* ont été traitées par l'UAR *Persée* avant *L'Espace Géogaphique*.

mais aussi des titres plus longs pour cette revue dont la reconnaissance tend à comporter moins d'erreurs[137].

Les résultats précédemment obtenus pour le deuxième corpus « ArticlePlusCr » sont très différents de ceux qui viennent d'être analysés. Les erreurs ne peuvent pas être traitées manuellement, car elles sont trop nombreuses. Une structure assez particulière commune aux deux revues explique une grande partie de ces erreurs. En effet, souvent le titre d'un compte-rendu reprend le titre du livre en question. En note de bas de page, les auteurs et les références exactes du livre sont précisées. Or, quand l'UAR *Persée* a référencé ces titres de comptes-rendus, les auteurs et les références exactes du livre ont été intégrés dans le nouveau titre. La méthode créée cherche ainsi une séquence textuelle qui n'est pas le titre donné originellement par la revue. Dans cette situation, il est assez compréhensible que la séquence trouvée ne corresponde presque jamais au titre réel.

L'analyse de ces résultats a aussi montré que contrairement aux articles, il n'y a pas d'inscription répétée du nom de la revue et de son adresse avant chaque compte-rendu. Ceci réduit l'objectif de la reconnaissance des titres à un simple remplacement par les titres exacts quand ces derniers ont été déformés par l'OCR. Il aurait été certes possible d'améliorer l'algorithme en essayant de mieux prendre en compte les cas des comptes-rendus, mais la détection des titres n'étant pas la finalité de cette thèse, il a été jugé préférable de laisser la méthode en l'état et de ne l'appliquer qu'aux articles. Le fait que le nom de la revue et son adresse ne soient pas répétés avant chaque compte-rendu rend cette décision moins problématique.

b. Traitements automatisés

À partir des explorations menées, une fonctionnalité de traitement automatisé des titres a été créée dans l'application avec 3 options. La première propose le choix d'un seuil à partir duquel les résultats obtenus avec la distance de Ratcliff/Obershelp pour les articles sont retenus. Pour les deux corpus, ce seuil a été fixé à 0,83 du fait du travail effectué précédemment. Si le résultat est supérieur à ce seuil, un enregistrement est créé dans la table « Transformer » correspondant au remplacement de la chaîne de caractères identifiée avec la méthode explicitée par le titre documenté par l'UAR *Persée*. Une deuxième option permet de retirer dans ce cas tout le texte qui précède la chaîne de caractères trouvée. Enfin, une

[137] Si un titre court comme « La Bresse » est mal transcrit par l'OCR et qu'il est répété au début du texte sans erreur de transcription cette fois-ci, cela provoque une erreur de détection. Le risque est moindre avec un titre du type : « La Bresse, étude d'un géosystème complexe ».

troisième option correspond à la possibilité d'enlever la partie anglaise des titres des articles des *Annales de Géographie* à partir de 1999. Ces deux dernières options, retirer la partie anglaise et la partie précédant le titre, ont été évidemment retenues pour les deux corpus.

Ces choix et leurs enregistrements se font par l'intermédiaire d'un formulaire qui peut être visualisé à partir des liens suivants : **@ArticleTraitAutoTitre** et **@ArticlePlusCrTraitAutoTitre** qui synthétisent les traitements réalisés pour les titres respectivement pour les corpus « Article » et « ArticlePlusCr ». Pour un utilisateur externe, les formulaires apparaissent grisés et ne peuvent donner lieu à aucun changement. Si une personne a demandé la création d'un corpus et a donc un droit d'écriture dessus[138], elle peut réaliser les explorations précédentes pour son propre corpus, remplir le formulaire et créer des enregistrements comme elle le souhaite. Par conséquent, une dimension « constructible » existe, car il est alors possible de choisir un autre seuil et d'autres options que celles ici privilégiées.

Toutefois, il est vrai également que cette dimension « constructible » peut être considérée comme réduite au sens où plusieurs paramètres faisant partie de la méthode (notamment le choix de la métrique utilisée pour calculer la similarité entre deux chaînes de caractères[139]) ne peuvent pas faire l'objet d'expérimentations. Idéalement, il aurait été pertinent de donner la possibilité de réaliser des expériences aussi à ce niveau. Plusieurs tests auraient pu être réalisés pour aboutir à un choix plus éclairé de tous les éléments de la méthode utilisée. Une telle perspective demande un temps de travail non négligeable et allongerait considérablement cette section. S'il est incontestable que certains choix de cette section sont expliqués et justifiés, il existe également des parties de la méthodologie qui sont en quelque sorte occultées de la perspective de constructibilité pour des raisons pratiques. L'explication de cette situation réside aussi dans le fait que l'objectif n'a pas été de déterminer vraiment la méthode optimale de détection de titre, mais d'effectuer une amélioration globale des textes à analyser dans un temps limité. Dans cette perspective, des choix méthodologiques ont été réalisés avec un certain pas de côté par rapport à l'idéal revendiqué d'une construction totalement transparente et justifiée.

Comme les traitements automatisés des titres réalisés ont conduit à ne pas prendre en compte tous les résultats inférieurs à 0,83 avec la méthode explicitée, ils ont été complétés par des traitements manuels.

---

[138] En se basant sur la variable « user restrict » (*cf.* section Chap4.IV.2)

[139] La distance de Ratcliff/Obershelp a été justifiée rapidement. Elle n'a pas été proposée pratiquement comme un choix possible parmi d'autres distances envisageables.

c. Traitements manuels

Ces ajouts manuels proviennent des examens précédemment menés des cas où la méthode utilisée obtenait manifestement de mauvais résultats (pour les *Annales de Géographie* **@ArticleAnnAnalyseCas** et pour *L'Espace Géographique* **@ArticleEspAnalyseCas)**. Au niveau du corpus « ArticlePlusCR », ce dernier contenant tous les articles du corpus « Article », j'ai décidé après réflexion d'ajouter les mêmes traitements manuels que ceux réalisés pour le corpus « Article ». Les explorations réalisées ayant montré qu'il n'était pas envisageable de corriger manuellement les titres des comptes-rendus, aucune autre correction manuelle n'a été ajoutée pour le corpus « ArticlePlusCR ».

3. Résumés

a. Explorations

La première fonction créée pour cette exploration permet d'afficher les contenus textuels des résumés référencés par l'UAR *Persée* comme étant en français. Si la décision d'exclure les résumés en langue étrangère a été facilement adoptée dans le cadre de cette recherche, le traitement à réserver aux résumés en langue française a suscité plus d'interrogations. Les résultats obtenus (**@ArticleAnnResumeFr** pour les *Annales de Géographie* et **@ArticleEspResumeFr** pour *L'Espace Géographique*) s'appuient sur ce qui a été documenté par l'UAR *Persée* directement dans les métadonnées des documents. La présentation chronologique adoptée pour ces résultats permet quelques premières analyses. Mise à part une exception de 1932, la véritable diffusion des résumés commence à partir de 1968 pour les *Annales de Géographie*. Pour *L'Espace Géographique*, ils sont présents dès le départ de la revue, c'est-à-dire dès 1972. Ces dates étant très proches, il est possible de penser que d'un point de vue formel, le fait de garder les résumés en français dans les textes à analyser ne crée pas une grande dissymétrie entre les deux revues. En revanche, une comparaison entre une époque antérieure à 1968 et des époques postérieures peut poser un problème d'hétérogénéité si la décision est prise d'inclure les résumés.

Un examen plus attentif des premiers résultats obtenus montre que, souvent, le terme « Résumé. » est inclus dans le contenu pour la revue des *Annales de Géographie* alors qu'il a été exclu pour *L'Espace Géographique*. Une erreur est aussi notable dès le neuvième résumé présenté pour les *Annales de Géographie* car ce dernier est en anglais alors qu'il a été référencé en français. Si la décision est prise d'inclure ces résumés, il y a donc un travail

d'homogénéisation à réaliser à un double niveau : tout d'abord, enlever les termes comme « Résumé. » qui introduisent ces parties et ensuite exclure les résumés en langue étrangère qui ont été mal référencés.

Par rapport à ce qui a été réalisé précédemment pour les titres, il faut souligner qu'il n'est pas utile de remplacer les parties dans l'OCR correspondant aux résumés, car l'UAR *Persée* n'a pas effectué de corrections sur ces éléments. Dans le cas où la décision retenue est de supprimer les résumés, le repérage de ces parties dans les textes issus de l'OCR est moins problématique que celui des titres puisqu'il est possible de s'appuyer sur les coordonnées graphiques enregistrées par l'UAR *Persée* lors de la phase de documentation infrapaginaire. Toutefois, il faut souligner que cette information n'a malheureusement pas été incluse dans le format XML-TEI qui a été utilisé pour cette recherche. Il est nécessaire d'aller chercher ces données dans le format Érudit (*cf.* section Chap4.I.3). Il n'existe pas de raison spécifique permettant d'expliquer rationnellement pourquoi ces données sont présentes dans un format et absentes dans un autre. Ce constat rappelle que ces différents formats sont des agrégations de multiples données et que ces agrégations dépendent en amont de choix qui ne sont pas forcément logiques jusque dans leurs moindres détails.

Dans le cas où le choix réalisé est d'exclure les résumés en français des textes à analyser, des petits travaux d'homogénéisation sont également nécessaires. En effet, dans certains cas, la fenêtre graphique tracée par l'UAR *Persée* pour délimiter les résumés contient la mention « Résumé. » alors que dans d'autres cas, cette mention peut être en dehors de la fenêtre graphique tracée. Ainsi, si tous les mots situés à l'intérieur de la fenêtre graphique tracée sont enlevés dans le texte de l'OCR suite à la décision d'exclure les résumés, il se peut que dans certains cas les annonces de résumé perdurent, car elles se situent en dehors des fenêtres graphiques tracées.

Malgré toutes ces explorations préalables, il est nécessaire de reconnaître que la décision finalement prise pour les résumés dans cette recherche n'a pas découlé de ces différentes considérations, mais a été une conséquence d'une difficulté apparue au moment du traitement des bibliographies (*cf.* section Chap5.IV.8). La résolution de cette difficulté est passée par la suppression de toutes les parties textuelles situées après la dernière bibliographie. Or, il existe de nombreux cas dans les *Annales de Géographie* où le résumé est situé en fin d'article. Pour homogénéiser les traitements, il a paru logique de supprimer tous les résumés des textes. Toutes les analyses liées aux explorations précédentes n'ont pas été perdues puisque l'objectif d'éliminer les annonces (« Résumé »…) qui peuvent se trouver avant les fenêtres graphiques tracées sur ce sujet par l'UAR *Persée* a été lui gardé.

Trois petites difficultés méritent d'être mentionnées. Tout d'abord, ces annonces peuvent prendre des formes multiples du fait des diverses langues. Ensuite, elles peuvent être en

majuscule ou non et accolées à une ponctuation (un point ou un double point) ou non. Enfin, elles sont généralement situées juste avant le résumé, mais ce n'est pas forcément toujours le mot situé juste avant la fenêtre graphique. Il est en soi possible que des mots s'intercalent. Face à cette situation, une décision peut consister à enlever tout ce qui ressemble à une annonce de résumé dans toutes les parties de textes situées avant les résumés. Toutefois, une telle décision n'est pas vraiment judicieuse du fait de la présence dans les *Annales de Géographie* de résumés situés en fin d'article. Le risque de détecter et de supprimer de faux positifs (des termes ressemblant à des annonces de résumés, mais n'en étant pas) est alors grand.

Pour faire face à ces difficultés, une interface d'exploration a été créée. Elle propose trois fenêtres à compléter. La première offre la possibilité de rentrer plusieurs formes comme « Résumé », « RESUME », « Abstract »… La deuxième propose de spécifier plusieurs ponctuations. Par exemple, « . », « : »… Enfin, la troisième permet de spécifier un nombre de termes avant les résumés[140] dans lesquels toutes les combinaisons « forme + ponctuation » qui viennent d'être entrées (par rapport aux exemples précédemment mentionnés : « Résumé. », « Résumé : », « RESUME. », « RESUME : ») vont être cherchées. La technique employée repose alors sur le fait de pouvoir mener une exploration exhaustive sur l'ensemble de formes probables à partir d'un seuil volontairement établi de manière large. Concrètement, ce seuil a été fixé à 20, car l'annonce d'un résumé est tout de même supposée être assez proche du résumé en question pour être efficiente. En fixant ce seuil, le risque est pris de détecter de faux positifs.

Les résultats obtenus sont disponibles à partir de ces liens : **@ArticleAnnResumeAnnonceCherche** pour les *Annales de Géographie* et **@ArticleEspResumeAnnonceCherche** pour *L'Espace Géographique*. Un extrait des résultats est présenté dans le tableau ci-dessous :

	"Résumé"	"RESUME"	"Abstract"	"ABSTRACT"
""	23	0	13	0
"."	0	0	0	0
":"	0	0	0	0

Tableau n°3 : Extrait des résultats de recherche de combinaisons [mot (en ligne) + ponctuation (en colonne)] avant un résumé pour la revue les *Annales de Géographie* dans le corpus « Article » avec un seuil de 20.

[140] Pour être plus précis, avec les fenêtres graphiques des résumés traçées par l'UAR *Persée*.

La première ligne de ce Tableau n°3 représente une absence de ponctuation. Dans ce cas, la combinaison « terme + ponctuation » recherchée correspond au terme en colonne. Chaque résultat supérieur à 0 se présente comme un lien qu'il est possible d'ouvrir dans l'application. Par exemple, en cliquant sur le 23, une nouvelle fenêtre s'ouvre présentant les 23 cas où le terme « Résumé » a été trouvé dans les 20 mots précédents une fenêtre graphique de résumé dans la revue les *Annales de Géographie*. Les résultats ont été triés automatiquement par ordre décroissant (des combinaisons « terme + ponctuation » les plus éloignées aux plus proches de la fenêtre graphique) pour permettre une vérification des cas les plus suspects de faux positifs (plus une combinaison « terme + ponctuation » est loin de la fenêtre graphique, moins elle a de chance d'être une annonce de résumés). Tous ces liens ont été ouverts pour les deux revues sur le corpus « Article ». Seul un cas a été trouvé où la distance entre la fenêtre graphique et la combinaison trouvée est supérieure à 2 (sur la revue les *Annales de Géographie* avec le terme « Résumé »). Une vérification manuelle a permis de confirmer qu'il s'agissait bien d'un cas d'annonce de résumé et qu'il ne s'agissait pas d'une combinaison « terme + ponctuation » faisant partie du cœur scientifique du texte qu'il aurait été dommageable d'enlever.

Les résultats obtenus permettent de prouver que ce phénomène d'annonces non annotées est finalement peu répandu (aucun cas dans *L'Espace Géographique* et quelques cas dans les *Annales de Géographie*). Enfin, les résultats pour le corpus « ArticlePlusCr » sont identiques, car les comptes-rendus ne comprennent pas de résumé.

b. Traitements automatisés

À partir de ces explorations, une fonctionnalité pour le traitement automatisé des résumés a été créée comme pour les titres avec 3 paramètres à préciser. Une première fenêtre permet de renseigner les combinaisons « terme + ponctuation » à enlever avant une fenêtre graphique de résumé. Un deuxième paramètre concerne le nombre de mots avant la fenêtre graphique dans lequel sont cherchés les termes ou combinaisons de termes avec une ponctuation, précédemment précisés. Enfin, une option a été créée pour pouvoir ajouter (ou

non) les résumés en français si jamais une recherche ultérieure souhaite prendre une décision différente sur ce point[141].

Les choix réalisés pour les deux corpus (**@ArticleTraitAutoResume** pour le corpus « Article » et **@ArticlePlusCrTraitAutoResume** pour le corpus « ArticlePlusCr ») se fondent sur les résultats précédents. La dernière exploration ayant montré que la distance maximale entre la fenêtre graphique et la combinaison « terme + ponctuation » trouvée avec un seuil de 20 est de 4, le seuil a finalement été fixé à 5 pour prendre en compte tous les cas.

Aucun traitement manuel n'a été effectué pour les résumés sur les deux corpus de cette thèse.

4. Mots-clés

a. Explorations

La section « Exploration » a la même structure générale que celle des résumés. Une différence réside dans les coordonnées des zones graphiques qui n'ont pas été stockées par l'UAR *Persée* dans le format Érudit. Étant également absentes du format XML-TEI, il a fallu s'appuyer sur des données récupérées au format CSV (*cf.* section Chap4.I.3). Au-delà de l'anecdote de devoir ainsi aller chercher des informations dans différents fichiers, il est intéressant de souligner que cette demande a permis de se rapprocher des données les plus basiques que possède l'UAR *Persée*. En effet, les formats XML-TEI ou Érudit sont déjà des reconstructions avancées. Les matériaux de base de l'UAR *Persée* sont les documents numérisés, les résultats des OCR et toutes les zones graphiques qui ont été consignées dans différentes bases de données. Les fichiers CSV obtenus correspondent à l'extraction de deux de ces bases de données, celles correspondant aux mots-clés pour les deux revues étudiées.

La première exploration permet d'afficher simplement tous les mots-clés en français trouvés. Les résultats montrent une dissymétrie entre les deux revues beaucoup plus importante que pour les résumés. Les mots-clés apparaissent seulement en 1984 pour les *Annales de Géographie* (**@ArticleAnnMotCleFr**) alors qu'ils sont présents dès 1974 pour *L'Espace Géographique* (**@ArticleEspMotCleFr**). Cette dissymétrie légitime le fait de ne

[141] L'algorithme actuel laisse le texte des résumés directement issus de l'OCR. Idéalement, il faudrait réaliser les travaux d'homogénéisation précédemment évoqués sur les textes stockés en métadonnées par l'UAR *Persée* et remplacer les parties de l'OCR concernées (qui peuvent être trouvées à partir des coordonnées graphiques) par les textes corrigés. Il conviendrait donc d'améliorer quelque peu l'algorithme de traitement si une telle décision était prise. Il faudrait également revenir sur la difficulté concernant les bibliographies qui a été évoquée dans cette partie.

pas prendre en compte les mots-clés dans les textes à analyser pour respecter une certaine homogénéité utile pour comparer ensuite les résultats entre les deux revues.

La deuxième exploration présente, comme précédemment, le résultat de la recherche de combinaisons « terme + ponctuation » spécifiques (« Mots-clés », « Mots-clés. »…) à proximité des zones graphiques de mots-clés délimitées par l'UAR *Persée*. Avec le même seuil de 20 mots précédents la zone graphique, il y a quelques points communs avec les résultats déjà obtenus. Il y a une quasi-absence d'annonce des mots-clés pour la revue *L'Espace Géographique* (**@ArticleEspMotCleAnnonceCherche**). Pour les *Annales de Géographie*, il y a quelques cas de combinaisons « terme + ponctuation » trouvées (**@tArticleAnnMotCleAnnonceCherche**).

Toutefois, contrairement aux résultats obtenus pour les résumés, beaucoup de cas présentent une distance assez importante par rapport à la fenêtre graphique (surtout pour le terme « Mots-clés »). Un examen attentif des résultats montre qu'il existe des doublons. Les articles apparaissent souvent deux fois : une fois avec une distance faible entre la zone graphique et le terme « Mots-clés » et une fois avec une distance plus importante. Cela provient du fait que les mots-clés en français et en anglais correspondent parfois à des zones graphiques très proches. Il arrive alors que la mention d'une zone antérieure soit trouvée avec un seuil de 20 mots.

Ces résultats empêchent de traquer facilement les faux positifs obtenus par l'algorithme utilisé comme précédemment pour les résumés. Toutefois, les zones situées avant les mots-clés étant dans la majorité des cas des résumés et les résumés ayant été enlevés, cela rend très peu probable la présence d'un faux positif ayant réellement un impact dans le cas de cette recherche. Enfin, si la table « Transformer » a un doublon visant à enlever deux fois la même combinaison textuelle[142], il a été prévu dans la suite qu'elle ne soit enlevée qu'une seule fois.

Ces réflexions expliquent pourquoi, malgré des résultats d'exploration très différents, un algorithme similaire à celui des résumés a été utilisé pour le traitement des mots-clés.

b. Traitements automatisés

Les liens suivants permettent de consulter les traitements automatisés réalisés : **@ArticleTraitAutoMotCle** pour le corpus « Article » et **@ArticlePlusCrTraitAutoMotCle** pour le corpus « ArticlePlusCr ». Les paramètres

[142] Une fois du fait de sa proximité avec la zone graphique qu'elle annonce, une autre fois du fait d'une autre zone de mot-clé un peu plus lointaine mais restant à moins de 20 mots.

choisis s'expliquent par les explorations précédemment menées. Comme ces dernières n'ont pas abouti à définir un nombre de mots optimal dans lesquels chercher les termes indiqués, le seuil initial très large de 20 a été conservé.

Aucun traitement manuel n'a été effectué pour les mots-clés sur les deux corpus.

5. Hauts de page

a. Explorations

L'objectif est ici de supprimer toutes les mentions souvent répétées en haut de chaque page du nom de l'article, des auteurs ou de la revue. Ces éléments n'ont malheureusement pas été annotés par l'UAR *Persée*. De plus, les fichiers XML-TEI utilisés ne permettent pas de récupérer directement une structuration par ligne. Toutefois, à partir des coordonnées graphiques des mots sur chaque page, des sauts de ligne peuvent être assez facilement détectés, car il y a une rupture spatiale (le mot suivant est situé bien en dessous du mot précédent). En fixant un seuil adapté[143], il est alors possible d'afficher toutes les premières lignes de chaque page pour tous les documents. Il est intéressant d'indiquer dans le même temps la coordonnée spatiale du mot le plus bas de chaque ligne trouvée. Les résultats obtenus sont très satisfaisants dans une première approche. Beaucoup de mentions de haut de page apparaissent pour l'ensemble des *Annales de Géographie* (**@ArticleAnnHtdePageCherche**) et avant 1990 pour *L'Espace Géographique* (**@ArticleEspHtDePageCherche**). En effet, il y a, après cette date, un changement de mise en page de cette revue qui supprime ces mentions de haut de page.

Du fait de l'aspect très répétitif de ces mentions de haut de page, des exceptions peuvent être facilement repérées. Il s'agit le plus souvent d'un titre d'une figure, parfois d'une référence bibliographique. Pour le corpus « ArticlePlusCR » dont les résultats sont disponibles à partir des liens suivants (**@ArticlePlusCrAnnHtDePageCherche** pour les *Annales de Géographie* et **@ArticlePlusCrEspHtDePageCherche** pour *L'Espace Géographique*), il existe des cas où cette technique détecte la première ligne des comptes-rendus et non une annotation de hauts de page. Ce phénomène provient du fait que les comptes-rendus sont souvent réunis dans une partie les présentant tous à la suite les uns des autres, sans forcément qu'un compte-rendu donne lieu à une nouvelle page. Dans ce cas-là, l'UAR *Persée* a délimité le texte de chaque compte-rendu, mais la mention détectée par la

[143] Après quelques essais, le seuil utilisé pour cette recherche pour détecter les sauts de lignes a été de 50 pixels. Il peut être modifié dans la fonction « DocMotPageLigne » du fichier « AllApps/Pretraitement/Persee/DelimitCorpus/outils/xpath.py »

méthode créée n'est plus le haut d'une page, mais la première phrase du compte-rendu en question. Le fait d'avoir ajouté la mention de la coordonnée spatiale la plus basse pour chaque ligne permet d'illustrer quantitativement que les exceptions sont très souvent situées bien plus bas sur la page que les annotations de haut de page. À la suite d'une telle constatation, l'objectif a été d'essayer de fixer un seuil spatial permettant de discriminer les premières lignes qui sont bien des annotations de haut de page de celles qui n'en sont pas. Une nouvelle exploration a été menée dans cette optique.

Les mentions de haut de page étant situées généralement entre 200 et 400 pixels[144], la deuxième fonction proposée n'affiche que les premières lignes dont les coordonnées spatiales de tous les mots sont supérieures à 400 pixels. Les résultats ne sont plus groupés par document, mais triés de manière croissante par rapport à la coordonnée spatiale maximale trouvée (*i.e*. la plus basse). L'objectif est ainsi de déterminer un seuil à partir duquel les premières lignes ne font plus référence à des mentions de haut de page. Les résultats obtenus (**@ArticleAnnHt+400Order** et **@ArticleEspt+400Order**) montrent qu'un tel seuil n'apparaît pas nettement dans le cas des *Annales de géographie*. Il y a entre les figures et les bibliographies beaucoup de bruit. Pour faire face à ce problème, il a été recherché dans tous ces résultats les mentions les plus lointaines ressemblant au syntagme « Annales de Géographie ». En effet, la répétition sur une des deux pages de ce nom est une constante des mentions de haut de page pour cette revue. La dernière mention ainsi repérée a pour valeur minimale 623. Il a été donc décidé de fixer un seuil à 625 pour cette revue. Pour *L'Espace Géographiqu*e, la moindre quantité de données permet de repérer plus facilement la dernière mention de haut de page. Elle a pour coordonnée minimale la valeur 471. Le seuil a par conséquent été fixé à 475 pour cette revue.

Il est vrai que dans les deux revues, et surtout dans les *Annales de Géographie*, il y a en dessous des seuils retenus des éléments qui ne sont pas des mentions de haut de page. Quand ces éléments relèvent de bibliographies, de figures ou de titres secondaires, leur présence n'est pas problématique, car ces parties spécifiques sont retirées ultérieurement (*cf.* sections Chap5.IV.8, 10 et 11). Cette réflexion permet de consulter les résultats précédemment obtenus en essayant de repérer les cas qui ne relèvent pas de ces parties spécifiques (bibliographies, figures et titres secondaires) et qui sont situés en dessous des seuils fixés. Le repérage des parties de texte risquant d'être effacées par erreur a été ainsi réalisé, car s'il est possible bien entendu d'imaginer des automatisations d'un tel travail, il a paru beaucoup

[144] Les *Annales de Géographie* présente de très nombreuses exceptions avec des articles qui commencent jusqu'en 1998 très bas sur la page. Ces exceptions ne touchent que les premières pages des articles qui ont déjà été traitées par ailleurs dans la partie sur les titres principaux. Il n'est donc pas nécessaire de traiter ces exceptions plus en profondeur.

plus simple et rapide de l'effectuer manuellement. Six cas seulement ont été trouvés pour les *Annales de Géographie*. Aucun cas n'a été repéré dans les résultats de *L'Espace Géographique*. Enfin, l'exercice n'a pas été mené pour le corpus « ArticlePlusCr ». Cette décision, qui témoigne de l'acception d'une qualité moindre pour ce corpus, s'inscrit dans la continuité de ce qui a été réalisé précédemment pour les titres (*cf.* section Chap5.IV.2).

b. Traitements automatisés

Les traitements automatisés réalisés (**@ArticleTraitAutoHtDePage** et **@ArticlePlusCrTraitAutoHtDePage**) s'appuient sur les seuils déterminés lors de la phase d'exploration.

c. Traitements manuels

Le traitement manuel du corpus « Article » (**@ArticleTraitMainHtDePage**) concerne les six cas d'erreurs repérés pour les *Annales de Géographie* dans la phase d'exploration. Ces traitements manuels ont été également réalisés pour le corpus « ArticlePlusCr ». L'affirmation précédente d'une moindre qualité de ce corpus aurait pu justifier le fait de ne pas effectuer ces traitements manuels. Toutefois, ceux-ci ayant été précédemment réalisés, il était particulièrement simple de les reproduire. Enfin, le fait de traiter la partie sur les articles du corpus « ArticlePlusCr » comme celle du corpus « Article » peut se justifier dans un objectif de comparabilité des deux corpus. Certes, la qualité de traitement des comptes-rendus n'est pas totalement identique à celle des articles pour des raisons pratiques. Toutefois, la comparaison entre les deux corpus a l'avantage de ne porter que sur cet ajout des comptes-rendus si, par ailleurs, les contenus et les traitements des parties concernant les articles ont été similaires.

6. Bas de page

a. Explorations

La problématique des bas de page est globalement similaire à celle des hauts des pages : il existe souvent des inscriptions situées en bas de page qui méritent d'être retirées par rapport à l'objectif poursuivi, mais qui n'ont pas été annotées par l'UAR *Persée*. Une première fonction a été construite sur le même modèle que celle réalisée pour les hauts de

page afin d'afficher toutes les dernières phrases de chaque page composant les documents d'un corpus. Les résultats obtenus pour les *Annales de Géographie* sur le corpus « Article » (**@ArticleAnnBasDePageCherche**) montrent qu'il existe plusieurs indications de bas de page qui méritent d'être enlevées. Elles commencent très fréquemment par la forme « ANN DE GEOG » qui a été parfois mal reconnue par l'OCR. De plus, il existe en 1993 quelques numéros avec des indications de bas de page plus systématiques commençant par « Gèo ». Après cette période, les indications de bas de page pour les *Annales de Géographie* se limitent à la première page de chaque numéro avec souvent la mention de la maison d'édition : « Armand Colin ». Les résultats pour *L'Espace Géographique* sur le même corpus (**@ArticleEspBasDePageCherche**) sont assez symétriques par rapport à ceux obtenus pour les hauts de page. En effet, il n'existe pas d'indication de bas de page avant 1990 et elles deviennent systématiques après cette date, à quelques exceptions près[145].

La difficulté pour enlever les indications de bas de page concerne surtout les *Annales de Géographie* car elles ne sont pas systématiques et prennent des formes irrégulières dues aux erreurs d'OCR. Un premier essai a été réalisé en essayant de déterminer, comme précédemment, un seuil spatial qui permettait de discriminer globalement les dernières phrases qui relèvent d'indications de bas de page à enlever par rapport à celles qui n'en sont pas. Le résultat obtenu pour le corpus « Article » (**@ArticleAnnBasDePageOrder)** présente trop de bruit pour déterminer un tel seuil spatial. L'expérimentation a été poussée plus loin en essayant d'enlever les éléments qui relèvent de notes ou de figures. Le nouveau résultat obtenu pour le même corpus (**@ArticleAnnBasDePageOrderSuppr**) ne permet toujours pas de déterminer de manière satisfaisante un seuil spatial. Suite à ces constatations, un changement d'optique a été réalisé en cherchant, non plus une détermination basée uniquement sur un seuil spatial, mais plutôt en utilisant le contenu textuel des lignes détectées. Un ensemble de combinaisons textuelles a été défini suite à plusieurs essais pour essayer de couvrir un maximum de cas tout en minimisant le nombre de faux positifs.

Les indications de bas de page sont retenues si elles correspondent à un des critères suivants :

- Dernière ligne qui commence ou finit par « ANN » ou « Ann »,
- Dernière ligne qui commence par « Gèo. » ou « Geo. »,
- Dernière ligne qui commence par « ARMAND » ou « Armand »,
- Dernière ligne qui inclue « Année. » ou « ANNEE. »,
- Dernière ligne qui commence par « DE », « DB », « DK », « DR » ou « DL »[146].

[145] Par exemple, sur la dernière page d'un l'article après la bibliographie, il n'existe pas forcément de note de bas de page.

[146] Ces cas prennent en compte quelques erreurs d'OCR dans la reconnaissance de ces bas de page.

Le résultat obtenu (**@ArticleAnnBasDePageTextualCombi**) est assez satisfaisant, même si quelques erreurs peuvent être facilement repérées. Toutefois, les multiples essais et les analyses de leurs résultats pour trouver les combinaisons textuelles les plus efficaces me laissent penser que les indications de bas de page qui se répètent le plus et qui sont le plus à même d'avoir un impact dans des traitements statistiques sont enlevées en utilisant cette méthode.

b. Traitements automatisés

Suite aux explorations menées, le traitement automatisé des bas de page a été construit en proposant deux options : la première offre la possibilité de supprimer toutes les dernières phrases détectées pour la revue *L'Espace Géographique* après 1990 ; la seconde permet de supprimer les dernières phrases détectées pour la revue les *Annales de Géographie* si elles relèvent des combinaisons textuelles précédemment exposées. Les traitements réalisés pour les deux corpus « Article » et « ArticlePlusCr » sont identiques avec des choix positifs concernant ces deux options.

c. Traitements manuels

Les erreurs facilement détectables dans le résultat précédemment obtenu (**@ArticleAnnBasDePageTextualCombi**) ont été analysées. Si ces erreurs relèvent de notes, de bibliographies, de figures ou de titres secondaires qui sont par la suite soit supprimés, soit remplacés à l'aide des informations de documentation infrapaginaire de l'UAR *Persée* (*cf.* sections Chap5.IV.7, 8, 10 et 11), elles n'ont pas été prises en compte. *In fine*, tous les cas analysés relevant de ces situations spécifiques, il n'y a pas eu de traitement manuel réalisé.

7. Notes

a. Explorations

Trois types de notes ont été délimités par l'UAR *Persée* : les notes de bas de page, les notes biographiques[147] et les notes de la rédaction[148]. Les notes de bas de page ont des statuts très variables suivant les auteurs et relèvent d'un genre textuel particulier avec des répétitions de termes comme « ibid. » et « op.cit. ». Cela explique pourquoi les notes de bas de page ont été supprimées dans les deux corpus créés pour cette recherche. Les notes biographiques et les notes de la rédaction relèvent également de genres textuels spécifiques et elles peuvent être considérées comme secondaires par rapport à l'objectif par cette recherche. Elles ont donc été aussi enlevées des deux corpus créés.

Lors d'un premier travail sur les évolutions lexicales (*cf.* section Chap2.II.5.c), j'avais pu m'apercevoir que certaines notes biographiques n'avaient pas été annotées par l'UAR *Persée*. Pour essayer de remédier à ce problème, une première fonction a été créée permettant d'afficher toutes les notes biographiques existantes. Les résultats obtenus (**@ArticleAnnNoteBioExist** pour les *Annales de Géographie* et **@ArticleEspNoteBioExist** pour *L'Espace Géographique*) montrent une structure commune de ces notes biographiques. D'abord, sur la première ligne, le nom et le prénom d'un auteur. Puis, sur une deuxième ligne, son grade ou son organisme d'affiliation. Ces résultats ont permis de répertorier les grades (« Professeur », « Maître de conférences », « Assistant »…) et les organismes d'affiliation (« Université », « Laboratoire », « CNRS »…) les plus utilisés.

Les listes de grades et d'organismes d'affiliation ainsi obtenues ont été ensuite utilisées dans l'élaboration d'une deuxième fonction pour chercher la structure précédemment identifiée (une première ligne avec le nom ou le prénom d'un auteur suivi d'une deuxième ligne avec un grade ou une affiliation) dans des articles où aucune note biographique n'avait été annotée par l'UAR *Persée*. Pour tenir compte de différences minimes d'écriture ou de petites fautes dues à l'OCR, les recherches de nom(s), de prénom(s), de grade(s) et d'organisme(s) d'affiliation ont eu recours à la distance de Ratcliff/Obershelp précédemment utilisée pour les titres (*cf.* section Chap5.IV.2). Seuls les taux de ressemblance supérieurs à 0,9 entre une entité (nom, prénom, grade ou organisme d'affiliation) et un terme ont été retenus. Les résultats obtenus (**@ArticleAnnNoteBioCherche** pour les *Annales de*

[147] Présentation brève de (ou des) auteur(s) en début (ou fin) d'article avec leur affiliation.
[148] *Ndlr* et *erratum*.

Géographie et **@ArticleEspNoteBioCherche** pour *L'Espace Géographique*) permettent de détecter plusieurs cas où des notes biographiques non annotées sont suspectées. Pour améliorer leur reconnaissance, chaque résultat est accompagné de sa localisation : première ou dernière page de l'article et numéro de la ligne où une note biographique est suspectée d'après cette fonction. Cela a permis d'examiner plus rapidement l'ensemble de ces cas.

Au niveau des notes de rédaction, ces dernières sont rares et peuvent être considérées comme marginales par rapport à la masse textuelle des documents dans les deux corpus créés. À l'inverse, les notes de bas de page sont globalement nombreuses (avec des disparités importantes suivant les documents). Cette considération explique pourquoi une exploration des notes de bas de page a été privilégiée. Toutefois, il n'est pas si facile de réaliser des explorations qui permettent d'améliorer les annotations réalisées par l'UAR *Persée* et de détecter des erreurs comme précédemment pour les notes biographiques. Ma fréquentation des données me conduit à penser qu'il est plutôt rare qu'une note de bas de page ait été oubliée d'être annotée. Ce que j'ai pu *a contrario* parfois observer est l'existence d'une fenêtre graphique tracée par l'annotateur un peu trop proche des mots. Il en résulte que les coordonnées graphiques d'un mot, coordonnées définies par l'OCR, peuvent alors dépasser le cadre de la fenêtre graphique tracée. Le mot n'est alors pas compté comme faisant partie de la note.

À la suite de ce constat, une exploration a été menée pour essayer d'afficher toutes les notes présentant une discontinuité par rapport au texte issu de l'OCR. La note est alors composée de plusieurs blocs textuels qui ne forment pas un tout continu dans le texte de l'OCR. Une fonction a été construite pour détecter tous les cas de notes présentant une telle discontinuité textuelle au sein d'un corpus. Les résultats obtenus (**@ArticleAnnNoteBasDiscon** et **@ArticleEspNoteBasDiscon**) montrent que ces cas sont assez nombreux : 782 pour les *Annales de Géographie* et 94 pour *L'Espace géographique*. Il n'y a pas eu de traitement automatisé ou manuel effectué par rapport à l'ensemble de ces cas. En conséquence, il existe un ou parfois quelques mot(s) appartenant(s) aux notes qui sont intégré(s) dans le texte et non pas supprimé(s). Le phénomène reste assez marginal par rapport à l'ensemble des notes : 0,04 % des notes de bas de page sont concernées pour les *Annales de Géographie* et 0,02 % pour *L'Espace Géographique.*

Les résultats ayant été classés par ordre décroissant par rapport à la discontinuité maximale[149] détectée dans chaque note, une différence très nette peut être observée entre les deux revues. *L'Espace Géographique* présente de nombreuses discontinuités assez importantes (supérieures à 10 mots) alors que ce n'est pas le cas des *Annales de géographie*. Cette différence s'explique par la mise en page sous forme de double colonne dans *L'Espace Géographique*. Une colonne s'intercale parfois dans une note, ce qui explique la présence de ces discontinuités importantes dans cette revue. En inversant la perspective, ces résultats montrent que les doubles colonnes ont plutôt été très bien reconnues par l'OCR, car sinon le nombre de discontinuités trouvées serait bien supérieur. Pour les *Annales de Géographie*, la présence accrue de petites discontinuités s'explique en partie par des pages numérisées penchées beaucoup plus nombreuses que pour *L'Espace Géographique*. Quand l'annotateur doit tracer une fenêtre graphique de forme rectangulaire sur un texte penché, il a plus de chance de finir très proche d'un bord du texte, et donc que les coordonnées graphiques de quelques mots dépassent de la fenêtre tracée.

b. Traitements automatisés

Le traitement automatisé présente un formulaire qui permet pour chaque type de note de choisir de leur maintien ou non dans les textes. Pour les deux corpus « Article » et « ArticlePlusCr », les notes biographiques, les notes de l'éditeur et les notes de bas de page annotées par l'UAR *Persée* ont toutes été supprimées.

c. Traitements manuels

Les traitements manuels pour les notes (**@ArticleTraitMainNote** et **@ArticlePlusCrTraitMainNote**) sont identiques pour les deux corpus. Ils proviennent de la détection précédemment réalisée de notes biographiques non annotées par l'UAR *Persée*.

[149] Cette taille est comptée en nombre de mots. Pour comprendre cette valeur, les intervalles composant les notes ont été ajoutés dans les résultats. Par exemple, une note ayant comme pour intervalle [(3,6), (8,10),(13,15)] commence au troisième mot du texte de l'OCR, avec deux trous entre le sixième et le huitième mot et entre le dixième et treizième mot (le septième, le onzième et le douzième mot du texte de l'OCR n'appartiennent pas à la fenêtre graphique correspondant à cette note) et une fin au quinzième mot. Dans ce cas, les deux trous sont de un mot et de deux mots. La discontinuité maximale a par conséquent pour valeur 2.

Le numéro de la note concernée, ainsi que les termes correspondant à ces intervalles ont été ajoutés, permettant de retrouver là où se localisent ces discontinuités.

8. Bibliographies

a. Explorations

Les bibliographies sont à l'instar des notes des éléments textuels spécifiques. Pour cette raison, j'ai décidé de les enlever des textes à analyser dans cette recherche. Si les bibliographies ont été globalement annotées par l'UAR *Persée*, leur retrait a été beaucoup plus complexe que prévu[150] à cause de la façon dont ces éléments sont représentés dans les fichiers XML-TEI. Les problèmes rencontrés concernent plus particulièrement les bibliographies qui s'étendent sur plusieurs pages d'affilée et qui sont suivies d'un texte. Dans ce cas, l'UAR *Persée* a créé dans les fichiers XML-TEI un premier bloc de données « bibliographie » qui compile les différentes pages de bibliographie avec l'indication des sauts de page dans ce bloc. La difficulté provient du bloc suivant qui représente le texte, quand il existe, venant à la suite de cette bibliographie de plusieurs pages. L'UAR *Persée* a fait le choix de réintégrer toutes les pages de bibliographies précédentes dans ce bloc textuel sans les indications de saut de page puisqu'elles avaient déjà été intégrées dans le bloc précédent[151].

La conséquence de ce choix est qu'en récupérant les données, si le premier bloc est ignoré, les sauts de page sont perdus, ce qui est problématique, car une grande partie des traitements précédemment réalisés reposent sur cette détection ; et si le second bloc est ignoré, des données textuelles ne sont pas prises en compte[152]. Le choix réalisé a donc été de récupérer le contenu des deux blocs ce qui a eu pour conséquence de créer des répétitions dans les textes sur ces parties spécifiques. Après pas mal de recherche pour essayer d'ôter ces doublons de la manière la plus efficace possible, une première exploration a été créée pour afficher les cas de documents où un contenu textuel existe après la dernière bibliographie et ne relève pas d'éléments déjà supprimés (résumés, mots-clés et notes).

Les résultats obtenus (**@ArticleAnnApresDernBiblio** pour les *Annales de Géographie* et **@ArticleEspApresDernBiblio** pour *L'Espace Géographique***)** montrent que, dans la plupart des cas, ce qui suit la dernière bibliographie n'appelle pas à être gardé pour cette recherche. Suite à ce constat, j'ai choisi de m'orienter vers un traitement automatisé où la

[150] Il peut être en effet pensé qu'en utilisant les coordonnées graphiques des différentes parties de bibliographies délimitées par l'UAR *Persée*, il sera possible de les enlever très facilement.

[151] Ce procédé peut se justifier, car dans le premier bloc de données, seules les bibliographies étaient représentées. Si d'autres données sont présentes sur ces pages, elles n'apparaissent pas dans ce premier bloc. En répétant l'ensemble du contenu textuel des pages contenant les bibliographies avant le bloc textuel suivant, cela permet d'avoir des données complètes.

[152] De manière évidente, le texte suivant la bibliographie, mais aussi tous les autres éléments textuels qui étaient sur les pages de bibliographies, mais n'ont pas été logiquement intégrés dans le premier bloc.

dernière bibliographie, mais aussi tout ce qui se trouve après celle-ci, est supprimé. Les résultats de cette première exploration peuvent alors être utilisés pour trouver les cas où un tel traitement supprime des contenus scientifiques qui devraient être gardés dans l'optique de la recherche menée. La phase de traitement manuel permet alors de corriger ces erreurs.

Une observation particulièrement étonnante pouvant être réalisée à partir des résultats précédents est la présence de bibliographies pour *L'Espace Géographique* alors que la méthode mise en œuvre est censée afficher le contenu textuel situé après la dernière bibliographie. Ceci provient du fait que l'UAR *Persée* a parfois répété dans ses données que la première partie d'une bibliographie (située sur une première colonne). Le texte correspondant à la seconde colonne apparait alors dans l'affichage des résultats de l'exploration, malgré une annotation correcte réalisée en amont par l'UAR *Persée.* C'est ici dans les étapes intermédiaires de création des fichiers XML-TEI à partir des annotations que se trouvent les erreurs. De nombreux autres éléments trouvés après les bibliographies dans cette revue relèvent de publicités ou de recommandations aux auteurs qui n'ont pas été annotées et qui méritent d'être supprimées dans le cadre de cette recherche. Pour deux documents seulement (spgeo_0046-2497_1988_num_17_2_2760 et spgeo_0046-2497_1984_num_13_4_3946), la suppression de toute la partie après la dernière bibliographie apparaît comme une erreur. Ces deux cas ont fait l'objet de corrections lors de la phase de traitement manuel.

Pour les *Annales de Géographie*, on remarque la présence de nombreuses notes biographiques, mais aussi de textes plus longs. Une grande partie de ces textes relève de comptes-rendus qui ont été définis de manière erronée comme « Article » par l'UAR *Persée* et qui mentionnent, dès le départ, en note bibliographique la référence traitée par le compte-rendu. Ces documents ont été enlevés du corpus « Article » et corrigés dans le corpus « ArticlePlusCr » en traitement manuel. Au-delà de ces comptes-rendus mal référencés, seulement deux articles (geo_0003-4010_1905_num_14_76_6413 et geo_0003-4010_1892_num_1_4_19938) et un document situé dans une limite floue entre l'article et le compte-rendu (geo_0003-4010_1907_num_16_88_6842) ont fait l'objet de corrections manuelles suite à l'application de la méthode précédemment explicitée.

Une deuxième exploration a été réalisée pour identifier les articles où il y a plusieurs bibliographies. En effet, la méthode qui a été exposée ne prend pas en compte ces cas particuliers puisqu'elle ne traite que de la dernière bibliographie. Les résultats obtenus (**@ArticleAnnPlusieursBiblios** pour les *Annales de Géographie* et **@ArticleEspPlusieursBiblios** pour *L'Espace Géographique*) montrent que peu d'articles possèdent plusieurs bibliographies. Après un examen de ces cas, j'ai choisi d'appliquer pour le traitement la méthode simple de retrait à partir des coordonnées graphiques quand la

bibliographie concernée n'est pas la dernière du document et de faire des modifications manuellement si nécessaire. Les cas où une bibliographie intermédiaire s'étend sur plusieurs pages ont été méthodiquement examinés.

Une troisième exploration a été menée pour vérifier si certains mots pouvant annoncer les bibliographies ne se trouvent pas en dehors des parties délimitées par l'UAR *Persée*. La même méthode que pour les résumés et les mots-clés a été utilisée avec une recherche exhaustive et un examen détaillé des résultats (**@ArticleAnnBiblioAnnonceCherche** pour les *Annales de Géographie* et **@ArticleEspBiblioAnnonceCherche** pour *L'Espace Géographique*) afin de déterminer d'éventuels faux positifs et de fixer la délimitation de la zone de recherche la plus pertinente possible.

Enfin, une dernière exploration traite plus spécifiquement des comptes-rendus. Il s'agit d'identifier tous les cas où un compte-rendu possède au moins une bibliographie : **@ArticlePlusCrAnnBiblioInCr** pour les *Annales de Géographie* et **@ArticlePlusCrEspBiblioInCr** pour *L'Espace Géographique*. Il est facile de trouver alors plusieurs comptes-rendus présentant la même difficulté que ce qui avait été repéré précédemment pour des comptes-rendus malencontreusement référencés comme article. Pour contourner cette difficulté, un retrait en utilisant les coordonnées graphiques a été privilégié. Les résultats de cette exploration ont été utilisés également pour vérifier que les comptes-rendus présentant des bibliographies s'étendant sur plusieurs pages successives, ne présentent pas le problème de répétitions précédemment rencontré dans le cadre des articles. Seulement quatre cas présentant ce problème ont été détectés. Ils ont été corrigés manuellement dans le corpus « ArticlePlusCr ».

b. Traitements automatisés

Le traitement des bibliographies peut être facilement réalisé en cochant la case en face de l'intitulé : « Supprimer les bibliographies par leurs coordonnées graphiques et tout ce qui suit la dernière bibliographie d'un article » dans la boîte de dialogue de l'application[153]. Pour faire suite à la troisième exploration, l'application propose également de préciser des termes et une zone de recherche pour effacer les annonces de bibliographies n'étant pas incluses dans les annotations de l'UAR *Persée*. Suite à l'examen des résultats obtenus lors de la troisième exploration, les termes « BIBLIOGRAPHIE », « Bibliographie », « Références »,

[153] Pour un utilisateur ayant des droits d'écriture sur un corpus.

« Orientation » et une zone de recherche de 20 mots[154] avant chaque bibliographie ont été retenus pour les deux corpus « Article » et « ArticlePlusCr ».

c. Traitements manuels

Tous les traitements manuels (**@ArticleTraitMainEltBiblio** et **@ArticleTraitMainExceptDocBiblio** pour le corpus « Article », **@ArticlePlusCrTraitMainEltBiblio** et **@ArticlePlusCrTraitMainExceptDocBiblio** pour le corpus « ArticlePlusCr ») proviennent des constats réalisés lors de la phase d'exploration.

9. Annexes

Les annexes s'étendant sur plusieurs pages ne présentent pas le problème précédemment explicité pour les bibliographies. Une exploration a été effectuée pour détecter d'éventuelles annonces des annexes avant les zones graphiques tracées par l'UAR *Persée* sur le modèle de ce qui a été réalisé pour les résumés, les mots-clés et les bibliographies. Les résultats obtenus (**@ArticleAnnAnnexeMotAnnonce** pour les *Annales de Géographie* et **@ArticleEspAnnexeMotAnnonce** pour *L'Espace Géographique*) montrent que seulement quelques documents des *Annales de Géographie* sont concernés. Le traitement des deux corpus s'est fait simplement en retirant les annexes à l'aide des coordonnées graphiques délimitées par l'UAR *Persée* et en tenant compte des résultats de l'exploration menée.

Il n'y a pas eu de traitement manuel réalisé pour les annexes.

10. Titres des parties

Il n'y a pas eu d'exploration spécifique menée pour cette section. Les titres des parties ont simplement été remplacés par leur version corrigée. Pour cela, les coordonnées graphiques des fenêtres et les corrections documentées par l'UAR *Persée* ont été utilisées.

Il n'y a pas eu de traitement manuel sur les titres des parties.

[154] Il n'existe que deux faux positifs sur le terme « Orientation » pour le document spgeo_0046-2497_1992_num_21_1_3038, mais cela renvoie à une figure dont le contenu est par la suite effacé. Il a ainsi paru pertinent de garder le seuil maximum testé. Il n'y a pas eu d'exploration avec des seuils plus importants, car les discontinuités maximales sont en fait dues à des phénomènes marginaux de double colonne. La fenêtre graphique correspondant à la deuxième colonne trouve l'annonce de la première colonne, mais cette annonce est bien dans la fenêtre graphique tracée par l'UAR *Persée* pour la première colonne comme le montre l'absence de doublon sur ces cas.

11. Figures

a. Explorations

Les figures ont été également délimitées par l'UAR *Persée* et leurs titres ont été aussi corrigés lors de la phase de documentation infrapaginaire. Ces titres peuvent être annoncés par des termes généraux comme « Fig » ou « Figure » et d'autres dénominations plus spécifiques comme « Tableau », « Carte », « Croquis »… Une exploration a tout d'abord été effectuée pour rechercher ces annonces les plus fréquentes. La méthode utilisée recense tous les premiers mots rencontrés dans l'ensemble des titres de chaque revue puis les classe par ordre décroissant de fréquence. Les résultats obtenus (**@ArticleAnnFreqAnnonceFig** pour les *Annales de Géographie* et **@ArticleEspFreqAnnonceFig** pour *L'Espace Géographique*) montrent une forte dissymétrie entre les deux revues. Le nombre de mots trouvé pour *L'Espace Géographique* est très inférieur, car cette revue a globalement réalisé des annonces plus normées pour les figures[155]. À partir de ces résultats, il est facile de repérer les termes qui relèvent d'annonces de figures et d'en dresser une liste assez exhaustive[156].

La suppression des annonces générales (« Fig », « Figure »…) s'explique facilement par le fait qu'elles sont variables entre les deux revues et les époques. Elles dépendent avant tout de choix éditoriaux. Le sort des dénominations spécifiques (« Tableau », « Carte », « Croquis »…) a suscité plus de questionnements dans la mesure où elles font référence plus explicitement à un contenu scientifique. Toutefois, elles présentent également des hétérogénéités dans le temps et entre revues. Un examen plus en détail permet de constater qu'elles ne sont pas essentielles par rapport aux contenus scientifiques. Par exemple, les titres, « Carte de la région parisienne » et « Région parisienne », peuvent tout aussi bien convenir pour la même figure. Ces réflexions ont conduit à supprimer également ces dénominations plus spécifiques des textes à analyser.

Une seconde exploration a consisté comme pour les titres à afficher tous les titres de figures contenant une ou des barres obliques. Le résultat obtenu pour les *Annales de Géographie* (**@ArticleAnnFigTitreSlash**) met en évidence qu'entre 1991 et 1998, beaucoup de titres de figure de cette revue ont été traduits en anglais avec une séparation par une barre oblique entre le titre français et anglais. Comme pour les titres des documents, il a été choisi de garder seulement la partie française.

---

[155] Le fait que *L'Espace Géographique* ne commence qu'en 1972 explique en partie cette différenciation. Une partie provient aussi d'une politique éditoriale moins changeante sur ce sujet.

[156] Pour les *Annales de Géographie*, seulement les 100 premiers termes ont été examinés.

L'OCR des figures donne lieu à deux situations très différentes selon que le logiciel utilisé pour cette tâche par l'UAR *Persée* a détecté ou non que la figure était une image. Dans le cas de nombreux tableaux, cartes ou graphiques par exemple, le logiciel a souvent continué à reconnaître automatiquement le texte. Ces parties présentent souvent des éléments textuels discontinus et parfois répétitifs dans le cas de certains tableaux. Ces éléments qui relèvent du contenu des figures n'ont pas été gardés même s'ils renvoient à des contenus scientifiques, car leur forme est souvent très différente de celles des corps de texte. Une difficulté rencontrée provient du fait que les fenêtres graphiques tracées par l'UAR *Persée* sur les figures contiennent parfois le titre, mais pas toujours. Il a été, par conséquent, nécessaire d'essayer de différencier ces deux cas pour savoir si après suppression du texte contenu dans les fenêtres graphiques tracées, il fallait ajouter ou non le titre corrigé.

Une troisième exploration a été menée dans cet objectif en reprenant la technique mise en place pour les titres. Il s'agit de rechercher pour chaque figure si dans le contenu textuel inclus dans la fenêtre graphique correspondante tracée par l'UAR *Persée* une suite de mots est proche du titre corrigé de la figure. À partir des résultats obtenus (**@ArticleAnnTitreFigCherche** pour les *Annales de Géographie* et **@ArticleEspTitreFigCherche** pour *L'Espace Géographique*), la technique utilisée pour les titres, à savoir la détermination d'un seuil de ressemblance à partir duquel le titre est très probablement trouvé, ne s'est pas révélée satisfaisante. En effet, l'OCR permet d'avoir un texte souvent discontinu sur ces zones correspondant à des figures. Le titre d'une figure se retrouve assez souvent en plusieurs parties entrecoupées par des mots du contenu de la figure. Par conséquent, la méthode consistant à rechercher une séquence continue n'est pas toujours pertinente. C'est pourquoi une quatrième exploration a été menée pour essayer de trouver une méthode plus adaptée.

La démarche utilisée consiste à évaluer pour chaque mot du titre s'il existe le même mot ou un terme lui ressemblant fortement[157] dans tout le contenu textuel inclus dans la fenêtre graphique de la figure correspondante. Il est ensuite possible d'obtenir un ratio entre le nombre de cas positifs et le nombre total de mots du titre de la figure. Les résultats obtenus (**@ArticleAnnTitreFigChercheDiscon** pour les *Annales de Géographie* et **@ArticleEspTitreFigChercheDiscon** pour *L'Espace Géographique*) montrent que des titres discontinus dans l'OCR peuvent ainsi être plus facilement détectés. Toutefois, ces résultats ont permis également de mettre en lumière un constat : sur les titres courts, le ratio obtenu peut varier de manière importante suivant la présence ou l'absence de quelques mots. Il est alors difficile de fixer un seuil permettant de discriminer les résultats corrects de ceux

[157] Toujours en utilisant la distance de Ratcliff/Obershelp avec un seuil cette fois-ci de 0,8 (L'OCR sur ces zones de figures étant souvent plus perturbé que sur le corps du texte).

erronés. Cette expérimentation et ce constat m'ont finalement fait privilégier pour le traitement une méthode mixte : pour les titres inférieurs ou égaux à 5 mots, la première technique explicitée (celle employée pour les titres) a été utilisée ; pour les titres supérieurs à 5 mots, la deuxième technique venant d'être explicitée, a été employée. Ce seuil de 5 mots n'a pas été déterminé à la suite d'une évaluation quantifiée, mais de manière approchée à partir des constats réalisés.

Une dernière difficulté importante a été rencontrée au niveau des figures pour lesquelles aucun texte n'a été détecté par l'OCR. Dans ce cas-ci, il est particulièrement complexe d'enlever les termes d'annonces de figures en privilégiant les zones textuelles situées juste avant ou après les figures parce qu'il est difficile de savoir où sont précisément localisées ces zones par rapport au texte. Après réflexion, j'ai décidé d'enlever tous les mots d'annonces de figure (définis à l'aide de la première exploration) de l'ensemble des pages contenant au moins une figure pour contourner ce problème. Ce qui m'a convaincu d'opter pour ce choix est le constat que tous ces termes d'annonces commencent par une majuscule, il y a donc très peu de chance qu'ils se retrouvent ailleurs que dans les titres de figures à l'exception des références du type « Cf. Fig ».

b. Traitements automatisés

Les traitements automatisés réalisés (**@ArticleTraitAutoFigure** pour le corpus « Article » et **@ArticlePlusCrTraitAutoFigure** pour le corpus « ArticlePlusCr ») ont permis d'enlever l'ensemble des mots d'amorces. Les titres (sans les parties anglaises) ont été ajoutés dans les figures où le titre a été détecté dans la fenêtre graphique délimitée par l'UAR *Persée* en utilisant la méthode mixte précédemment explicitée. Concernant la première démarche utilisée pour les titres inférieurs ou égaux à 5 mots, le seuil retenu pour juger si le titre a été trouvé est de 0,75. Concernant la deuxième démarche pour les titres supérieurs à 5 mots, le seuil est de 0,7. Ces seuils ont été déterminés à l'aide des résultats obtenus lors de la troisième et quatrième exploration, mais le nombre de cas et d'exceptions n'a pas permis de déterminations aussi précises que celles menées pour les titres (*cf.* section Chap5.IV.2).

Il n'y a pas eu de traitement manuel ajouté pour les figures.

12. Fins de documents

a. Explorations

La première exploration menée précédemment pour les bibliographies (*cf.* section Chap5.IV.8) a montré que plusieurs documents de *L'Espace Géographique* se terminent par la présence de publicités ou de recommandations aux auteurs. À la suite de cette exploration, les traitements réalisés ont permis de retirer ces éléments pour l'ensemble des documents contenant une bibliographie. Toutefois, sur tous les documents n'ayant pas de bibliographie, certains peuvent présenter également des publicités ou des recommandations aux auteurs et n'ont de fait pas été traités. L'observation des cas trouvés à partir des résultats concernant les bibliographies a permis d'identifier une structure commune de ces éléments. Alors que le texte scientifique se caractérise dans *L'Espace Géographique* par une mise en page sur deux colonnes, ces éléments particuliers se différencient très souvent par leur extension sur une seule colonne. Les lignes correspondant à ces parties se retrouvent centrées sur la page. Il a par conséquent été décidé de rechercher tous les documents sans bibliographie présentant sur leur dernière page au moins deux lignes d'affilée centrées par rapport à la page. Le résultat obtenu (**@ArticleEspLignesCentrFin**) permet de détecter plusieurs contenus qui méritent en effet d'être retirés des textes à analyser, car relevant de publicités ou de recommandation aux auteurs.

Toutefois, ce résultat présente aussi plusieurs cas de faux positifs qu'il n'est pas pertinent de supprimer des textes à analyser, car ils relèvent du contenu scientifique des documents. Comme pour les notes de bas de page, j'ai décidé de coupler cette recherche d'une structure spatiale particulière à une détermination faisant également appel au contenu textuel. En effet, il peut être observé à partir des résultats obtenus que les éléments recherchés peuvent être aussi caractérisés à l'aide d'un petit ensemble de termes spécifique : « GIP », « collection », « recommandation », « Colloque »… Après avoir recensé ces derniers, la technique mise au point pour le traitement repose donc sur la recherche à la fois d'une ligne centrée sur la dernière page d'un document et sur la détection d'un de ces mots spécifiques dans la ligne en question.

Une deuxième exploration a enfin été menée suite au constat de deux mentions récurrentes, « Manuscrit reçu le… » et « Manuscrit prêt le… » en fin de document de la plupart des articles de *L'Espace Géographique* jusqu'aux années 1980. Le résultat obtenu (**@ArticleEspManuscritFin**) présente toutes les lignes commençant par « Manuscrit » dans cette revue. Ils correspondent tous aux mentions recherchées. Il n'y a pas eu de recherche

équivalente menée pour les *Annales de Géographie* car il n'y a pas eu d'observation de mentions récurrentes marquant la fin des documents.

b. Traitements automatisés

Le traitement automatisé des fins de document fait suite aux explorations menées. En effet, l'utilisateur peut opter pour la suppression des contenus situés sur la dernière page de *L'Espace Géographique* et après une ligne centrée contenant des termes spécifiques que l'utilisateur peut définir. De plus, il est possible de supprimer tous les passages de cette revue situés après une ligne commençant par le terme « Manuscrit ». Ces deux options ont été utilisées pour les deux corpus de cette recherche. La même liste de termes spécifiques a été utilisée dans les deux cas (**@ArticleTraitAutoFinDoc** pour le corpus « Article » et **@ArticlePlusCrTraitAutoFinDoc** pour le corpus « ArticlePlusCr »).

Il n'y a pas eu de traitement manuel ajouté pour les fins de documents.

Suite à cet examen méthodique des différentes parties des documents, une approche complémentaire centrée plus sur l'échelle des mots a été effectuée.

V. Améliorations réalisées grâce à l'échelle des mots

1. Travail préalable et décisif pour aborder cette nouvelle échelle

L'objectif d'un travail à l'échelle des mots est de corriger plus directement des erreurs d'OCR. Une difficulté rencontrée provient de l'absence d'une référence numérique satisfaisante quant au vocabulaire de la géographie. Il existe bien des sites s'attelant à une telle entreprise comme Hypergéo[158]. Toutefois, une prise en compte des références existantes sous forme papier montre qu'il existe une pluralité de dictionnaires de référence (Bonnamour 2004). D'une manière plus générale, un recensement exhaustif du vocabulaire de la géographie est difficilement réalisable au vu des frontières floues de cette discipline et de son évolution permanente[159]. Pour essayer de faire au mieux dans un temps raisonnable, l'ensemble des termes de cinq dictionnaires a été collecté :

[158] http://www.hypergeo.eu/ consulté le 18/09/2023.
[159] Les rééditions de plusieurs dictionnaires témoignent de ce phénomène.

- P. George et F. Verger (2013), *Dictionnaire de la géographie*, 3e édition 2009, (1ère édition 1970), Quadrige / Presses Universitaires de France, Paris.
- R. Brunet, R. Ferras et H. Théry dir (1992), *Les mots de la géographie*, GIP RECLUS et la documentation française, Montpellier-Paris.
- Y. Lacoste (2003), *De la géopolitique aux paysages : dictionnaire de la géographie*, Armand Colin, Paris.
- J. Lévy et M. Lussault dir (2003), *Dictionnaire de la géographie et de l'espace des sociétés*, Belin, Paris.
- *Hypergéo*, l'encyclopédie électronique moissonnée en 2018.

L'application permet de télécharger les termes recensés pour chaque dictionnaire en utilisant les liens suivants : **DicoGeorgeVerger**, **DicoBrunet**, **DicoLacoste**, **DicoLevyLussault** et **DicoHypergeo**.

Le travail sur les mots ne se limitant pas aux termes géographiques, des ressources électroniques assez exhaustives ont été utilisées pour avoir une base des termes français et des noms propres :

- *Morphalou*[160], Analyse et traitement informatique de la langue française, UMR 7118, 2019.
- *Prolex*[161], Laboratoire d'informatique de l'université François-Rabelais de Tours, 2018.

Pour les noms de géographes, la base constituée par la Géothèque[162] a été récupérée. Bien que cette base ne soit plus mise à jour et qu'elle soit reconnue comme incomplète, elle a été utilisée du fait de sa facilité d'accès et de l'impossibilité une nouvelle fois d'être exhaustif sur ce sujet.

Cet ensemble de ressources a servi de base de référence pour le travail mené sur les mots en réalisant une fusion de ces dictionnaires en une seule liste globale (avec une suppression des doublons et un travail spécifique concernant les pluriels des termes géographiques[163]). La présentation de ce travail débute par un examen des mots présentant des « ç », car ces derniers étaient particulièrement mal reconnus dans l'OCR de nombreux documents.

[160] https://www.ortolang.fr/market/lexicons/morphalou consulté le 18/09/2023.
[161] https://www.ortolang.fr/market/lexicons/prolex consulté le 18/09/2023.
[162] https://geotheque.org/ consulté le 18/09/2023.
[163] La base « Morphalou » intègre déjà l'ensemble des pluriels et toutes les formes conjuguées des verbes.

2. Les « ç »

a. Explorations

Si dans certains textes, les mots avec des « ç » ont été globalement bien reconnus, dans d'autres, cette reconnaissance peut être qualifiée de médiocre. Cette hétérogénéité se retrouvant au sein même d'une même revue sur des documents dont la typographie est similaire, l'explication de ce phénomène provient de la version du logiciel d'OCR utilisé par l'UAR *Persée*. Tous les documents n'ayant pas été océrisés au même moment, cela a créé des différences non négligeables au sein des textes obtenus. Quand la reconnaissance est médiocre, le c cédille est le plus souvent remplacé par un blanc. Par exemple, « français » devient une suite de deux mots : « fran » et « ais ». À partir ce constat, l'idée a été d'examiner l'ensemble des mots de tous les documents en testant si en reliant deux mots successifs par un « ç », la combinaison créée existe dans la base de référence utilisée.

Le problème est qu'il existe potentiellement des cas où il n'est pas forcément pertinent de remplacer deux mots successifs par la combinaison créée en les reliant par un « ç »[164]. La solution trouvée a résidé dans une exploration recherchant tous les mots avec une cédille dont la première partie (avant la cédille) et la seconde partie (après la cédille) sont des termes existants dans la base de référence utilisée. Le résultat obtenu (**@MotCedilleAvtApres**) montre que le nombre de cas litigieux est très limité. Ce sont surtout les termes finissants par « a » dans l'ensemble trouvé qui peuvent éventuellement poser problème[165]. Cette exploration tend à montrer que ces cas problématiques sont très marginaux. Il a tout de même été donné la possibilité d'exclure ces cas litigieux dans les traitements réalisés, car cela était particulièrement simple à effectuer après la réalisation de ce travail exploratoire.

b. Traitements automatisés et manuels

Le même traitement a été réalisé pour les deux corpus en remplaçant tous les mots successifs (mot_i, mot_i+1) par la combinaison « mot_i + 'ç' + mot_i+1 » si et seulement si cette combinaison se trouve dans la base de référence utilisée et ne fait pas partie des cas litigieux précédemment explicités.

---

[164] Par exemple, si l'OCR a mal reconnu l'accent de l'expression « mena à », la transformation en « menaça » n'est pas opportune.

[165] Par exemple, « menaça » peut être décomposé en « mena » et « a » dont la présence peut s'imaginer avec une faute d'OCR sur la reconnaissance du « à » sur l'expression « mena à ».

3. Mots avec un appel de note

Les appels de note de bas de page, quand ils sont reconnus par l'OCR, se traduisent par le mot concerné et le chiffre de l'appel de note directement accolé. Ce phénomène pose problème, car certains textes sont composés de nombreuses notes et certains termes sont plus fréquemment que d'autres associés à une note de bas de page. Il a donc été décidé de passer en revue automatiquement l'ensemble des mots situés sur une page présentant au moins une note de bas de page. Tous ces termes ont été testés quand ils finissent par un chiffre ou deux chiffres pour voir si leur forme réduite (sans ce ou ces chiffre(s)) est dans la base de référence utilisée. Dans ce cas, ils ont été remplacés par leur forme réduite dans les textes à analyser.

4. Mots coupés par une fin de ligne

Sur ce sujet de coupure des mots en fin de ligne, les textes présentent comme pour les cédilles de fortes disparités en fonction de la version du logiciel d'OCR utilisé. Il a été essayé de reconstituer ces fins de lignes en suivant la même technique que celle utilisée pour les cédilles, c'est-à-dire en regardant si l'association de deux termes (le premier en fin de ligne et le deuxième au début de la ligne d'après) forme un mot présent dans la base de référence utilisée et s'il ne fait pas partie d'une liste de mots litigieux. Cette liste a été établie en cherchant tous les termes dans la base de référence utilisée pouvant être constitués par l'association de deux termes de cette même base de référence. Un travail a également été réalisé pour normaliser certaines formes. Par exemple, si l'expression « l'espace » est coupée (« l'es » - « pace »), elle ne peut pas être trouvée, car « l'espace » n'appartient pas à la base de référence utilisée.

La phase de traitement réalise, pour faire face à cette difficulté et à d'autres, un ensemble de normalisations correspondant aux cas suivants : mot commençant par « l' » ou « L' », mot tout en majuscule ou avec une première lettre seulement en majuscule, mot commençant ou finissant par une parenthèse ou des guillemets, mot finissant par une virgule ou un point, ainsi que les mots ayant un tiret d'ajouté entre les deux formes. Les deux corpus ont été traités de la même manière en choisissant d'effectuer ces normalisations avant la détection.

Enfin, pour terminer l'explicitation de ces travaux d'amélioration, certains ont été envisagés sans être réalisés. Ils sont tout de même évoqués dans la partie suivante, car ils représentent des pistes concrètes qu'il aurait été pertinent d'envisager plus tôt dans ce travail. De plus, ils peuvent éventuellement servir à l'avenir pour des chercheurs qui voudraient également travailler à partir de sources provenant de l'UAR *Persée*.

VI. Améliorations envisagées, mais non réalisées

Il est nécessaire de reconnaître que dans les textes l'hétérogénéité liée à des versions différentes du logiciel d'OCR n'est que partiellement corrigée par les traitements précédents : il existe d'autres éléments présentant des disparités, comme la reconnaissance de la ponctuation ou de mots spécialisés, qui n'ont pas été corrigés. Ces constats invitent à penser qu'une réocérisation avec un transfert des données de documentation déjà existantes est une solution pertinente à envisager. Une demande a été effectuée à l'UAR *Persée* dans ce sens, mais leur chaîne de production ne permettait pas une telle opération au moment de cette demande. J'ai par conséquent envisagé de réaliser cette action de réocérisation à l'aide des services de l'infrastructure de recherche *Huma-Num*. Plusieurs difficultés ont été rencontrées et expliquent l'inachèvement de ce travail.

Tout d'abord, l'UAR *Persée*, pour générer les documents PDF, s'appuie sur des images enregistrées en 200 dpi. Or, lors de la numérisation et de l'océrisation, les images utilisées sont en 400 dpi. Cette meilleure qualité permet d'améliorer significativement les résultats de l'OCR. Une fois ce travail effectué, les images en 400 dpi ne sont pas gardées par l'UAR *Persée,* mais sont stockées au Centre Informatique National de l'Enseignement Supérieur (*CINES*) pour des raisons de sécurité. Il est bien sûr possible de récupérer ces documents, mais cette action ne s'est pas révélée aussi simple que prévu. La demande effectuée est restée sans suite.

Un autre problème vient du processus d'océrisation qui réalise un ensemble de prétraitements destiné à améliorer la reconnaissance. Parmi ces derniers, il y a le redressement du texte si la page a été numérisée de travers. En refaisant un OCR, le redressement n'est pas exactement le même et par conséquent les coordonnées graphiques des mots peuvent présenter des petites variations. Or, une grande partie du travail d'amélioration qui a été effectué dans cette thèse dépend de ces coordonnées graphiques[166]. Trois solutions sont envisageables : soit réaligner les nouvelles coordonnées graphiques des mots sur les anciennes en écrivant un algorithme adapté à cette tâche[167] ; soit ne rien faire, mais il faudrait dans ce cas montrer que le type d'erreur venant d'être explicité est marginal ; soit récupérer les coefficients de redressement de chaque page et réaliser ce prétraitement avant l'océrisation. La troisième solution semblant *a priori* la plus simple et la plus sûre, une

[166] C'est surtout l'utilisation des coordonnées graphiques des fenêtres tracées par l'UAR *Persée* lors de la phase de documentation infrapaginaire qui a été mise en avant dans le travail présenté. Toutefois, pour récupérer le contenu textuel de ces fenêtres, il faut s'appuyer sur les coordonnées graphiques des mots. Si ces dernières changent, des erreurs peuvent apparaître. Certains mots peuvent se retrouver hors de leur fenêtre d'origine.

[167] Une librairie python comme *Difflib* pouvant servir comme base pour trouver les séquences de mots similaires et aider aux alignements entre les anciens et nouveaux mots.

demande a été effectuée à l'UAR *Persée* sur ce sujet. Là encore, cette demande est restée sans suite.

L'échec de ces deux demandes s'explique par le fait qu'elles ont été effectuées au moment de la refonte de la chaîne de production de l'UAR *Persée* qui avait à ce moment-là des priorités plus importantes à gérer. De plus, de mon côté, il a fallu boucler ce long travail d'amélioration des données pour envisager la production de résultats ainsi que leurs analyses et interprétations. Ces raisons expliquent pourquoi ces demandes n'ont pas été renouvelées et le choix d'abandonner ces pistes de travail.

VII. Synthèse méthodologique et réflexive

L'application permet d'accéder pour chaque corpus à un récapitulatif des choix effectués pour chaque traitement automatisé et à l'ensemble des traitements manuels réalisés (**@ArticleSynthese** et **@ArticlePlusCrSynthese**). De telles synthèses se sont révélées particulièrement utiles lors de l'effectuation des travaux d'amélioration. Elles ont joué le rôle de tableaux de bord permettant de savoir quelles étapes avaient été réalisées et avec quels paramètres. De plus, une fois ces travaux d'amélioration terminés, ces visualisations permettent de ne pas oublier et de rappeler le caractère construit des corpus ainsi que la complexité de ces constructions. Il faut ici souligner qu'il n'est pas possible dans la plupart des publications d'effectuer de tels développements, mais seulement de résumer les actions réalisées de la manière suivante :

Le travail d'annotation produit en amont par l'UAR *Persée* a été utilisé pour sélectionner les corps des textes. Les notes de bas de page, bibliographies, annexes, résumés, mots-clés et contenus des figures ont été exclus des textes à analyser. Un effort a été effectué pour retirer les termes annonçant ces éléments spécifiques. Les titres des figures ont été conservés en ôtant également les termes les annonçant (« Figure », « Fig »…). Un travail d'identification des annotations de haut de page, de bas de page, de publicités et de certaines notes biographiques a été effectué afin de supprimer ces éléments des textes à analyser. Enfin, tout un travail a été mené à l'échelle des mots pour corriger certaines erreurs d'OCR.

Un tel résumé, s'il reste exact dans ses grandes lignes, supprime toutes les difficultés rencontrées et tous les compromis et solutions trouvés. Ce qui a été effectué dans cette partie peut être alors lu comme une remise en cause d'un effet de « plain-pied » couramment pratiqué par le travail quantitatif. Au-delà de la dimension relevant de la déconstruction et d'une prise de recul utile sur les documents utilisés, ces travaux d'amélioration ont alimenté ainsi toute une réflexion par rapport à la thèse du plain-pied (Orain 2003).

Les développements menés pour produire l'application permettent en effet de mettre en avant de manière détaillée le processus de construction des corpus tout en montrant à divers endroits ses limites. Plusieurs moments/décisions faisant appel à une certaine subjectivité s'expliquent à la fois par un temps limité, mais aussi par l'impossibilité d'une objectivation totale. Le travail mené est évidemment perfectible, mais des justifications complètes de tous les choix successifs opérés sont particulièrement difficiles à tenir, à partir du moment où le chercheur rentre dans les détails. Il faut souligner qu'il existe toujours une certaine tension à révéler ces moments où la subjectivité du chercheur se trouve mobilisée à l'intérieur d'une démarche quantitative. La réalisation d'une application favorise un tel processus en forçant à rendre visible et à justifier toutes les étapes. Toutefois, la construction d'une application n'entraîne pas automatiquement la mise en avant des moments de subjectivité. Leur construction comme des moments de recherche plutôt que leur minimisation, voire leur invisibilisation, dépend avant tout de la volonté du chercheur.

Par rapport aux travaux d'amélioration ici menés, il est certain que certaines parties peuvent être jugées *a posteriori* par d'autres chercheurs comme étant anecdotiques et non nécessaires. Par exemple, il est possible de penser que la partie corrigeant les mots avec les « ç » ne conduit qu'à des améliorations négligeables compte tenu du poids réduit de ces occurrences par rapport à l'ensemble des mots les plus utilisés. C'est pourtant l'erreur qui ressortait le plus de premiers résultats statistiques que j'ai obtenus. C'est suite à ces résultats que cette partie a été ajoutée. De plus, ce qui peut être parfois un détail d'un point de vue quantitatif, peut se révéler loin d'être négligeable dans la critique et la compréhension induite des sources.

Enfin, une analyse globale des difficultés rencontrées permet de distinguer plusieurs niveaux dans l'origine des problèmes. Premièrement, au niveau des revues en elles-mêmes. Il existe certes de grandes formes qui se répètent (article, compte-rendu), mais ces formes ne sont pas entièrement formatées. Il subsiste toujours des variations, des exceptions et des évolutions chronologiques. Puis, le deuxième niveau de difficulté provient du travail de documentation réalisé par l'UAR *Persée.* Même si ce travail est globalement d'une bonne qualité, certains manques par rapport à l'objectif poursuivi dans cette thèse et certaines erreurs ont pu être identifiés. Enfin, le troisième niveau est constitué par les premiers formats de fichier acquis avec des données manquantes par rapport à l'ensemble réellement existant et quelques complexités sur certains points spécifiques (par exemple les bibliographies).

Ce travail d'amélioration, et en particulier celles envisagées, mais non réalisées (*cf.* section Chap5.VI), permet de saisir qu'il existe au fur et à mesure des avancées une remontée progressive dans la chaîne de production de l'UAR *Persée*. Les difficultés rencontrées sur

les mots-clés (*cf.* section Chap5.IV.4) m'avaient déjà conduit au-delà de mes sources originellement utilisées (les fichiers XML-TEI et Érudit) en me faisant découvrir une partie des bases de données que l'UAR *Persée* obtient directement en sortie de sa phase de documentation infrapaginaire. Le travail envisagé, mais non réalisé (*cf.* section Chap5.VI) propose une remontée jusqu'à l'océrisation elle-même. La volonté de construire des données de plus en plus adéquates mène ainsi à une déconstruction des données initialement obtenues de plus en plus approfondie.

Ce processus rejoint les réflexions d'Andreas Fickers à propos d'une actualisation de la pensée critique liée à l'ère numérique : « Pour mener une critique des sources prétendant à une validité scientifique, il nous faut comprendre comment les données sont codées, indexées et enrichies de métadonnées. Sans critique des sources numériques, nous abandonnons une compétence clé du travail des historiens et historiennes »[168]. Dans cette optique, ce travail peut être vu comme un exemple concret illustrant la complexité que ce processus peut recouvrir.

La dernière étape a été l'écriture d'une fonction qui permet de prendre en compte l'ensemble des améliorations stockées dans la base de données pour tous les documents d'un corpus d'étude et de produire les textes correspondants. Cette fonction[169] n'est accessible que pour les utilisateurs ayant des droits sur un corpus.

Après avoir réalisé toutes ces étapes de délimitations et d'améliorations des textes, plusieurs opérations ont encore été nécessaires pour finaliser les deux corpus.

---

[168] Andreas Fickers, *Entre altérité et familiarité : pour une herméneutique numérique en sciences historiques.* Conférence dans le cadre du cycle « Les jeudis de l'Institut historique allemand », IHA, Paris, (25 avril 2019). Le texte cité provient de la présentation de conférence https://dhiha.hypotheses.org/2611 publié le 18 mars 2019, consulté le 31 juillet 2019.

[169] Disponible dans le fichier « Interface/AllApps/PreTraitement/Persee/AmeliorText/views.py » sous le nom « Extract ».

Chapitre 6 :
Finalisation et évaluation des corpus d'étude

I. Finalisation des corpus d'étude

L'étape de finalisation des corpus est constituée de deux moments :

- L'ajout et le retrait de documents par rapport aux corpus jusqu'ici construits.
- Les choix concernant un ensemble de prétraitements terminaux : segmentation, lemmatisation et suppression de mots-outils.

1. Ajout et retrait de documents

Cette phase de travail est due tout d'abord à la volonté d'intégrer aux corpus les quelques documents dont l'UAR *Persée* a limité l'accès (*cf.* section Chap4.III.6). Ces documents ont été scannés et océrisés avec l'aide de la MSH Lyon Saint-Étienne, puis corrigés et compilés dans deux fichiers : un premier pour compléter le corpus « Article »[170], un deuxième pour le corpus « ArticlePlusCr »[171]. Les textes obtenus sont de meilleure qualité que ceux issus de l'UAR *Persée,* car ils ont été océrisés avec un logiciel plus récent et corrigés manuellement[172]. La présence de plusieurs articles de deux auteurs mobilisés par Olivier Orain dans sa thèse, Antoine S. Bailly et Jean-Bernard Racine, a joué un rôle dans le choix d'intégrer ces textes malgré l'hétérogénéité constatée. Pour rester cohérent par rapport à la construction précédemment menée (*cf.* section Chap2.II.5.a)[173], l'ajout ne s'est pas limité à ces deux auteurs. Il a concerné l'ensemble des documents qui auraient fait partie du corpus sans ces limitations d'accès.

Symétriquement, le retrait de certains documents s'est imposé après l'observation d'articles en anglais dans les corpus créés malgré le travail précédemment effectué sur la métadonnée 'Langue' (*cf.* section Chap4.III.5). Une expérimentation a été réalisée en soumettant les textes précédemment obtenus à un algorithme de détection de langue existant en *Python*[174]. Les résultats suivants (**@ArticleDetectLang** et **@ArticlePlusCrDetectLang**) présentent tous les cas où la langue détectée par l'algorithme n'est pas le français. Une vérification manuelle a permis de confirmer que les résultats obtenus étaient exacts. Il y a donc des erreurs sur la métadonnée « Langue » de ces documents. À la suite de cette

[170] Ce premier fichier forme le corpus additionnel nommé « ArticleAdd ». Il est constitué des articles suivants : **@ArticleAdd**.

[171] Ce deuxième fichier forme le corpus additionnel nommé « ArticlePlusCrAdd ». Il est constitué des articles et comptes-rendus suivants : **@ArticlePlusCrAdd**.

[172] Ces corrections manuelles sont dues à la difficulté rencontrée pour scanner correctement du fait des reliures. Les textes proches de cette dernière tendent à être mal reconnus ensuite par l'OCR. L'UAR *Persée* découpe les revues scannées pour faire face à cette difficulté. Cette action n'était pas envisageable dans le cadre de mon travail.

[173] Avec notamment la volonté de ne pas sélectionner des textes en fonction d'auteurs spécifiques.

[174] https://pypi.org/project/langdetect/ consulté le 18/09/2023.

exploration, ils ont été retirés des corpus créés. Les deux corpus « ArticleComplet » et « ArticlePlusCrComplet » ont été ainsi constitués[175].

Après cette étape, de nombreuses questions se sont posées concernant des prétraitements classiques (mais toujours discutés) dans les pratiques d'analyse textuelle : niveau de segmentation, lemmatisation, suppression des mots-outils.

2. Choix, améliorations et limites des prétraitements effectués

a. Réflexions et choix des prétraitements

Pour chaque prétraitement, le choix à réaliser ne se résume pas à une alternative entre deux possibilités, mais présente de multiples variantes. Par exemple, la segmentation des textes en unités distinctes (servant ensuite comme base dans les traitements) est réalisable à différents niveaux : la forme graphique[176], le mot, le mot composé (avec des différences existantes de degrés entre ceux se limitant aux composés à apostrophe[177] et à trait d'union[178] et ceux prenant aussi en compte des composés détachés[179]) ou encore la sélection des n-gram[180] les plus fréquents… Au niveau de la lemmatisation[181], il existe une controverse depuis les années 1980 sur la pertinence de ce prétraitement (Lafon 1984; Brunet 1999). Aucun consensus n'a été trouvé et une diversité de pratiques continue d'exister en la matière. Il faut souligner que pour ceux choisissant la lemmatisation se pose la problématique du choix du lemmatiseur puisqu'il en existe de nombreux (*TreeTagger*[182], *Cordial*[183]…). Le choix le plus radical d'une racinisation[184] peut également être envisagé. Enfin, au niveau des

[175] Les choix réalisés dans l'application peuvent être consultés pour ces deux corpus avec les liens suivants : **@ArticleComplet** et **@ArticlePlusCrComplet.**

[176] « Aujourd'hui » est un mot composé de deux formes graphiques.

[177] « Aujourd'hui » est un cas de mot composé à apostrophe.

[178] « Après-midi » est un cas de mot composé à trait d'union.

[179] « Pomme de terre » est un cas de mot composé détaché. La limite est alors assez subjective entre un simple syntagme et un mot composé. Par exemple, « carte géographique » est-il un mot composé ?

[180] Une séquence continue de n-mots.

[181] La lemmatisation consiste à transformer chaque forme en un lemme qui est le plus souvent la « forme canonique » enregistrée dans le dictionnaire. Par exemple, la forme « espaces » est transformée en « espace ». La forme « allons » en « aller ».

[182] https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (consulté le 18/09/2023) qui est intégré notamment au logiciel *TXM*.

[183] https://www.atala.org/content/cordial-universit%C3%A9s-ou-cordial-analyseur (consulté le 18/09/2023) qui est intégré au logiciel *Hyperbase*.

[184] Chaque terme est transformé en sa racine. Par exemple, le verbe « chercher » en « cherch ». Le nom « chercheur » est également transformé en « cherch » à la différence de la lemmatisation qui maintient une différence entre ces deux termes.

mot-outils[185], les exclusions peuvent être bien entendu différentes selon les logiciels et les choix des chercheurs.

Plusieurs recherches ont déjà traité spécifiquement de la question de l'impact de ces prétraitements classiques dans le cas de l'utilisation de plongements de mots statiques (méthodes que la partie méthodologique a conduit à privilégier pour cette recherche : *cf.* section Chap3.IV.3). Fares *et al*. (2017) obtiennent par exemple de meilleurs résultats sur des tâches de résolution d'analogies après lemmatisation. Toutefois, il n'existe pas de consensus scientifique autour de la réalisation de ces prétraitements. Par exemple, Camacho-Collados *et al.* (2017) concluent à une meilleure efficacité d'un simple découpage en « token » (unité lexicale) par rapport à l'utilisation de techniques plus complexes comme la lemmatisation ou le regroupement des n-grams. Cette opposition dans les conclusions des deux études peut s'expliquer par des perspectives et des objectifs différents[186]. Les spécificités du corpus peuvent aussi jouer comme l'indiquent Camacho-Collados *et al.* (2017) qui nuancent leur conclusion en affirmant qu'elle ne s'applique selon eux qu'à des textes ne relevant pas de domaine spécifique[187].

Ces lectures m'ont conduit à adopter une position très pragmatique consistant à penser que les prétraitements les plus adaptés dépendent avant tout des objectifs fixés et des corpus. La théorie et l'état des lieux ne permettent pas de déterminer entièrement les prétraitements à privilégier. Pour faire face à cette difficulté, j'ai construit à partir du corpus « Article » deux corpus : un lemmatisé et un non lemmatisé. Il aurait été certes possible de construire de nombreux autres corpus avec tout le panel des prétraitements possibles. Cela n'a pas été réalisé pour des raisons de temps. Il n'y a pas eu de recherche expérimentale du meilleur prétraitement possible.

Le fait de ne pas pouvoir fonder théoriquement le choix des prétraitements explique également l'importance prise par le travail précédemment mené sur l'évolution lexicale (*cf.* section Chap2.II.5.c). Ce dernier m'avait conduit à faire des prétraitements avec des découpages en unités à partir du logiciel *IRaMuTeQ* : découpage en formes graphiques et recomposition des mots composés à partir d'un dictionnaire d'expressions. La lemmatisation avait aussi été faite à partir du logiciel *IRaMuTeQ* (qui utilise pour cela un dictionnaire mots/lemmes) ainsi que la suppression des mots-outils (en ne gardant que les « mots

[185] Un mot-outil est un mot dont le rôle sémantique est secondaire. Son rôle est avant tout syntaxique. C'est le cas par exemple de la majorité des prépositions, conjonctions, pronoms et déterminants.

[186] Catégorisation de textes et analyse de sentiments pour Camacho-Collados *et al*. (2017) / Résolutions d'analogies et de similarités sémantiques pour Fares *et al.* (2017).

[187] Camacho-Collados *et al.* (2017) citent l'exemple d'un corpus de textes relevant du domaine médical sur lequel une lemmatisation et un regroupement en n-gram est plus efficace qu'un simple découpage en « token ».

pleins » : noms, verbes, adjectifs et adverbes). Ces prétraitements avaient été adoptés essentiellement pour trois raisons :

- La première relève d'un objectif d'homogénéisation des textes. En effet, comme cela a été précédemment explicité dans le chapitre sur l'amélioration des textes (*cf.* Chap5), il reste une hétérogénéité liée à des qualités inégales d'océrisation. Ces prétraitements permettent de lisser une partie de ces différences. Si la ponctuation et les mots-outils sont éliminés, toutes les erreurs liées à ces contenus spécifiques le sont aussi. Cette technique permet donc de produire une homogénéisation des textes.
- Le choix de la lemmatisation est aussi lié à une problématique de visualisation des résultats. Devant la contrainte de ne pouvoir faire apparaitre qu'un nombre limité de termes sur une figure, il m'a très souvent semblé plus intéressant de faire apparaitre des lemmes différents plutôt que plusieurs termes pouvant être proches, car appartenant au même lemme[188]. Cette réflexion explique pourquoi un statut privilégié a été accordé au corpus lemmatisé. Le corpus non lemmatisé sert de point de comparaison pour évaluer si les résultats obtenus dépendent ou non d'un tel prétraitement.
- Plus spécifiquement, le choix du logiciel *IRaMuTeQ* avait été effectué après une comparaison avec le lemmatiseur *TreeTagger*. Ce dernier peut être facilement appelé à partir d'un programme *Python* mais ne peut pas être facilement ouvert pour analyser son fonctionnement et corriger des erreurs. À l'opposé, le système de dictionnaires[189] utilisé par le logiciel *IRaMuTeQ* laisse la possibilité de facilement corriger la lemmatisation si nécessaire. Cette potentialité a paru d'autant plus intéressante ici qu'elle permettait de réinvestir une partie du travail sur les dictionnaires précédemment réalisé (*cf.* section Chap5.V.1).

b. Amélioration des dictionnaires et des prétraitements

Les mots extraits des différents dictionnaires de géographie et les noms propres issus du dictionnaire *Prolex* ont été ajoutés au lexique et aux expressions françaises proposés par défaut dans le logiciel *IRaMuTeQ*. Pour les expressions, ce sont les mots composés à trait d'union qui ont été retenus. Ce choix s'inscrit dans la continuité du dictionnaire proposé par

[188] Par exemple, constater qu'« espace » et « espaces » sont proches sur une AFC a été jugé d'un intérêt moindre par rapport au fait de pouvoir compléter la figure par un autre lemme.
[189] Dictionnaire mots/lemmes et dictionnaire d'expressions.

le logiciel *IRaMuTeQ* puisque 92 % des expressions consignées dans ce dictionnaire initial sont des mots composés à trait d'union. À partir des dictionnaires obtenus (**@IraMotsLemmesAmelior1** et **@IraExpressionsAmelior1**), une expérimentation a été réalisée pour voir quelles formes graphiques présentes dans les textes n'étaient pas reconnues, car absentes des dictionnaires. Ces formes non reconnues ont été ensuite classées par ordre décroissant de fréquence. Le résultat obtenu (**@FNR1**) permet d'identifier différents cas de formes non reconnues : des chiffres, des mots coupés (malgré le travail réalisé sur les fins de ligne), des erreurs d'OCR et des mots existants, mais qui ne sont pas recensés dans les dictionnaires utilisés. Ces identifications ont eu pour conséquence plusieurs actions :

- Les dictionnaires ont été complétés pour prendre en compte une partie des mots non reconnus et les transformer de manière adéquate en lemmes[190]. Deux nouveaux dictionnaires ont été ainsi obtenus : **@IraMotsLemmesAmelior2** et **@IraExpressionsAmelior2**. La possibilité d'ajouter des dictionnaires enrichis mots/lemmes et d'expressions a été implémentée dans l'application.
- Une liste de « fin de mots » a été établie pour essayer de reconstituer une partie des termes coupés. En effet, à partir du recensement précédent des formes non reconnues, il est assez facile de déterminer celles correspondant à des fins de mots[191]. La liste ainsi construite (**@FinMots1**) peut être ensuite utilisée pour reconstituer les mots. Lors de l'établissement de cette liste de « fin de mots », l'objectif a été de sélectionner des formes n'étant pas également des mots pour ne pas créer des agrégations erronées. La seule exception qui a été réalisée correspond à la forme « ment ». En effet, cette dernière peut renvoyer au verbe « mentir » conjugué. Toutefois, l'examen des textes a montré que la majorité des cas correspond à la fin d'un adverbe coupé. Cette forme a été par conséquent intégrée à la liste de « fin de mots » même si elle peut provoquer quelques agrégations erronées[192].
- Au niveau des quelques erreurs d'OCR qui apparaissent dans les formes non reconnues les plus fréquentes, il est possible d'en corriger une grande partie par l'ajout de la forme corrigée dans le dictionnaire[193]. Ce travail s'est fait à partir de la liste précédemment construite en allant jusqu'aux formes non reconnues

[190] Ce travail a été effectué jusqu'aux formes non reconnues apparaissant au moins 15 fois dans l'ensemble du corpus « ArticleComplet ».

[191] La forme non reconnue « tions » par exemple est une fin de mot puisqu'il n'existe pas en français de terme correspondant à cette forme.

[192] Par exemple, une phrase comme « La carte ment » est automatiquement transformée en « La cartement ».

[193] Par exemple, si le premier accent d'épistémologie est souvent mal reconnu, il suffit d'ajouter une forme « epistémologie » qui renvoie au lemme « épistémologie ».

apparaissant au maximum 15 fois dans l'ensemble de corpus « ArticleComplet » pour des raisons de temps.

- Enfin, le choix d'exclure les chiffres dans les textes finaux à analyser a été retenu, car ils ont un intérêt mineur dans cette recherche.

Une fois ces décisions prises et ces actions réalisées, la partie de l'application pour produire ces prétraitements a été construite.

c. Réalisation et limite des prétraitements effectués

En plus du choix des dictionnaires (mots/lemmes, expression) et de la liste de « fin de mots », l'application permet de choisir d'effectuer (ou non) un premier ensemble de transformations. Les textes sont alors convertis entièrement en minuscule. Les caractères non conventionnels sont enlevés. La liste de « fin de mots » est utilisée pour reconstituer les mots coupés. Les expressions sont identifiées à l'aide du dictionnaire correspondant.

Le second grand choix que permet à ce stade l'application porte sur la réalisation (ou non) de la lemmatisation. Concernant le corpus d'articles non lemmatisés, intitulé « ArticleNonLem », le choix réalisé a été d'effectuer tout de même le premier ensemble de transformations. La différence avec le corpus d'articles lemmatisés, intitulé « ArticleLem » porte ainsi que sur la lemmatisation[194]. Les options utilisées pour la construction de chaque corpus sont accessibles dans l'application à partir des liens suivants : **@ArticleLem**, **@ArticleNonLem** et **@ArticlePlusCrLem**.

Par rapport à la lemmatisation, il faut reconnaître que la méthode utilisée par le logiciel *IRaMuTeQ* n'est pas optimale du fait du mécanisme assez basique de dictionnaire qui attribue un lemme à un mot sans tenir compte du contexte. Par exemple, le mot « porte » peut renvoyer au lemme « porte » correspondant au nom commun, mais aussi au lemme « porter » suivant les situations. Le fonctionnement du logiciel *IRaMuTeQ* privilégie une seule solution, toujours la même, ce qui engendre inévitablement quelques erreurs. Il n'y a pas eu d'amélioration implémentée pour faire face à ce problème pour des raisons de temps. Il s'agit évidemment d'un axe d'amélioration possible par rapport au travail mené dans le cadre de cette thèse. Comme cela a été précédemment dit, la solution adoptée n'a pas la prétention d'être optimale. Sa simplicité a permis de produire des améliorations, mais ces dernières restent limitées.

[194] Dans le cas de cette recherche, la lemmatisation implique aussi l'enlèvement des mots-outils (pronoms, prépositions, articles, conjonctions…) et des formes non reconnues.

Suite à cette étape, les corpus définitifs sont construits. Une évaluation a été alors effectuée pour avoir un aperçu plus précis de la qualité des données utilisées.

II. Évaluation des corpus d'étude

Pour chaque corpus, une méthode automatique de sélection au hasard de 20 documents, puis de 100 formes[195] consécutives pour chacun de ces documents, a été mise en place. Ces quantités, 20 et 100, n'ont pas été déterminées à la suite d'un calcul pour construire un échantillon représentatif. Elles résultent de l'objectif de réaliser ces évaluations dans un temps raisonnable. Par rapport à la taille des corpus obtenus (« ArticleLem » : 3223 textes et 10 280 408 lemmes, « ArticleNonLem » : 3223 textes et 19 619 153 mots et formes de ponctuations, « ArticlePlusCrLem » : 9701 textes et 14 453 978 lemmes), les parties évaluées peuvent être considérées comme assez réduites. En effet, s'il m'a semblé nécessaire de réaliser une évaluation pour chaque corpus, il ne m'a pas semblé indispensable de réaliser un travail exhaustif sur ce sujet. Au-delà du problème du temps disponible et des priorités qu'il est nécessaire de fixer dans tout travail de recherche, ce choix s'explique aussi par la problématique de la multiplicité des mesures possibles pour une telle évaluation.

Le fait de reconnaître qu'il s'agit d'une évaluation réduite permet de mieux justifier l'utilisation d'une seule mesure : le taux d'erreur de mots (WER[196]). Le taux d'erreur de caractères (CER[197]) qui est souvent aussi mis en avant pour évaluer des systèmes d'OCR (Margner, Pal, et Antonacopoulos 2018) n'a pas été retenu. Ce choix s'explique par la finalité de cette recherche. Les évolutions sémantiques, qu'il s'agit de caractériser, se jouent à l'échelle des mots (ou lemmes) et non des caractères. Par conséquent, une mesure prenant comme référence cette unité de base du mot (ou du lemme suivant les corpus) a été privilégiée.

Le taux d'erreur de mots est défini par la formule suivante :

$$\mathrm{WER}=\frac{\mathrm{S+D+I}}{N}$$

[195] Mots, expressions et ponctuations pour les corpus non lemmatisés. Lemmes et expressions pour les corpus lemmatisés.
[196] *Word Error Rate* en anglais.
[197] *Caracter Error Rate* en anglais.

Où S est le nombre de substitutions (mots incorrectement reconnus), D est le nombre de suppressions (mots omis) et I le nombre d'insertions (mots ajoutés) par rapport à un texte de référence (constitué de N mots). Les textes de référence sont constitués en réalisant l'ensemble du travail d'amélioration manuellement. Plus le taux d'erreur de mots est proche de 0, plus les textes évalués sont considérés de qualité.

Les résultats obtenus pour chaque corpus sont les suivants :

	Substitutions	Suppressions	Insertions	Nombre de mots du texte de référence	WER
ArticleLem	64	54	36	2027	0,08
ArticleNonLem	35	127	36	2093	0,09
ArticlePlusCrLem	56	75	23	2052	0,08

Tableau n°4 : Calcul du taux d'erreur de mots (WER) pour chaque corpus.

Les taux d'erreur de mots obtenus pour les trois corpus sont très proches. Globalement, il existe moins d'une erreur (substitution, suppression ou insertion) toutes les dix formes évaluées. La qualité un peu supérieure en valeur absolue des corpus lemmatisés provient d'un nombre moindre de mots omis (suppression). En effet, plusieurs documents non lemmatisés présentent plusieurs formes de ponctuation ou/et des petits articles (à, de, l'…) oubliés par l'OCR. Dans le détail[198], il y a une certaine hétérogénéité entre des textes sans erreur (ou presque) et d'autres qui peuvent présenter jusqu'à deux fois plus d'erreurs que la moyenne. La lemmatisation permet de réduire une partie de cette hétérogénéité en supprimant la ponctuation et ces petits articles. Toutefois, elle est à l'origine aussi de nouvelles erreurs.

Une quantification des causes d'erreur a été réalisée. Trois causes principales ont été identifiées : délimitation[199], océrisation, lemmatisation. Une catégorie « autre » a été ajoutée pour prendre en compte des cas non attendus.

[198] Les évaluations détaillées sont consultables à partir des liens suivants : **@ArticleLemEval1**, **@ArticleNonLemEval1** et **@ArticlePlusCrLemEval1**. Le dernier item pour chaque passage est un lien permettant de consulter l'évaluation détaillée du passage en question.

[199] La délimitation renvoie au fait que des parties de textes peuvent être présentes alors qu'elles ne devraient pas l'être (ou inversement). Par exemple, si le retrait automatique des publicités, des hauts de page, des bas de page a échoué, les parties correspondantes sont liées à des erreurs de délimitation. Cette erreur peut aussi toucher des documents entiers si ces derniers se révèlent être dans un corpus alors qu'il ne devrait pas.

	Délimitation	Océrisation	Lemmatisation	Autre
ArticleLem	26	30	95	2
ArticleNonLem	38	157	0	3
ArticlePlusCrLem	4	29	106	14

Tableau n°5 : Causes des erreurs détectées pour chaque corpus.

Les erreurs de délimitation sont essentiellement dues à deux raisons. La première est liée à des légendes de figures qui sont situées sous l'image après le titre et qui n'ont pas été intégrées dans les fenêtres graphiques tracées par l'UAR *Persée*. Elles se retrouvent alors dans les textes à analyser. Ceci est d'autant plus gênant qu'il y a une hétérogénéité de traitement par rapport à des figures où la légende se trouve dans les fenêtres graphiques tracées par l'UAR *Persée*. Ce problème n'est pas simple à résoudre, car il nécessite de mettre au point des méthodes de détection automatique des légendes. Une deuxième cause des erreurs de délimitation identifiées lors des évaluations menées provient de hauts de page non reconnus. Il s'agit de trois cas sur l'ensemble des documents analysés. La méthode mise au point n'est pas parfaite, mais elle fonctionne tout de même de manière satisfaisante comme le montrent aussi plusieurs retraits réussis de hauts de page sur les passages évalués.

Les erreurs d'océrisation concernent surtout la ponctuation et des petits articles comme le montre la différence entre les corpus lemmatisés et non lemmatisés. La reconstitution des mots coupés en fin de phrase n'est pas parfaite, mais permet de corriger plusieurs erreurs sur les passages évalués. Si l'opération de lemmatisation réduit l'hétérogénéité des documents par rapport aux erreurs d'océrisation, elle ne la supprime pas. Ces erreurs d'océrisation ont un impact sur les erreurs de lemmatisation qu'elles tendent à augmenter.

Par rapport à ces dernières, plusieurs cas doivent être distingués. Il y a tout d'abord un grand nombre d'erreurs de lemmatisation liées aux formes non reconnues. Si le texte présente des mots peu fréquents et mal référencés dans les dictionnaires utilisés, il y a alors beaucoup de fautes de lemmatisation. L'importance de ces erreurs peut être relativisée en pensant que par rapport à la problématique de détection de changement sémantique majeur, ces mots rares jouent un rôle assez secondaire. Quelques erreurs de lemmatisation apparaissent toutefois comme problématiques. Par exemple, quand le terme « sens » est transformé en verbe « sentir » alors qu'il est utilisé dans sa forme nominale. Même si la

lemmatisation a déjà été reconnue comme imparfaite, cette évaluation vient confirmer qu'il s'agit d'un important axe d'amélioration par rapport au travail effectué.

Enfin, la catégorie « Autre » présente un score plus élevé pour le corpus avec les comptes-rendus. L'explication réside dans la découverte de répétitions de lemmes provenant de termes dans une balise dans le format « XML-TEI » qui n'avaient pas été pris en compte lors de l'extraction des mots. Il a été décidé pour rendre ce corpus de comparaison plus valide de construire un corpus « ArticlePlusCrLem2 » en retirant seulement ces répétitions erronées[200].

III. Réflexions conclusives

Cette évaluation marque la fin de cette première partie qui a permis de choisir une orientation méthodologique (les plongements de mots non contextualisés) et de préparer les données aux analyses envisagées. Au-delà de ces résultats pratiques, c'est ma conception même du travail de « préparation des données » qui a évolué. Le fait de se servir des difficultés rencontrées pour commencer à développer une critique des sources est, à mon sens, l'apport le plus important de cette partie.

De plus, en développant de manière approfondie ces deux plans (méthodes et données), il s'est produit, dans les deux cas, le développement d'un phénomène ambivalent : d'un côté, l'acquisition d'une certaine force, d'une certaine assise au sens où cette recherche peut se réfléchir et faire valoir une connaissance approfondie (aussi bien des données utilisées que du spectre de méthodes possibles par rapport à la problématique retenue). De l'autre, la reconnaissance évidente d'une certaine fragilité, au sens où les choix réalisés se doivent de reconnaitre leurs limites : les données sont imparfaites et même si les résultats de l'OCR étaient optimaux, certains choix de délimitations du corpus (intégration ou non des comptes-rendus, prise en compte des titres de figures…) n'ont pas de justifications fortes.

Par rapport à cette ambivalence, il me semble possible de mettre en avant tout autant la rupture que le maintien d'un réalisme. En effet, en reconnaissant que les données et les méthodes sont imparfaites, il existe évidemment une rupture avec un réalisme fort. Dans le même temps, tout ce travail réalisé pour l'amélioration des données et pour choisir la méthode la plus appropriée a pour objectif que les résultats ne soient pas des fictions, mais permettent un discours fondé empiriquement c'est-à-dire renvoyant à une forme de réalisme.

---

[200] La fonction « CreateCorpusArticlePlusCrNonLem2 » dans le « view.py » de « AllApps/PreTraitement/Persee/EvaluateCorpus » permet de réaliser la construction de ce corpus à partir du corpus « ArticlePlusCrLem ».

Cette idée ayant trait à la problématique d'Olivier Orain (2003), est ici simplement esquissée et méritera ultérieurement des développements plus amples. Avant cela, nous nous engageons, à partir des données précédemment obtenues et de la piste méthodologique retenue, dans la production de résultats et leurs analyses.

Deuxième partie

Exploration, construction, analyse et discussion des résultats

Chapitre 7 :
Exploration et construction des résultats

Rappelons que l'objectif poursuivi dans cette partie est d'analyser les évolutions sémantiques de plusieurs termes (notamment « espace » et « milieu ») pour mettre à l'épreuve les affirmations de la thèse d'Olivier Orain à ce propos (*cf.* section Chap2.II.3.c). L'enjeu est *in fine* d'éprouver la lecture kuhnienne de la géographie française soutenue par cet auteur. En effet, Olivier Orain défend l'idée dans sa thèse que la mise en évidence d'évolutions sémantiques dont il affirme l'existence, peut permettre d'appuyer la validité de sa lecture kuhnienne. En se basant sur une approche quantitative objectivée, il y a donc dans la confirmation (ou dans la réfutation) de ces évolutions sémantiques, un enjeu majeur pour donner un appui (ou non) à la lecture épistémologique proposée par Olivier Orain.

Toutefois, la production des résultats et leurs analyses ont été loin d'être simples et immédiates. De nombreuses expérimentations empiriques et réflexions ont été effectuées lors de cette phase de recherche. Il m'a semblé important de ne pas seulement présenter les résultats finaux, mais aussi de rendre compte du processus et des interrogations ayant participé à leur élaboration. Nous nous engageons ainsi dans la poursuite de la même perspective que celle revendiquée et suivie pour les parties précédentes.

L'état des lieux précédemment réalisé (*cf.* Chap3) a orienté le choix méthodologique vers trois techniques différentes de plongement de mots : *Word2Vec*, *Glove* et *FastText*. Sachant que les méthodes *Word2Vec* et *FastText* présentent deux variantes différentes (*CBOW* et *Skip-gram*), cinq variantes méthodologiques sont dès lors envisageables. De plus, comme chacune de ces méthodes oblige à choisir un nombre de paramètres non négligeables (taille du contexte, taille des plongements, nombre d'itérations…)[201], il existe par conséquent un grand nombre de traitements possibles. Assez rapidement, la question s'est donc posée de savoir quelle méthode et quels paramètres permettaient d'obtenir les résultats les plus intéressants.

I. Choix de la méthode et des paramètres

La stratégie suivie a été celle d'une décomposition de ce problème en deux sous-problèmes. Dans un premier temps, des paramètres ont été choisis sans qu'il y ait aucune recherche d'optimisation. Les résultats obtenus, avec ces paramètres et en utilisant successivement les 5 variantes méthodologiques, ont été ensuite comparés pour essayer de

[201] Pour les méthodes *Word2Vec* et *FastText* par exemple, l'ensemble des paramètres disponibles avec la librairie utilisée (*Gensim*) est disponible à cette adresse : https://github.com/RaRe-Technologies/gensim/blob/develop/gensim/models/word2vec.py (consulté le 18/09/2023). Pour la méthode *GloVe*, c'est l'implémentation de l'université de Standford qui a été utilisée : https://github.com/stanfordnlp/GloVe (consulté aussi le 18/09/2023).

choisir la meilleure méthode. Dans un second temps, plusieurs expérimentations ont été réalisées avec la méthode retenue en faisant varier les principaux paramètres les uns après les autres pour mieux comprendre leurs effets et tenter de choisir les valeurs optimales.

La section suivante présente et justifie en amont quelques choix réalisés avant d'aborder les deux sous-problèmes explicités.

1. Choix initial des données et des paramètres

Le corpus « ArticleLem »[202] et le terme « espace » ont été choisis pour réaliser cette première expérimentation. Le choix d'étudier en priorité l'évolution sémantique de ce terme s'explique par l'importance qu'Olivier Orain lui accorde dans la conclusion de sa thèse (*cf.* section Chap2.II.3.c).

Une réflexion a été menée pour définir le découpage temporel semblant le plus approprié pour cette première expérimentation. Sur ce sujet, il faut souligner que le chapitre *La révolution dans les textes* de la thèse d'Olivier Orain (2003) met en avant la date de 1971 du fait d'une ébauche de *L'Espace Géographique* (appelée « numéro zéro ») qui a servi à annoncer la revue à venir. Toutefois, les corpus constitués ne contenant pas ce numéro zéro[203], il a paru pertinent de choisir 1972 comme date de rupture dans le découpage temporel. Cette date se justifie pleinement du fait de la structure des corpus construits : une seule revue (Les *Annales de géographie)* compose les corpus avant 1972 ; deux revues (Les *Annales de Géographie* et *L'Espace Géographique*) à partir de cette date.

Ensuite, une rupture autour des années 1960-1961 a été privilégiée, car Olivier Orain mentionne les années 1960 comme période d'importation du terme « espace » tout en affirmant qu'il n'y a pas eu de changement sémantique durant cette décennie. La sémantique du terme « espace » est selon lui restée proche du terme « milieu ». Il m'a paru ensuite pertinent d'avoir un point de comparaison avec la période antérieure. Cette réflexion a abouti à la construction de la période 1950-1960. La date de 1950 ne correspond pas à un changement majeur. Elle n'a été retenue que pour créer une période comparable à 1961-1971.

Après 1972, le milieu des années 1980 est cité par Olivier Orain comme marquant la fin de la « phase révolutionnaire » avec « la généralisation d'une sémantique particulière sur les bases du répertoire adopté depuis déjà quelque temps » (Orain 2003, 354). La prise en

[202] Le choix de ce corpus s'explique aisément par la partie de construction des corpus (*cf.* section Chap5) qui lui a conféré une certaine primauté.

[203] Ce numéro zéro n'est formé que de l'éditorial du premier numéro à paraître, d'une plaquette de présentation et des sommaires des numéros à venir. Cette composition justifie le fait qu'il ne fasse pas partie des corpus constitués.

compte de cette réflexion a conduit à la création d'une période 1972-1984. Dans un processus symétrique à celui de la justification de la période 1950-1960 (après avoir construit 1961-1971), il m'a paru intéressant de créer la période 1985-1995 pour avoir un point de comparaison par rapport à 1972-1984. Comme précédemment pour 1950, 1995 ne correspond pas à une rupture avancée par Olivier Orain. C'est le pas de temps des autres périodes[204] qui permet de comprendre cette dernière date.

Ce découpage temporel (1950-1960,1961-1971,1972-1984,1985-1995) défini, il a fallu déterminer les paramètres utilisés pour chacune des méthodes de plongement de mots. Les travaux déjà existants ne sont pas d'une grande utilité sur la question. Ils sont nombreux et ne présentent pas une grande homogénéité. La taille des plongements varie souvent de 50 à 300. La taille du contexte et le nombre d'itérations sont aussi très variables. Un seul paramètre proposé par ces méthodes m'a semblé pouvoir être fixé à l'aide d'un argument se référant à la construction des corpus. Il s'agit du choix d'un seuil permettant de ne pas prendre en compte les mots en dessous d'un certain nombre d'occurrences. Le traitement des formes non reconnues ayant été précédemment effectué jusqu'aux formes apparaissant au moins 15 fois (*cf.* section Chap6.I.2.b), cette valeur pour ce paramètre a été choisie.

Pour les autres paramètres, les choix n'ont pas pu être justifiés aussi précisément. Par exemple, il a été privilégié dans un premier temps une taille de contexte ni trop grande, ni trop petite. C'est ainsi que la taille de 10 a été retenue pour ce paramètre. De la même manière, sans justification raffinée, le nombre d'itérations a été fixé à 100 et la taille des plongements à 50. Il s'agit là bien entendu de choix pour une première expérimentation. L'impact de ces différents paramètres sera discuté par la suite (*cf.* section Chap7.I.3).

Pour la méthode *FastText*, il est nécessaire de fixer en plus deux autres paramètres définissant la taille minimale et maximale des n-grams. Les valeurs par défaut de la librairie utilisée ont été employées pour cette première expérimentation : 3 pour la valeur minimale et 6 pour la valeur maximale. Une fois les plongements réalisés, la mesure de similarité cosinus a été utilisée pour comparer les différents vecteurs. En effet, contrairement aux autres paramètres, il y a une certaine convergence dans la littérature existante autour de l'emploi de cette mesure pour détecter les termes les plus proches d'un terme donné à partir des vecteurs obtenus.

Avant de présenter les résultats, il est nécessaire de rappeler que les plongements sont calculés indépendamment pour chaque période et revue. Il y a par conséquent des situations

[204] Une dizaine d'années à l'exception de 1972-1984 (exception justifiée par la volonté d'être proche des ruptures mises en avant par Olivier Orain).

très différentes au niveau du nombre d'occurrences du lemme « espace » dans les corpus distingués[205].

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	214	861	2195	2291
L'Espace géographique			6452	5147

Tableau n°6 : Nombre d'occurrences du lemme « espace » dans le corpus ArticleLem.

L'émergence et la montée en puissance du lemme « espace » fortement promu par la revue *L'Espace Géographique* sont indiscutables. Il est possible à partir de ce tableau de penser également une certaine stagnation dans l'usage de ce lemme sur la dernière période. Cet usage reste toutefois élevé. De plus, il faut préciser que la mise en avant d'une baisse sur la période 1985-1995 dans *L'Espace Géographique* doit être relativisée puisqu'une partie importante de l'infériorité observée provient de l'intervalle temporel plus réduit de cette période par rapport à celui de 1972-1984[206].

2. Premiers résultats obtenus à partir des cinq variantes méthodologiques

Les cinq tableaux suivants (n°7, 8, 9, 10 et 11) présentent les lemmes les plus proches d'« espace » selon les calculs de similarité cosinus réalisés à partir des plongements avec les cinq variantes méthodologiques (*Word2Vec CBOW*, *Word2Vec Skip-gram*, *GloVe*, *FastText CBOW* et *FastText Skip-gram*) et les paramètres précédemment explicités[207]. Après chaque lemme, entre parenthèses, la similarité cosinus obtenue entre le vecteur de ce lemme et le vecteur du lemme « espace » est indiquée. Pour des raisons de place, les résultats présentés sont ici limités aux dix premiers lemmes[208].

[205] Ce terme provient de la terminologie de Bénédicte Pincemin (*cf.* section Chap4.II).

[206] En réalisant le calcul sur la période 1985-1997 (pour avoir un pas de temps exactement similaire à 1972-1984), les résultats suivants sont obtenus : 2727 pour les *Annales de Géographie* et 6215 pour *L'Espace Géographique*. La baisse dans *L'Espace Géographique* sur cette dernière période est par conséquent très réduite.

[207] Minimum d'occurrences : 15, nombre d'itérations : 100, taille du contexte : 10 et taille des plongements :50. Pour les deux variantes utilisant *FastText*, taille minimale des n-gram : 3 et taille maximale : 6.

[208] Il n'existe pas dans la littérature de seuil permettant de discriminer à partir du calcul de la similarité cosinus si deux termes sont sémantiquement proches ou non. Il est courant de présenter les valeurs obtenues les plus grandes tout en précisant que le résultat d'un calcul de similarité cosinus entre deux vecteurs est au maximum égal à 1.

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	stationnement (0,58) rempart (0,56) bâtir (0,51) incendie (0,51) voirie (0,51) enceinte (0,5) boulevard (0,49) palais (0,49) quartier (0,47) urbanisme (0,46)	périmètre (0,61) territoire (0,55) organisation (0,55) fonctionnel (0,53) spatial (0,51) tissu (0,51) unité (0,49) structurer (0,49) intégrer (0,49) aire (0,49)	interdépendance (0,59) entité (0,58) spécificité (0,55) dialectique (0,54) territoire (0,52) aire (0,51) spatialement (0,51) conception (0,5) adéquation (0,5) vision (0,5)	territoire (0,64) spatial (0,56) abstrait (0,54) noyau (0,52) tissu (0,52) acteur (0,52) com (0,52) aire (0,52) milieu (0,51) pratique (0,51)
L'Espace Géographique			territoire (0,6) support (0,54) optique (0,54) territorialité (0,53) articulation (0,53) enveloppe (0,51) organisation (0,5) univers (0,5) social (0,49) mode (0,48)	territoire (0,61) social (0,52) univers (0,5) mode (0,5) notion (0,5) champ (0,5) société (0,5) contexte (0,49) représentation (0,49) milieu (0,49)

Tableau n°7 : Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace » d'après la méthode *Word2Vec CBOW* et les paramètres spécifiés.

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	bâtir (0,71) vert (0,67) jardin (0,61) quartier (0,61) vaste (0,6) enceinte (0,6) boulevard (0,59) pavillon (0,58) stationnement (0,58) voirie (0,58)	organisation (0,71) fonctionnel (0,71) urbain (0,7) intégrer (0,68) spatial (0,68) structurer (0,67) vacance (0,65) humaniser (0,64) harmonieux (0,64) parc (0,63)	spatial (0,68) géographique (0,67) structuration (0,66) organisation (0,64) fonctionnel (0,64) définir (0,63) vision (0,63) concevoir (0,63) territoire (0,62) urbain (0,61)	spatial (0,68) territoire (0,66) tendre (0,65) organisation (0,63) paysage (0,62) social (0,62) territorialité (0,62) urbain (0,61) imaginaire (0,61) perception (0,61)
L'Espace Géographique			organisation (0,74) concevoir (0,73) fonctionnel (0,71) notion (0,7) spatial (0,69) dimension (0,68) inscrire (0,68) vivre (0,68) organiser (0,68) définir (0,67)	façonner (0,73) représentation (0,72) organiser (0,7) inscrire (0,69) spatial (0,68) vivre (0,68) définir (0,67) société (0,67) territoire (0,67) structurer (0,67)

Tableau n°8 : Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace » d'après la méthode *Word2Vec Skip-gram* et les paramètres spécifiés.

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	espacé (0,83) espacement (0,74) champs_élysées (0,52) enceinte (0,49) bâtir (0,48) herbacé (0,46) rue (0,45) tuile (0,45) rempart (0,43)	espacé (0,81) organisation (0,58) ensemble (0,54) harmonie (0,54) unité (0,53) spatial (0,52) densément (0,52) intégration (0,52) armature (0,5) périmètre (0,5)	espacé (0,79) territoire (0,57) spatialement (0,57) interdépendance (0,56) spatial (0,53) interrelation (0,52) interurbain (0,51) conception (0,5) entité (0,5) dimension (0,5)	espacé (0,71) territoire (0,61) spatialement (0,58) socio_spatial (0,58) territorialité (0,55) spatial (0,55) entité (0,55) modalité (0,55) milieu (0,54) tissu (0,52)
L'Espace géographique			sous_espace (0,73) espace_temps (0,7) sous_espaces (0,62) territoire (0,58) notion (0,54) territorialité (0,54) archétype (0,52) trame (0,51) forme (0,5) concrètement (0,5)	espace_temps (0,72) sous_espaces (0,7) spatialité (0,69) espacement (0,65) sociétal (0,59) spatialement (0,57) territoire (0,57) territorialité (0,56) métastructure (0,55) matérialité (0,54)

Tableau n°9 : Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace » d'après la méthode *FastText CBOW* et les paramètres spécifiés.

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	bâtir (0,68) vert (0,63) quartier (0,63) implantation (0,62) pavillon (0,62) parcellaire (0,62) parc (0,62) rue (0,61) libre (0,61) palais (0.61)	intégrer (0.76) organisation (0.76) fonctionnel (0.71) structurer (0.71) urbain (0.69) aménagement (0.67) harmonieux (0.67) concevoir (0.67) urbaniste (0.66) spatial (0.66)	relationnel (0.76) structurer (0.76) structuration (0.75) vision (0.74) spatial (0.74) concevoir (0.73) géographique (0.73) entité (0.72) dialectique (0.71) spatialement (0.7)	territoire (0.75) relationnel (0.72) structurer (0.71) authentique (0.71) territorialité (0.69) représentation (0.68) tendre (0.68) fonctionnel (0.67) perception (0.67) spatial (0.67)
L'Espace géographique			indissociable (0.79) notion (0.79) concevoir (0.79) fonctionnel (0.78) organisation (0.78) travers (0.77) vivre (0.76) relationnel (0.75) territorialité (0.75) transposition (0.74)	inscrire (0.79) façonner (0.78) société (0.78) représentation (0.77) organisation (0.76) spatial (0.76) vivre (0.76) appropriation (0.75) territoire (0.75) champ (0.74)

Tableau n°10 : Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace » d'après la méthode *FastText Skip-gram* et les paramètres spécifiés

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	bâtir (0.66) vert (0.55) vaste (0.51) libre (0.49) urbain (0.47) reconstruction (0.47) extension (0.46) occuper (0.46) entourer (0.46) quartier (0.45)	organisation (0.69) urbain (0.62) fonctionnel (0.6) ensemble (0.51) aménagement (0.5) vaste (0.5) unité (0.49) définir (0.48) territoire (0.48) touristique (0.47)	urbain (0.64) définir (0.63) géographique (0.63) spatial (0.6) humain (0.57) social (0.56) relation (0.56) organisation (0.55) dimension (0.54) aménagement (0.54)	territoire (0.65) tant (0.62) urbain (0.61) spatial (0.6) social (0.6) pratique (0.57) définir (0.57) nouveau (0.55) géographique (0.54) rapport (0.53)
L'Espace géographique			urbain (0.64) organisation (0.6) vivre (0.57) spatial (0.57) définir (0.57) forme (0.56) social (0.56) notion (0.55) territoire (0.54) considérer (0.53)	représentation (0.65) territoire (0.65) organisation (0.63) géographique (0.61) spatial (0.61) agir (0.6) définir (0.57) vivre (0.56) pratique (0.55)

Tableau n°11 : Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace » d'après la méthode *GloVe* et les paramètres spécifiés.

Pour ne pas limiter la visualisation de ces résultats à 10 lemmes (ou 10 termes) et pour faciliter les comparaisons, un module de visualisation a été construit dans l'application. Il permet de choisir le ou les expérimentation(s) à visualiser : **@CompareExpeGeneral**. Pour les cinq expérimentations précédentes, il faut suivre le lien suivant : **@CompareExpe1**. L'utilisateur a alors accès aux résultats précédents avec une extension jusqu'à 50 lemmes (ou termes) et la possibilité de comparer facilement les différentes périodes. Une option a été ajoutée pour limiter la visualisation à une seule revue permettant d'avoir un accès plus simple à la barre de défilement horizontal (située en bas) pour passer d'une période à une autre. Le nombre de lemmes (ou termes) peut également être réduit pour privilégier une comparaison entre les revues, si nécessaire. Le nombre de tableaux affichés peut enfin être réduit pour s'adapter à des écrans plus petits ou favoriser la visualisation simultanée de plusieurs périodes.

Dans les résultats obtenus, ceux de la méthode *FastText CBOW* se différencient assez nettement des autres. En effet, c'est la seule méthode où ressortent explicitement dans les premiers résultats des formes graphiques très proches d'espace comme « espacé » ou « sous-espace ». Il est exact que ces formes sont sémantiquement proches d'« espace », mais ce

n'est pas ce sémantisme de forme qui est important par rapport à la recherche menée. En effet, les évolutions mises en avant par Olivier Orain portent plus sur un sémantisme qui est indépendant de ces proximités de formes. Un des objectifs de cette recherche étant de suivre dans le temps les évolutions sémantiques, il est possible de penser que ce sémantisme de forme, en répétant les mêmes lemmes à chaque époque, risque de générer des continuités non pertinentes pour cette recherche. L'ensemble de ces réflexions m'a conduit à ne pas retenir cette méthode.

L'analyse des résultats m'a conduit également à exclure la méthode *Word2Vec CBOW*. Ce qui a été décisif dans cette décision, ce sont surtout les résultats obtenus pour *L'Espace Géographique* sur 1972-1984. Les premiers lemmes obtenus (« territoire », « support », « optique », « territorialité », « articulation », « enveloppe ») sont, à mon avis, moins représentatifs des notions liées au lemme « espace » pour cette période et cette revue que ceux obtenus par exemple avec *GloVe* (« urbain », « définir », « géographique », « spatial », « humain », « social »). La montée en puissance du terme « territoire » est postérieure à cette époque[209]. Ce faisant, il faut reconnaître qu'un tel argument introduit un autre facteur que la seule proximité sémantique.

Cette réflexion m'a conduit à réfléchir à la construction d'une évaluation plus objectivée. Plutôt qu'une comparaison des résultats les uns par rapport aux autres, la technique la plus sûre serait de fixer une référence permettant d'évaluer indépendamment chacun des résultats. Le problème est que cette référence n'existe pas. En effet, il n'existe pas de recensement des significations d'« espace » permettant d'évaluer quantitativement et de manière indépendante les résultats obtenus. D'un point de vue théorique, cette affirmation est évidemment en adéquation avec le choix du cadre théorique précédemment réalisé : un positionnement affirmé dans des sémantiques différentielles et non dans des sémantiques référentielles ou componentielles (*cf.* section Chap3.I).

Cette réflexion sur l'absence de référence et ses conséquences d'un point de vue épistémologique peut être développée avec les analyses du philosophe Ludwig Wittgenstein et des penseurs du cercle de Vienne[210]. En effet, le problème ici rencontré est une conséquence logique de la difficulté des êtres humains à objectiver le langage puisque leurs

[209] Lévy (2003, 907) fait remonter l'« entrée 'officielle' » du terme de « territoire » dans la production francophone à 1982 avec la rencontre *Géopoint* intitulée *Les territoires de la vie quotidienne*. Sa montée en puissance est par conséquent postérieure à cette date.

[210] Le cercle de Vienne est un groupe de savants qui se sont réunis dans la capitale de l'Autriche du début des années 1920 jusqu'en 1936. Il est composé notamment par Rudolf Carnap, Kurt Gödel, Otto Neurath… Leur objectif − la refondation de la métaphysique − explique des réflexions poussées sur ce problème de la référence. L'émigration de ces penseurs due à la montée du nazisme et à la seconde guerre mondiale a nourri l'école pragmatique américaine, avec notamment les pensées d'Hilary Putnam dont Olivier Orain se sert comme appui philosophique (*cf.* section Chap11.II).

pensées sont prises dans ce même langage. Otto Neurath a proposé dans cette optique une métaphore qui peut être résumée ainsi : « il faut réparer le bateau en pleine mer et au coup par coup »[211]. Cette perspective remet en cause de façon profonde une épistémologie cherchant et revendiquant traditionnellement une fondation stable des connaissances.

Face à ces réflexions qui pourraient être critiquées comme une façon de s'en sortir à moindres frais, il est nécessaire de rappeler que ce qui est ici exprimé est un positionnement fort, loin d'être sans conséquence dans la problématique d'analyse de la thèse d'Olivier Orain (2003). En effet, l'argument qui soutient la lecture kuhnienne de cet auteur est l'affirmation d'une rupture dans le rapport connaissance / réalité signant la fin d'un paradigme réaliste. Or, ce positionnement concerne ce rapport connaissance / réalité. Il est même d'une certaine manière favorable à l'évolution mise en avant par Olivier Orain, dans la mesure où précisément les sémantiques référentielles sont intrinsèquement liées à une approche réaliste (*cf.* section Chap3.I.1), notamment à travers « le positivisme logique qui exprime un idéal de correspondance si l'on peut dire terme à terme entre un mot, un concept et un objet » (Rastier 1995).

Est-ce qu'en affirmant ce positionnement je quitte le paradigme réaliste ? Ici, il est important de signaler que quand j'ai exclu précédemment la méthode *Word2Vec CBOW* en considérant ses résultats comme moins représentatifs que la méthode *GloVe*, j'ai convoqué à mon sens un argument réaliste. En effet, la qualité de la représentation se réfère certes à ma connaissance, mais en deçà présuppose que cette connaissance n'est pas totalement erronée. Ce n'est donc pas un saut hors du réalisme que j'effectue ici, mais bien plutôt un positionnement complexe pour gérer au mieux une difficulté.

Affirmer l'impossibilité d'une évaluation directe par rapport à un référentiel n'empêche en rien de réaliser des évaluations indirectes. Pour essayer d'étayer le choix entre les trois méthodes restantes (*Word2Vec Skip-gram*, *FastText Skip-gram* et *GloVe*), une fonctionnalité a été construite dans l'application permettant de mieux visualiser et comparer quantitativement les lemmes (ou termes) communs entre différents résultats. Pour cela, l'utilisateur choisit en amont n-expérimentations. Si n est inférieur à 5, les n résultats peuvent être directement visualisés. Pour chaque période et revue, les lemmes en communs des résultats affichés sont surlignés en gris. Par exemple, en comparant les méthodes *Word2Vec Skip-gram* et *FastText Skip-gram* à partir des résultats précédemment obtenus sur les *Annnales de Géographie*, les trois tableaux suivants (n°12, 13 et 14) sont produits :

[211] Cette métaphore a été popularisée par W. V. O. Quine dans *Word And Object* (1960).

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	bâtir (0.71) vert (0.67) jardin (0.61) quartier (0.61) vaste (0.6) enceinte (0.6) boulevard (0.59) pavillon (0.58) stationnement (0.58) voirie (0.58)	organisation (0.71) fonctionnel (0.71) urbain (0.7) intégrer (0.68) spatial (0.68) structurer (0.67) vacance (0.65) humaniser (0.64) harmonieux (0.64) parc (0.63)	spatial (0.68) géographique (0.67) structuration (0.66) organisation (0.64) fonctionnel (0.64) définir (0.63) vision (0.63) concevoir (0.63) territoire (0.62) urbain (0.61)	spatial (0.68) territoire (0.66) tendre (0.65) organisation (0.63) paysage (0.62) social (0.62) territorialité (0.62) urbain (0.61) imaginaire (0.61) perception (0.61)

Tableau n°12 : Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace » d'après la méthode *Word2vec Skip-gram* et les paramètres spécifiés.

(Coloration des lemmes en commun par rapport aux résultats suivants)

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	bâtir (0.68) vert (0.63) quartier (0.63) implantation (0.62) pavillon (0.62) parcellaire (0.62) parc (0.62) rue (0.61) libre (0.61) palais (0.61)	intégrer (0.76) organisation (0.76) fonctionnel (0.71) structurer (0.71) urbain (0.69) aménagement (0.67) harmonieux (0.67) concevoir (0.67) urbaniste (0.66) spatial (0.66)	relationnel (0.76) structurer (0.76) structuration (0.75) vision (0.74) spatial (0.74) concevoir (0.73) géographique (0.73) entité (0.72) dialectique (0.71) spatialement (0.7)	territoire (0.75) relationnel (0.72) structurer (0.71) authentique (0.71) territorialité (0.69) représentation (0.68) tendre (0.68) fonctionnel (0.67) perception (0.67) spatial (0.67)

Tableau n°13 **:** Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace » d'après la méthode *FastText Skip-gram* et les paramètres spécifiés.

(Coloration des lemmes en commun par rapport aux résultats précédents)

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	0,4	0,7	0,5	0,5

Tableau n°14 : Rapport entre le nombre de lemmes en commun et le nombre total de lemmes pour chaque époque entre les deux résultats précédents.

Dans le Tableau n°14, pour les *Annales de Géographie* sur la période 1950-1960, le résultat « 0,4 » s'explique par le fait qu'il y a 4 lemmes en commun sur 10 en comparant les deux méthodes précédentes (*cf.* Tableau n°12 et 13). Il est possible de calculer une moyenne sur l'ensemble des périodes, ici[212] 0,525 (soit 52,5 %).

Le lien suivant **@CompareExpe1a** permet de comparer les résultats obtenus lors de cette expérimentation avec un affichage de 50 lemmes pour les méthodes *Word2Vec Skip-gram* et *FastText Skip-gram*. Il existe une proximité de ces deux résultats (plus de 60 % en moyenne de lemmes en commun). Les similarités des résultats entre *GloVe* et *Word2Vec Skip-gram* (**@CompareExpe1b**) ainsi que *GloVe* et *FastText Skip-gram* (**@CompareExpe1c**) sont bien moindres, avec en moyenne 45 % pour la première comparaison et 30 % pour la seconde. Un fait marquant dans une comparaison entre ces trois méthodes est que les premiers lemmes obtenus avec la méthode *GloVe* font beaucoup plus souvent partie des ressemblances que ceux obtenus avec les méthodes *Word2Vec Skip-gram* ou *FastText Skip-gram*. Le tableau suivant reproduit les résultats obtenus pour les *Annales de Géographie* 1950-1960. Pour des raisons de place, seuls les 10 premiers lemmes sont présentés, mais la coloration correspond à une comparaison des résultats entre les trois résultats (*Word2Vec Skip-gram*, *FastText Skip-gram* et *GloVe*) sur 50 lemmes.

[212] (4 + 7 + 5 + 5) / 40

Word2Vec Skip-gram	*FastText Skip-gram*	*GloVe*
bâtir (0.71)	bâtir (0.68)	bâtir (0.66)
vert (0.67)	vert (0.63)	vert (0.55)
jardin (0.61)	quartier (0.63)	vaste (0.51)
quartier (0.61)	implantation (0.62)	libre (0.49)
vaste (0.6)	pavillon (0.62)	urbain (0.47)
enceinte (0.6)	parcellaire (0.62)	reconstruction (0.47)
boulevard (0.59)	parc (0.62)	extension (0.46)
pavillon (0.58)	rue (0.61)	occuper (0.46)
stationnement (0.58)	libre (0.61)	entourer (0.46)
voirie (0.58)	palais (0.61)	quartier (0.45)

Tableau n°15 : Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace » dans les *Annales de Géographie* sur la période 1950-1960.

Les lemmes en communs (sur 50 lemmes retenus pour chaque méthode) ont été surlignés.

Le Tableau n°15 montre que les premiers lemmes obtenus à partir de la méthode *GloVe* intègrent dans ce cas une plus grande partie de l'ensemble des lemmes obtenus avec les trois méthodes. Ce fait n'est certes pas toujours aussi marqué suivant les époques, les revues et le nombre de lemmes choisis. Cependant, les expérimentations réalisées m'ont conduit à observer que ce pourcentage de lemmes en commun (dès les premiers lemmes par rapport à l'ensemble des lemmes obtenus) est plus souvent supérieur avec la méthode *GloVe*. Cette observation, couplée à l'idée que les lemmes obtenus étaient globalement plus représentatifs (car moins spécifiques d'une seule méthode), permet de comprendre pourquoi la méthode *GloVe* a été privilégiée.

La justification de ce choix reste de mon point de vue assez faible. C'est plus la nécessité pratique de sélectionner quelques résultats à analyser qui a été déterminante. En effet, j'ai multiplié les expérimentations jusqu'à identifier un argument jugé assez pertinent pour justifier un choix. Toutefois, il m'est impossible d'affirmer que les résultats des deux autres méthodes (*Word2Vec Skip-gram* et *FastText Skip-gram*) sont moins pertinents que les résultats issus de *GloVe*. L'application créée, en permettant de consulter les résultats qui auraient été obtenus si ces deux autres méthodes avaient été privilégiées, doit aussi être pensée comme une cause et une conséquence[213] de la reconnaissance qui vient d'être

[213] La possibilité de faire des liens vers l'application supprime l'obligation de présenter seulement quelques résultats. Une fois celle-ci créée, j'ai été également entraîné par ses possibilités d'actions et de présentations. La notion de milieu, telle que la définit Augustin Berque (2014), à la fois comme « matrice » et « empreinte » mérite, à mon avis d'être invoquée. L'application est à la fois une « matrice » et une « empreinte ». Une fois créée, elle n'a pas été qu'un outil technique, mais a joué un rôle dans l'affirmation du positionnement épistémologique.

effectuée. Même si les résultats de ces deux autres méthodes ne seront pas analysés dans cette thèse, ils peuvent être facilement être produits et analysés. Cette application permet donc ainsi aussi de laisser ouvertes les expérimentations et les réflexions.

À cette première phase d'expérimentations succède une seconde phase dont l'objectif est de déterminer les paramètres à privilégier au sein de la méthode retenue.

3. Seconde phase d'expérimentation sur les paramètres

La technique utilisée a été de faire varier plusieurs paramètres de la méthode *GloVe* l'un après l'autre et d'examiner les effets sur les résultats obtenus. Le premier paramètre testé est le minimum du nombre d'occurrences.

a. Minimum du nombre d'occurrences

L'expérimentation menée a consisté à faire varier ce paramètre entre 5 et 20 avec un pas de 5. Toutes les valeurs n'ont donc pas été testées. L'objectif est de déterminer si un choix optimal peut être revendiqué au sein des valeurs testées. Le lien suivant **@CompareExpe2a50** présente les résultats obtenus pour 50 lemmes affichés en conservant pour les autres paramètres les valeurs adoptées pour la première expérimentation.

Nous reproduisons ci-dessous les pourcentages de lemmes en commun entre les différents résultats obtenus pour chaque période et revue :

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	0,42	0,6	0,72	0,7
L'Espace Géographique			0,64	0,7

Tableau n°16 : Rapport entre le nombre de lemmes en commun et le nombre total de lemmes en faisant varier le minimum du nombre d'occurrences (5,10,15,20).
(Méthode *GloVe* avec nombre d'itérations : 100, taille du contexte : 10 et taille des plongements : 50)

Dans ce tableau, par exemple, pour les *Annales de Géographie* sur la période 1985-1995, « 0,7 » signifie que 70 % des lemmes se retrouvent dans les résultats que le minimum du nombre d'occurrences soit fixé à 5, 10, 15 ou 20 (pour cette époque et revue avec la méthode et les valeurs spécifiées pour les autres paramètres).

Ce tableau permet de constater un pourcentage de lemmes en commun bien inférieur pour la période 1950-1960. Ceci s'explique facilement par le nombre beaucoup plus faible d'occurrences du lemme « espace » sur cette période. Une faible variation comme le retrait plus ou moins important des mots de basses fréquences engendre alors plus de changements dans les résultats. En rentrant dans le détail des résultats à partir du lien (**@CompareExpe2a50**), il est possible de caractériser cette spécificité de la période 1950-1960 d'une autre manière : alors que les lemmes en commun[214] deviennent beaucoup plus rares après le quinzième lemme, ils se maintiennent pour les autres périodes et revues jusqu'au trentième, voire au quarantième lemme. Ces scores dépendent bien entendu du nombre de lemmes retenus, mais des expérimentations avec 25 ou 75 lemmes (**@ComparaExpe2a25** et **@CompareExpe2a75**) montrent toujours une plage réduite de résultats en commun de la période 1950-1960 lors de la variation du minimum du nombre d'occurrences.

Concernant le problème initial, à savoir le choix d'une valeur optimale du nombre d'occurrences, l'observation dans le détail des résultats (à l'aide des liens et que ce soit pour 25, 50 ou 75 termes retenus) ne permet pas de répondre de manière satisfaisante. Tout d'abord, à la différence de l'argumentation qui a conduit précédemment à exclure la méthode *Word2Vec CBOW*, il n'y a pas un lemme ou un ensemble de lemmes obtenus permettant d'affirmer qu'une valeur testée pour le minimum du nombre d'occurrences semble plus pertinente que les autres. Ensuite, à la différence de l'argumentation qui a conduit précédemment à retenir plutôt la méthode *GloVe* par rapport à *Word2vec Skip-gram* et *FastText Skip-gram*, il n'y a pas un ensemble de lemmes en commun plus important pour une valeur spécifique sur plusieurs époques et revues permettant de justifier un choix. J'ai donc pris la décision par la suite de garder ce paramètre fixé à 15 comme il avait été fixé initialement.

b. Nombre d'itérations

Le second paramètre testé étant le nombre d'itérations, une variation de leur nombre a été réalisée entre 100 et 250 avec un pas de 50. Les pourcentages suivants de lemmes en commun sont obtenus :

[214] Pour rappel, après avoir utilisé la méthode *GloVe*, fixé successivement le minimum d'occurrences à 5,10,15 et 25 et retenu 50 lemmes.

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	0,8	0,9	0,94	0,92
L'Espace Géographique			0,92	0,84

Tableau n°17 : Rapport entre le nombre de lemmes en commun et le nombre total de lemmes en faisant varier le nombre d'itérations (100,150,200,250).
(Méthode *GloVe*, min nombre d'occurrences : 15, taille du contexte : 10 et taille des plongements : 50)

Ce tableau se lit de la même manière que ce qui a été explicité pour le Tableau n°16. De plus, un phénomène identique est observable avec la présence d'une valeur plus faible pour les *Annales de Géographie* 1950-1960. L'explication précédemment avancée reste valable, à savoir une quantité de données moindre[215] pour cette revue et époque, permettant de comprendre une plus grande instabilité des résultats. Globalement, les scores obtenus pour ce paramètre sont très élevés, ce qui montre une bonne convergence de l'algorithme.

Enfin, de la même manière que précédemment, un examen dans le détail des résultats (**@CompareExpe2b25**, **@CompareExpe2b50** et **@CompareExpe2b75**) en s'appuyant sur le contenu des lemmes ou en utilisant le nombre de lemmes en commun, ne permet pas d'identifier une valeur optimale. La valeur de ce paramètre a donc été maintenue à 100 pour les prochaines expérimentations.

c. Taille du contexte

De la même manière que précédemment, pour une variation de la taille du contexte, entre 5 et 20 avec un pas de 5, les pourcentages suivants de lemmes en commun sont obtenus :

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	0,3	0,48	0,54	0,6
L'Espace géographique			0,62	0,66

Tableau n°18 : Rapport entre le nombre de lemmes en commun et le nombre total de lemmes en faisant varier la taille du contexte (5,10,15,20).
(Méthode *GloVe*, min nombre d'occurrences : 15, nombre d'itérations : 100 et taille des plongements : 50)

[215] Quantité de données moindre par rapport au lemme ici testé : « espace ».

Même si ces résultats quantitatifs montrent des variations importantes suivant le choix de ce paramètre, il est difficile d'identifier une valeur optimale. Après un examen en détail des résultats (**@CompareExpe2c25**, **@CompareExpe2c50**, **@CompareExpe2c75**), la répétition de cette situation m'a conduit à essayer de proposer une explication pour mieux comprendre les causes de cette indétermination. Une difficulté majeure, à mon avis, est qu'il n'y a pas eu UN sens du terme « espace » qui s'est maintenu dans la durée et qui a été remplacé de manière nette et durable par un autre sens. Il y a plutôt une multiplicité de sens évoluant dans le temps. Cette plasticité sémantique, au-delà du fait qu'elle est un résultat intéressant en soi, permet aussi de mieux comprendre un aspect de la difficulté d'évaluation. Comme il y a beaucoup de lemmes proches sémantiquement d'« espace » avec des orientations sémantiques variées, il est difficile d'établir une discrimination entre des lemmes pertinents et non-pertinents.

Suite à ces observations et ces analyses, la valeur de 10, initialement choisie comme taille de contexte, a été conservée pour les prochaines expérimentations.

d. Taille des plongements

Dans les mêmes conditions que les trois expériences précédentes et pour une variation de ce paramètre entre 50 et 200 avec un pas de 50, les pourcentages suivants de lemmes en commun sont obtenus :

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	0,32	0,54	0,6	0,66
L'Espace Géographique			0,68	0,7

Tableau n°19 : Rapport entre le nombre de lemmes en commun et le nombre total de lemmes en faisant varier la taille des plongements (50,100,150,200).
(Méthode *GloVe*, min nombre d'occurrences : 15, nombre d'itérations : 100 et taille du contexte : 10)

L'examen en détail des résultats (**@CompareExpe2d25**, **@CompareExpe2d50**, **@CompareExpe2d75**) permet d'observer que l'augmentation de la taille du plongement au-delà de 50 permet aux lemmes « structure » et « fonction » d'être beaucoup mieux placés[216] pour la période 1972-1984 dans *L'Espace Géographique.* L'importance de ces

[216] Entre la vingtième et la vingt-cinquième position.

concepts à cette période et dans cette revue m'a conduit à privilégier une taille de plongement supérieure à 50. Il est possible également d'observer sur ces mêmes résultats que le passage d'une taille de plongement de 100 à 150 (ou 200) a pour effet de faire monter le lemme « temps » de la trente-cinquième position à la douzième position, mais également dans le même temps de provoquer une baisse du lemme « échelle ». Faut-il alors plutôt privilégier une taille de plongement de 150 ou 200 ? La question peut être à mon avis largement débattue.

Si le gain en passant d'une taille de plongement de 50 à 100 me semble tout à fait justifiable, le passage de 100 à 150 reste plus discutable. Pour cette raison, j'ai choisi de retenir la valeur 100 pour ce paramètre. Sur le fond, au-delà des interrogations sur la pertinence des montées/descentes de quelques lemmes, se pose la question du risque d'une sélection des résultats attendus à travers le choix des paramètres. Il existe là une réelle tension entre la volonté de faire émerger des résultats du corpus et la reconnaissance que ces résultats dépendent de paramètres fixés de manière réfléchie sans qu'une nécessité forte puisse justifier de manière décisive ces choix.

De plus, il faut mentionner qu'il est bien entendu possible de tester également l'effet des variations combinées de plusieurs de ces paramètres. Par conséquent, cette thèse reconnaît un travail incomplet d'optimisation. Les choix réalisés doivent par conséquent être compris, non pas comme une optimisation totale, mais comme une amélioration partielle.

Suite à ces choix de valeurs pour les paramètres, la section suivante présente et analyse les premiers résultats obtenus.

II. Premiers résultats et leurs analyses

1. Évolution sémantique du lemme « espace » et ses deux interprétations

Le premier résultat présenté découle du travail précédent sur la méthode *GloVe* et les valeurs des paramètres retenues.

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	bâtir (0.55) vert (0.49) vaste (0.47) libre (0.44) extension (0.41) urbain (0.4) occuper (0.39) temps (0,39) vide (0.37) paysage (0.36) entourer (0.34) quartier (0.34) limite (0.34) immense (0.34) reconstruction (0.34)	organisation (0.63) urbain (0.5) vaste (0.48) ensemble (0.45) aménagement (0.43) fonctionnel (0.43 définir (0.43) territoire (0,43) régional (0.43) fonction (0.4) touristique (0.4) organiser (0.4) intégrer (0.39) temps (0.39) géographique (0.38)	géographique (0.61) urbain (0.55) organisation (0.54) définir (0.53) spatial (0.49) aménagement (0.48) humain (0.48) territoire (0.48) social (0.48) dimension (0.47) rural (0.47) milieu (0.46) relation (0.45) régional (0.45) nouvelle (0.45)	spatial (0.57) territoire (0.56) social (0.55) tant (0.54) urbain (0.53) géographique (0.5) nouveau (0.5) organisation (0.49) rural (0.46) rapport (0.46) pratique (0.45) aussi (0.44) forme (0.44) représentation (0.44) société (0.43)
L'Espace Géographique			organisation (0.54) urbain (0.51) social (0.51) spatial (0.5) géographique (0.5) vivre (0.49) définir (0.48) rural (0.46) territoire (0.46) organiser (0.46) ainsi (0.45) notion (0.45) économique (0.44) ensemble (0.44) considérer (0.44)	géographique (0.59) organisation (0.56) vivre (0.55) territoire (0.51) spatial 0.5) représentation (0.5) définir (0.49) ainsi (0.49) agir (0.48) organiser (0.46) social (0.46) société (0.46) pouvoir (0.45) urbain (0.45) monde (0.45)

Tableau n°20 : Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace » par époques et revues avec la méthode et les paramètres retenus.
(Méthode *GloVe* avec minimum nombre d'occurrences : 15, nombre d'itérations : 100, taille du contexte : 10 et taille des plongements : 100)

Ce premier résultat est caractérisé pour la période 1950-1960 des *Annales de géographie* par deux pôles sémantiques. Le premier est lié aux espaces bâtis. Ce pôle est évidemment marqué par la présence du lemme « bâtir » en première position, mais aussi ensuite par les lemmes « urbain », « quartier », « reconstruction »… Le second pôle est lié aux espaces

naturels ou agricoles avec les lemmes « vert », « vide » ou encore « champs », « doline »[217]. Par rapport à ces deux pôles, il faut souligner que plusieurs lemmes peuvent appartenir à l'un et à l'autre comme « libre », « extension », « occuper »… De plus, il serait possible de considérer que le lemme « temps » relève par exemple d'un autre pôle sémantique. Toutefois, il n'existe pas vraiment, de mon point de vue, de lemmes sémantiquement proches de ce lemme « temps » formant un agrégat facilement identifiable.

En passant de 1950-1960 à 1961-1971, un changement sémantique important est observable. De nouveaux lemmes s'affirment comme « organisation », « aménagement » ou « fonctionnel », transformant une partie du pôle sémantique lié aux espaces « bâtis ». De plus, il y a un net recul du pôle sémantique lié aux espaces naturels et agricoles. De nouveaux lemmes apparaissent comme « touristique ». L'échelon « régional » s'affirme. Enfin, il est possible de remarquer que l'ensemble de la sémantique liée au lemme « espace » est plus conceptuel et moins concret qu'en 1950-1960. Les lemmes comme « vert », « entourer », « quartier » reculent et des lemmes comme « fonction », « définir », « intégrer » deviennent plus présents.

Les années 1970-1984 confirment cette évolution sémantique. Dans les deux revues, il y a une montée en puissance des lemmes « géographique », « spatial » et « social ». Un nouveau trait sémantique apparaît dans *L'Espace Géographique* avec le verbe « vivre » qui renvoie à l'« espace vécu ». Toujours dans cette revue, une conceptualisation plus importante est notable avec des lemmes comme « notion » ou encore « forme », « relation », « concept », « structure » et « type »[218]. La sémantique liée aux espaces naturels et agricoles se maintient dans les *Annales de Géographie* avec « rural », « nature » et « milieu ». Elle est moins présente dans *L'Espace Géographique* à l'exception du lemme « rural ».

La période suivante, 1985-1995, marque une certaine continuité avec l'importance des lemmes « organisation », « géographique », « spatial », « social », mais aussi l'affirmation de nouveaux traits avec les lemmes « territoire », « représentation », « système » et « pratique » dans les deux revues. Le terme « rural » se maintient dans les *Annales de Géographie* (ainsi que « nature »), mais décline dans *L'Espace Géographique*.

Ces résultats peuvent donner lieu à deux grandes interprétations opposées par rapport à la thèse d'Olivier Orain :

[217] En consultant les résultats jusqu'au cinquantième lemme avec le lien suivant **@Result1n50**, les lemmes « champ » et « doline » sont situés respectivement à la 19ème et 20ème position. Sur ce cas spécifique, il n'est pas pertinent d'aller jusqu'à la cinquantième position, car les valeurs de similarité-cosinus deviennent assez faibles et les résultats peuvent être légitimement jugés comme sémantiquement moins pertinents.

[218] Ces derniers lemmes ne sont observables qu'à partir du lien précédemment cité.

- La première remet en cause de manière importante les analyses développées par cet auteur. En effet, Olivier Orain affirme l'importation d'une nouvelle terminologie dans les années 1960 sans qu'il y ait de changement sémantique. Or, les résultats tendent à montrer le contraire. Une grande partie du changement sémantique a déjà eu lieu au sein même des *Annales de Géographie* dès 1961-1971. Cette observation est loin d'être anecdotique, car elle tend à accréditer la thèse d'un changement plus continu et moins révolutionnaire. Une telle perspective s'accorde mieux avec les résultats obtenus. Il est en effet particulièrement difficile de réduire les résultats obtenus au passage d'un premier ensemble de sens bien défini par un paradigme à un nouvel ensemble de sens significatif d'un nouveau paradigme.
- La seconde interprétation met en avant au contraire une adéquation entre les résultats obtenus et les analyses développées par Olivier Orain. En effet, ce dernier affirme l'importance d'une « anomalie » dans les années 1960 liée à la rencontre de la géographie avec la scène « aménagementiste ». Cette « anomalie » pourrait être incarnée et justifiée par le changement sémantique observé sur cette décennie. La « crise » des années 1970, au sens kuhnien du terme, développée par Olivier Orain peut alors être reliée à l'approfondissement du changement sémantique observé lors de la période 1972-1984. La progression de la conceptualisation qui a été soulignée peut alors servir à accréditer la thèse d'une modification d'un rapport au « réel » chez les géographes. L'essor du terme « représentation » sur la période 1985-1995 peut être aussi utilisé pour affirmer la fin du « plain-pied » caractéristique pour Olivier Orain de la géographie post-vidalienne.

  Dans cette perspective, une objection aisée à la première interprétation concerne la temporalité du changement sémantique. En effet, la modification sémantique qui est observée sur la période 1961-1971 n'est peut-être qu'un effet de la fin des années 1960 et du début des années 1970. Dans ce cas, l'argumentation de la première interprétation s'en retrouve affaiblie, car il est possible de relier l'évolution sémantique au début de la « crise de la géographie » (assimilée par Olivier Orain à une « crise » kuhnienne).

Ces réflexions et ce questionnement expliquent la production et la présentation d'un résultat à partir d'un découpage temporel plus fin.

2. Résultat avec un découpage temporel plus fin

Le même corpus (ArticleLem) et les mêmes paramètres ont été conservés par rapport au résultat précédent à l'exception du découpage temporel. L'objectif étant de préciser la date du changement sémantique précédemment caractérisé, la division suivante a été privilégiée : 1950-1955, 1956-1960, 1961-1965, 1966-1971, 1972-1976

	1950-1955	1956-1960	1961-1965	1966-1971	1972-1976
Annales de Géographie	libre (0.49) bâtir (0.46) vaste (0.41) urbain (0.4) vert (0.39) temps (0.35) vide (0.34) occuper (0.33) progression (0.33) rapidité (0.31) ouvert (0.31) perfectionner (0.31) banlieue (0.3) cultivé (0.3) ville (0.3)	bâtir (0.53) vert (0.49) vaste (0.44) libre (0.4) temps (0.39) extension (0.37) détruire (0.36) plan (0.36) débouché (0.35) pratiquement (0.35) demeuré (0.33) immense (0.33) sinistré (0.33) rayon (0.32) largement (0.32)	urbaniste (0.47) urbain (0.47) concevoir (0.46) organisation (0.45) vaste (0.45) régional (0.44) échelle (0.43) occupation (0.42) fonctionnel (0.42) géographique (0.41) ensemble (0.4) aménagement (0.4) relation (0.39) temps (0.39) territoire (0.38)	organisation (0.65) urbain (0.48) aménagement (0.46) fonctionnel (0.45) vert (0.44) vaste (0.43) fonction (0.42) territoire (0.41) bâtir (0.41) touristique (0.4) social (0.4) économique (0.4) créer (0.39) régional (0.39) temps (0.39)	organisation (0.56) géographique (0.52) régional (0.5) territoire (0.49) urbain (0.48) vaste (0.47) aménagement (0.45) relation (0.44) politique (0.43) homme (0.42) système (0.41) planification (0.4) fonction (0.4) étudier (0.4) organiser (0.4)

Tableau n°21 : Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace » pour les *Annales de Géographie* 1950-1976 avec un découpage tous les 4,5 ans.
(Méthode *GloVe* avec minimum nombre d'occurrences : 15, nombre d'itérations : 100, taille du contexte : 10 et taille des plongements : 100)

Ces résultats[219] montrent que les premiers lemmes obtenus sont très proches dans les années 1950-1955 et 1956-1960. Cette observation permet de penser qu'il existe une certaine continuité sémantique du lemme « espace » sur ces deux périodes. À l'inverse, une discontinuité est facilement observable entre 1956-1960 et 1961-1965 avec un changement notable des lemmes obtenus témoignant d'un changement sémantique important. Enfin, une nouvelle continuité se dessine entre les années 1961-1965, 1966-1971 et 1972-1976. Par conséquent, l'argument précédent qui consistait à penser que le changement sémantique était

[219] Consultables de manière plus complète avec ce lien : **@Result2n50**.

surtout lié à la fin des années 1960 et au début des années 1970, ne tient pas. Il y a lors de la période 1960-1965, considérée par Olivier Orain comme la phase d'« anomalie », un net changement sémantique du lemme « espace ». Cette observation est particulièrement problématique, car la « bascule sémantique » est attendue dans le cas d'une déclinaison sémantique simple du schéma kuhnien pendant la « crise » et même plutôt en sortie de « crise ».

La lecture de la thèse d'Olivier Orain ne permet pas, à mon avis, de comprendre directement l'origine de cette affirmation d'un changement sémantique tardif autour du terme « espace ». La formule qu'il emploie par exemple dans sa conclusion : « C'est ainsi que j'ai montré dans cette thèse qu''espace' a désigné pendant longtemps la même chose que 'milieu' ou 'paysage' précédemment » (Orain 2003, 353) ne renvoie pas à un point précis de sa thèse. Il est sûrement possible d'évoquer la petite partie quantitative mobilisée au milieu de sa thèse où des calculs de rapport entre les occurrences du terme « espace » et celles du terme « milieu » sont réalisés dans plusieurs ouvrages. Toutefois, cet appui est relativement fragile. La composition du corpus et la méthodologie employée sont très peu interrogées. Cette insatisfaction pour comprendre l'origine des affirmations sémantiques d'Olivier Orain, m'a donc conduit à une recherche allant au-delà de ce qui est directement exprimé dans sa thèse.

L'élément trouvé le plus intéressant provient de ma lecture de plusieurs articles de Philippe Pinchemel. Rappelons rapidement que ce géographe a fondé le Centre de géohistoire en 1969, centre qui est à l'origine de l'équipe *Épistémologie et Histoire de la GéOgraphie* (EHGO). Cette équipe a été ensuite dirigée à partir de 1991 par Marie-Claire Robic, c'est-à-dire la directrice de thèse d'Olivier Orain. Ce dernier est lui-même maintenant co-directeur de l'équipe d'EHGO depuis 2018. Si la filiation institutionnelle est indéniable, celle des idées est plus complexe. En effet, Philippe Pinchemel est à l'origine d'un projet de refondation disciplinaire, formalisé dans l'ouvrage *La Face de la Terre* (1988), qui vise l'unité de la géographie par une synthèse d'une partie des méthodologies historiques qu'il identifie comme centrées sur le concept de « milieu » et une partie des nouvelles démarches centrées sur le concept d'« espace ». Ce contexte permet de mieux comprendre d'où vient l'affirmation d'Olivier Orain quant à cette proximité sémantique initiale d'« espace » et de « milieu ».

En rentrant plus en détail dans la pensée de Philippe Pinchemel (1923-2008), une certaine ironie dans la reprise de ses réflexions par Olivier Orain se fait jour. En effet, il y avait une volonté forte de la part de Philippe Pinchemel de ne pas opposer, mais de proposer une synthèse entre la géographie post-vidalienne et la « nouvelle géographie ». Olivier Orain, en s'appuyant sur le schéma kuhnien, se place à l'opposé de cette position. La « révolution

scientifique » et l'« incommensurabilité » n'est pas conciliable avec une « synthèse » des deux approches. Il est vrai que l'insuccès du projet unitaire de *La Face de la Terre* par rapport à la hauteur des espoirs entretenus par Philippe Pinchemel quant à la refondation de la discipline[220], est déjà acté au moment de la thèse d'Olivier Orain. Toutefois, les résultats quantitatifs que j'ai obtenus m'amènent à être sceptique sur cette affirmation d'une correspondance sémantique entre « milieu » et « espace ». Cette dernière doit à mon avis être fortement recontextualisée et comprise par rapport à la volonté de Philippe Pinchemel de défendre son projet intellectuel d'unité disciplinaire.

Tout ce développement permet de mieux comprendre, mais aussi de mieux remettre en cause, les affirmations sémantiques d'Olivier Orain puisque celles-ci peuvent être considérées comme le fruit d'un contexte spécifique. Si ces réflexions permettent d'interroger sa lecture kuhnienne, elles ne la remettent pas directement en cause. Il est en effet possible de reconnaître la faiblesse de la déclinaison sémantique qu'effectue Olivier Orain du schéma kuhnien tout en affirmant que cela n'invalide pas le reste de sa lecture kuhnienne.

La section suivante développe un élément favorable à la thèse d'Olivier Orain.

3. Réflexions à partir de la dynamique sémantique d'« espaces »

Le fait d'avoir construit précédemment plusieurs corpus m'a conduit à expérimenter sur d'autres corpus qu'« ArticleLem ». L'objectif a été tout d'abord d'éprouver l'observation qui vient d'être détaillée et analysée, à savoir un changement sémantique du lemme « espace »[221] beaucoup plus précoce que ce qu'avait affirmé Olivier Orain. Les liens suivants (**@ResultArtNonLemEspace50** et **@ResultArtPlusCrLem2Espace50**) montrent que cette observation n'est pas remise en cause par les corpus « ArticleNonLem » et « ArticlePlusCrLem2 ».

Toutefois, les expérimentations réalisées sur le corpus « ArticleNonLem » m'ont conduit à changer de position sur un point. Il s'agit du choix du corpus « ArticleLem » comme référence. Cette remise en cause découle d'une expérimentation menée sur le terme « espaces » (au pluriel) dans le corpus « ArticleNonLem » toujours avec la méthode *GloVe* et les paramètres précédemment utilisés. Au niveau du découpage temporel, j'ai privilégié

[220] Dans une interview accordée en 1997, Philippe Pinchemel affirme : « Il n'y a pas de science, de démarche, ni de travail scientifique sans une réflexion toujours plus profonde sur les concepts. Or, à l'heure actuelle, la géographie utilise des concepts extrêmement flous et ambigus » (Pinchemel 1997, 48). Il est reconnu par la même que les propositions intellectuelles effectuées dans *La Face de la Terre* n'ont pas été reprises comme socle disciplinaire.

[221] Du lemme ou du terme « espace » pour le corpus non lemmatisé « ArticleNonLem ».

celui utilisé initialement, car l'objectif n'est pas ici de préciser le moment d'une discontinuité. Les résultats obtenus[222] sont les suivants :

	1950-1960	1961-1971	1972-1984	1985-1995
Annales de Géographie	libres (0.59) verts (0.58) territoires (0.49) bâtis (0.47) vastes (0.47) vides (0.45) troncs (0.41) étendues (0.41) immenses (0.4) genève (0.4)	verts (0.77) vastes (0.64) secteurs (0.48) étendues (0.47) périphériques (0.47) libres (0.46) parcs (0.46) grands (0.45) terrains (0.44) disponibles (0.43)	verts (0.61) ensembles (0.58) ruraux (0.57) vastes (0.55) urbains (0.54) milieux (0.53) piétonniers (0.51) géographiques (0.51) périurbains (0.51) régionaux (0.49)	ruraux (0.59) nouveaux (0.57) secteurs (0.53) urbains (0.52) certains (0.52) vastes (0.52) naturels (0.51) industriels (0.5) quartiers (0.49) géographiques (0.49)
L'Espace géographique			ruraux (0.77) types (0.65) urbains (0.6) régionaux (0.6) différents (0.58) vécus (0.58) paysages (0.57) milieux (0.56) ensembles (0.54) homogènes (0.52)	systèmes (0.62) types (0.61) urbains (0.6) industriels (0.57) milieux (0.57) périphériques (0.55) verts (0.55) naturels (0.54) territoires (0.54) ruraux (0.53)

Tableau n°22 : Termes dont les vecteurs sont les plus proches du vecteur du terme « espaces » dans le corpus « ArticleNonLem ».

(Méthode *GloVe* avec nombre minimum d'occurrences : 15, nombre d'itérations : 100, taille du contexte : 10 et taille des plongements : 100)

Les résultats pour les *Annales de Géographie* 1950-1960 montrent une proximité avec les résultats précédemment obtenus pour le lemme « espace » sur le corpus lemmatisé. Autrement dit, à cette période, soit les deux dynamiques sémantiques des termes « espaces » et « espace » sont très proches, soit la sémantique du terme « espaces » domine et occulte celle du terme « espace ». Le lien **@ResultArtNonLemEspace50** permet de consulter les résultats obtenus pour le terme « espace » dans le corpus « ArticleNonLem » avec la méthode *GloVe* et les mêmes paramètres. Ces résultats permettent d'affirmer qu'en 1950-1960 dans les *Annales de Géographie*, ce n'est pas la sémantique du terme « espaces » qui domine et occulte celle du terme « espace », mais qu'il existe alors une proximité sémantique

[222] Consultables de manière plus complète avec ce lien : **@Result4Espaces50**.

des termes « espaces » et « espace ». À partir des années 1961-1971, la sémantique d' « espace » se différencie et s'affirme nettement en quittant cette base sémantique liée aux espaces naturels et bâtis sur laquelle reste globalement le terme « espaces ».

L'emploi précédent du corpus « ArticleLem » n'était donc pas dans ce cas optimal, car il y avait l'agrégation de ces deux dynamiques sémantiques méritant d'être différenciées. Il faut souligner que dans les cas où il y a assez peu d'occurrences du terme étudié et où les dynamiques sémantiques des formes au singulier et au pluriel sont proches, il y a à l'inverse un intérêt à utiliser le corpus lemmatisé puisqu'il y a une augmentation des données permettant d'avoir *a priori* des résultats plus signifiants. Il y a eu par conséquent sur cet aspect de l'usage des corpus une évolution de ma pratique de recherche avec le passage d'une conception marquée par une dichotomie « corpus de référence / autres corpus » à une approche plus réflexive tentant de sélectionner le corpus le plus adapté en fonction de la situation.

Peu après cette observation des dynamiques sémantiques différenciées d'« espace » et d'« espaces », j'ai lu l'ouvrage *Temps et récit* de Paul Ricoeur (1983). Dans le troisième tome, *Le temps raconté*, cet auteur revient sur l'évolution qu'a connue « le vocabulaire de l'histoire dans la deuxième moitié du XVIII^{ème} allemand ». Un fait marquant est le passage des « histoires de » à l'« histoire ». Paul Ricoeur souligne l'importance de ce processus en écrivant : « Les significations nouvelles souvent attribuées à des mots anciens vont servir ultérieurement à identifier l'articulation en profondeur de la nouvelle expérience historique, marquée par un rapport nouveau entre espace d'expérience et horizon d'attente » (Ricoeur 1983, 377). Une analogie avec le passage d'« espaces » à « espace » est réalisable[223]. Cette analogie révèle toute l'importance de ce changement sémantique. En la poursuivant (et en reprenant les formules de Paul Ricoeur), il est possible de penser que ces nouvelles significations ont pu servir ultérieurement à définir une nouvelle expérience géographique, « marquée par un rapport nouveau entre espace d'expérience et horizon d'attente »[224]. Il existe bien entendu des limites à cette analogie au sens où les études sur « les espaces de » n'étaient sûrement pas aussi développées que « les histoires de ». Toutefois, il y a dans cette analogie un fort potentiel pour redonner de la force à l'hypothèse sémantique réalisée par Olivier Orain en affirmant une révolution de la discipline qui s'appuie sur un « rapport

[223] De surcroît parce que le temps et l'espace sont deux catégories structurantes et problématiques dans le déploiement des SHS.

[224] L'espace d'expérience désigne « une structure feuilletée qui fait échapper le passé ainsi accumulé à la simple chronologie » (Ricoeur 1983, 376) ; quant à l'*horizon d'attente*, « c'est le futur-rendu-présent, tourné vers le pas-encore » (*Ibid* 1983, 379). Ces concepts sont repris par Paul Ricoeur à l'historien Reinhart Koselleck.

nouveau entre espace d'expérience et horizon d'attente » pour reprendre les termes de Paul Ricoeur.

Cette citation mérite à mon avis d'être mise au pluriel en affirmant plutôt : « des rapports nouveaux entre espaces d'expériences et horizons d'attentes ». Le conflit entre les tenants de la géographie appliquée et ceux de la géographie active illustre bien cette conception plurielle, aussi bien des espaces d'expériences que des horizons d'attentes. Cette remarque minore le caractère révolutionnaire, car il ne s'agit pas d'un remplacement d'une matrice par une autre, mais plutôt d'une multiplication des manières de penser, de faire et de revendiquer la géographie. De plus, ce rapprochement avec les réflexions de Paul Ricoeur sur l'histoire, ne valide en rien le transfert du schéma kuhnien. Les analogies développées n'ont pas valeur de « preuve ». Il n'est pas possible par exemple d'affirmer qu'il y a une « anomalie » telle que Thomas Kuhn la développe. Il n'est pas possible non plus à partir de ces analogies de soutenir que les quatre grands éléments de la matrice kuhnienne (les généralisations symboliques, la métaphysique, les valeurs et les exercices-types : *cf.* section Chap2.II.1.b) sont présents dans la géographie post-vidalienne.

Avant d'aborder ces points plus qualitativement (*cf.* Part3), il m'importe de développer d'autres résultats : ceux obtenus par exemple sur l'adjectif « spatial » permettent de mettre en évidence une modification sémantique quelque peu ultérieure à celle d'« espace ».

4. Réflexion à partir de l'évolution sémantique de « spatial »

Le corpus « ArticleLem » a été privilégié, car il y a dans les premières périodes (1960-1965 et 1966-1971) peu d'occurrences du lemme « spatial » et les différenciations sémantiques entre les formes au singulier et au pluriel sont peu significatives. La méthode *GloVe* et les paramètres précédemment utilisés ont été repris.

	1960-1965	1966-1971	1972-1975	1976-1982	1983-1988
Annales de Géographie	fonctionnel (0.37) expansion (0.36) répartition (0.35) maintien (0.35) uniformité (0.35) commission(0.34) particule (0.34) vide (0.34) arrosage (0.33) sécurité (0.32)	structure (0.5) répartition (0.45) modèle (0.44) signification (0.42) complexe (0.39) analyser (0.39) géographique (0.37) analyse (0.37) cohésion (0.36) circonscription (0.36)	organisation (0.53) structure (0.52) politique (0.51) processus (0.5) géographique (0.48) analyse (0.46) conséquence (0.45) économique (0.43) modèle (0.42)	certaine (0.54) concentration (0.54) géographique (0.53) espace (0.52) répartition (0.51) différent (0.51) dimension (0.5) structure (0.5)	organisation (0.63) structure (0.58) modèle (0.56) analyse (0.55) espace (0.51) distribution (0.49) social (0.48) échelle (0.47) forme (0.46) répartition (0.45)
L'Espace géographique			structure (0.59) géographique (0.51) attribut (0.5) analyse (0.49) considérer (0.48) phénomène (0.47) échelle (0.47) ensemble (0.47) organisation (0.47) économique (0.45)	géographique (0.54) social (0.54) échelle (0.53) dimension (0.51) modèle (0.51) structure (0.5) processus (0.49) phénomène (0.49) problème (0.48) analyse (0.47)	structure (0.59) organisation (0.56) comportement (0.53) espace (0.51) système (0.5) modèle (0.49) dynamique (0.49) analyse (0.48) interaction (0.47) géographique (0.47)

Tableau n°23 : Lemmes dont les vecteurs sont les plus proches du vecteur du lemme « spatial » dans le corpus ArticleLem.

(Méthode GloVe avec nombre minimum d'occurrences : 15, nombre d'itérations : 100, taille du contexte : 10 et taille des plongements : 100)

Ces résultats[225] montrent qu'il y a un changement sémantique important du lemme « spatial » dans les *Annales de Géographie* pendant les années 1966-1971. Ce changement est postérieur à celui observé pour « espace ». Pendant les années 1960-1965, « organisation de l'espace », « aménagement de l'espace », « espace urbain », « espace régional » sont des syntagmes utilisés sans que plusieurs termes liés à l'« analyse spatiale » s'affirment. L'expression d'« organisation spatiale » se développe plus tardivement (1972-1975). Il y a sûrement dans ces observations des explications de ce qu'Olivier Orain souhaitait mettre en avant quand il parlait d'une période de changement lexical sans changement sémantique.

[225] Consultables de manière plus complète avec ce lien : **@Result5Spatial50**.

Il faut ici souligner que toutes les expérimentations menées n'ont pas permis de mettre en évidence un changement sémantique important du milieu des années 1980 qui marquerait le passage d'un paradigme « réaliste » à des démarches constructivistes et nominalistes. Pour conclure cette première présentation de résultats, il est possible de remarquer que dans la « conjecture socio-sémantique » réalisée par Olivier Orain, il y a l'adjectif « socio ». Or, ce qui a été développé jusqu'ici est uniquement la dimension sémantique. Le dernier résultat présenté dans cette partie vise donc à combler cette lacune.

5. Réflexions à partir des auteurs ayant le plus utilisé « espace » et « spatial »

Comme cela a été précédemment justifié, la sémantique du terme « espace » a été étudiée à partir du corpus « ArticleNonLem » (pour ne pas prendre en compte les usages d'« espaces » qui relèvent d'une autre dynamique) alors que le lemme « spatial » a été étudié à partir du corpus « ArticleLem » (pour regrouper les différentes formes de l'adjectif). Si un article a été écrit par plusieurs auteurs, le nombre d'occurrences d'« espace » ou de « spatial » dans le document a été affecté à chaque auteur. Dans les tableaux suivants, n°24 et 25, les chiffres entre parenthèses après chaque auteur indiquent le nombre total d'occurrences trouvé pour cet auteur à la période et dans la revue en question. Ces tableaux reproduisent les résultats seulement pour les *Annales de Géographie* afin d'essayer de mieux comprendre ce qui explique la précocité des dynamiques sémantiques d' « espace » et de « spatial », notamment par rapport à ce qui est affirmé dans la thèse d'Olivier Orain.

	1960-1965	1966-1971	1972-1975	1976-1982	1983-1988
Annales de Géographie	P George (52) E Juillard (33) M Santos (10) C Chaline (9) M Rochefort (9) F Durand-Dastès (8) J Bastié (7) R Rochefort (6) R Lebeau (5) M Dubois (5)	B Kayser (55) Y Barbaza (53) J-B Racine (36) P George (26) P Claval (23) M Santos (20) C Chaline (18) F Cribier (15) C Camara (13) P Pinchemel (12)	H Isnard (69) W Hartke (43) A-L Sanguin (32) B Bret (32) M Grésillon (31) À Thibault (30) P Claval (27) J B-Garnier (24) C Manzagol (23) B Greer-Wootten (16)	H Isnard (108) A-L Sanguin (108) P George (92) J-O Simonetti (77) P Limouzin (65) H Bakis (49) G Schneier (48) A Fischer (40) R Lambert (33) S Liszewski (32)	G Di Méo (183) H Isnard (61) M Michel (44) M Sebti (40) V Rey (39) M-C Robic (39) F Doumenge (30) P Claval (25) R Pébayle (24) À Bourgey (24)

Tableau n°24 : Auteurs ayant le plus utilisé le terme « espace »
dans les *Annales de géographie.*

	1960-1965	1966-1971	1972-1975	1976-1982	1983-1988
Annales de Géographie	E Juillard (6) J Bastié (5) W Steigenga (5) J Labasse (4) Y Lacoste (4) M Rochefort (3) P George (3) P Riquet (2) P Georges (2) J Hautreux (2)	F Cribier (23) J-B Racine (18) P George (5) E Brutzkus (5) P Pinchemel (5) J Labasse (4) Y Lacoste (4) HSmotkine (3) M Santos (3) J Soppelsa (3)	A-L Sanguin (24) B Greer-Wootten (23) G M Gilmour (23) P George (11) J Dresch (10) J Beaujeu-Garnier (9) J-P Besancenot (9) M Rousset-Deschamps (9) M Santos (7) H Isnard (7)	À Fischer (75) M-C Maurel (34) J-O Simonetti (26) A-L Sanguin (21) P George (17) P Riquet (17) H Isnard (16) M Michel (13) À Bernard (11) G Schneier (10)	G Di Méo (40) V Rey (35) M-C Robic (35) M Michel (34) R Pébayle (30) M Sebti (29) J-P Donnay (22) A-S Bailly (15) B Marchand (14) J Malézieux (13)

Tableau n°25 : Auteurs ayant le plus utilisé le lemme « spatial »
dans les *Annales de géographie.*

Pour la période 1960-1965, dans les *Annales de Géographie*, Pierre George, et dans une plus faible mesure Etienne Juillard, apparaissent comme les deux principaux utilisateurs du terme « espace ». En regardant dans le détail, pour Pierre George, plus des trois quarts des occurrences sont situés dans un seul article paru en 1965 intitulé : *Géographie et urbanisme.* Dans ce papier, Pierre George rapproche la conception de l'espace des urbanistes de celle des géographes[226], tout en soulignant une différence importante : « Il reste qu'hétérogène par

[226] « Il est sans discussion que l'urbaniste prend acte d'un espace actuel qui est un produit de l'action des collectivités humaines sur un milieu naturel » (George 1965, 644).

nature, l'espace est un pour l'urbaniste » (George 1965, 644). Au contraire, pour les géographes, l'héritage d'une conception basée sur une approche différenciée des espaces est revendiqué. Dans le détail, la distinction entre l'espace géographique et urbanistique n'est pas si évidente. Pierre George développe et revendique un travail analytique et synthétique du géographe plus important[227]. L'urbaniste est présenté par comparaison, de manière assez caricaturale, comme saisissant l'espace dans « la globalité une de ses données » (George 1965, 644). Cette opposition n'est pas plus développée pour insister sur la complémentarité des deux approches. Le travail préalable du géographe assure « la transmission dans les meilleures conditions du "témoin" dans la course de relai entre géographie et urbanisme » (*Ibid* 1965, 648).

De la même manière, pour Etienne Juillard, la très grande majorité des occurrences d'« espace » provient d'un seul article : *La région, essai de définition* publié en 1962. Cet auteur détaille l'évolution conceptuelle se jouant alors autour du concept de « région ». Son introduction est une bonne illustration de la notion de plain-pied défendue par Olivier Orain autant sous la forme d'une critique de la géographie passée[228] que d'un désir pour la géographie à venir[229]. L'importance des réflexions menées en aménagement et en économie est soulignée dans l'évolution conceptuelle qui part de la « région-paysage » pour aller vers la région fonctionnelle. Les perspectives inverses de l'économie et de la géographie sont soulignées. La première part d'une modélisation qui se complexifie au fur et à mesure pour tenter de comprendre le particulier. L'économiste « aborde finalement, avec quelque gêne, un cas concret » (Juillard 1962b, 484). Le géographe prend appui *a contrario* sur des cas particuliers et essaye au fur et à mesure de généraliser et d'aller vers une forme d'abstraction. Même si les difficultés d'un dialogue entre les disciplines sont reconnues[230], c'est plutôt l'horizon d'une collaboration qui est présenté.

Il faut souligner que cette collaboration n'est pas exactement la même que celle présentée par Pierre George à propos de l'urbanisme. En effet, dans le cas d'Etienne Juillard, il y a à la fois la reconnaissance et le désir de transformation d'un concept central de la géographie traditionnelle, celui de « région ». À l'inverse, dans le cas de Pierre George, la perspective de coopération ne remet pas en cause l'outillage conceptuel de la géographie.

---

[227] « Chacun des éléments de l'espace est une donnée dont il [le géographe] doit tenir compte séparément dans son analyse avant de procéder aux synthèses successives qui conduisent à une vue globale de l'espace » (*Ibid* 1965, 644).

[228] « La région est le plus souvent conçue comme une sorte de "donné" dont on s'efforce au seuil de l'étude de justifier les limites » (Juillard 1962b, 483).

[229] « Grâce à la géographie générale, on met maintenant des réalités précises derrière des mots tels que pédiment, forêt-galerie, openfield, banlieue… On ne saurait en dire autant du mot région. » (*Ibid* 1962b, 483).

[230] « [les] différences d'optique et de vocabulaire ont rendu difficile le dialogue » (*Ibid* 1962b, 484).

En allant dans le sens de la distinction réalisée précédemment entre « espace » et « spatial », il est possible de faire remarquer qu'Etienne Juillard utilise de manière précoce cette forme adjectivale annonciatrice d'évolutions plus importantes à venir. Le faible nombre d'occurrences en question (6) et le danger d'interpréter trop le passé avec la connaissance de ce qui est advenu par la suite doivent ici être rappelés. Les occurrences de « spatial » pour Françoise Cribier (23) et Jean-Bernard Racine (18) sur la période suivante (1966-1971) me semblent plus représentatives. Elles sont d'autant plus intéressantes à analyser qu'elles représentent des situations restant marginales comme le montrent les nombres d'occurrences beaucoup plus réduits pour les auteurs suivants. Il est intéressant de remarquer que ces deux articles datent de 1971 et traitent tous deux de la sphère anglo-saxonne. Pour Françoise Cribier, l'article s'intitule *La géographie de la récréation en Amérique anglo-saxonne* et pour Jean-Bernard Racine, *Le modèle urbain américain, les mots et les choses*.

Ces deux articles présentent aussi des positionnements différenciés dans leurs conclusions. Françoise Cribier reconnaît plusieurs avantages aux approches anglo-saxonnes,[231] mais développe en contrepoint un nombre important de défauts : la « sécheresse » des approches, la faiblesse des études de cas et des descriptions ainsi que le manque de prise en compte des aspects sociaux et politiques. À l'inverse, l'approche de Jean-Bernard Racine est moins critique et met en avant le changement de paradigme qui a eu lieu de son point de vue aux États-Unis avec « l'ensemble des travaux particulièrement stimulants récemment fécondés par l'apparition de l'ordinateur » (Racine 1971, 425). Un parallèle avec les changements d'« épistémè » chez Michel Foucault est réalisé, bien souligné par le sous-titre choisi. (« les mots et les choses »). Jean-Bernard Racine en appelle au développement des modèles en concluant : « Pour une fois, pourquoi ne pas prendre le train en marche ? ». Remis dans le contexte des *Annales de Géographie* en 1971, le discours est très critique. Il est fort probable que la parution imminente de *L'Espace Géographique* ne soit pas étrangère au laissez-passer d'une telle charge critique. La différence de positionnement entre Françoise Cribier et Jean-Bernard Racine permet de comprendre ce qui a constitué le changement et, ce faisant, une partie de la thèse d'Olivier Orain.

Une part importante du discours sur l'« espace » dans les *Annales de géographie* pour cette période 1966-1971 provient de sources différentes : l'aménagement rural avec Bernard Kayser, le tourisme avec Yvette Barbaza. Les géographes qui participeront de manière forte à la « rénovation » de la géographie sont présents, mais faiblement représentés : Jean-Bernard Racine, Paul Claval, Yves Lacoste et François Durand-Dastès. Les sémantiques

[231] « la clarté des exposés - on explique ce qu'on cherche, on explicite sans fausse pudeur les hypothèses, les raisonnements -, la compétence dans l'utilisation des techniques de traitement des informations. L'approche n'est pas facile, mais elle vaut la peine et le profit qu'on en tirera est indéniable » (Cribier 1971, 660).

d'« espace » sur la période 1961-1965 et 1966-1971 et de « spatial » sur 1966-1971 ne sont donc pas uniquement portées par leurs discours. Mise à part l'exception notable de Jean-Bernard Racine, ce discours de la « nouvelle géographie » n'est d'ailleurs pas présent sur ces périodes dans cette revue. Revenir aux textes même permet de comprendre que la sémantique autour de l'« organisation de l'espace » n'est pas si développée sans relever toutefois de l'exceptionnel. Le mouvement de rénovation des études régionales (René Lebeau, Michel Rochefort…) et de l'aménagement du territoire (Etienne Juillard, Bernard Kayser…) a été pris en compte par les *Annales de géographie.*

Sur la période 1972-1975, le cas d'Hildebert Isnard mérite aussi d'être développé, car son positionnement sur ce concept d'espace a abouti à la parution d'un livre intitulé *L'espace géographique* paru en 1978. Son positionnement est celui d'une relecture de la géographie humaine traditionnelle, non pas dans une perspective théorico-quantitativiste mais dans une orientation systémique de réactualisation de la combinaison vidalienne. La prise de position est d'autant plus intéressante qu'Hildebert Isnard (1904-1983) a dirigé des travaux quantitatifs et théoriques. Il s'agit d'une position réfléchie qui est discutée par Jean-Bernard Racine et Antoine S Bailly dans un article paru en 1979 dans la revue *L'Espace Géographique*. Dans l'expression de leurs différences se retrouve un point faisant incontestablement partie de la thèse d'Olivier Orain :

> « N'oublions pas en effet l'essentiel des approches nouvelles : un changement d'attitude face au monde étudié, l'adoption d'une démarche qui ne se réduit pas à la seule observation, mais qui suppose d'abord une interrogation d'ordre théorique puis l'emploi d'une démarche transparente et rigoureuse s'appuyant au besoin sur la formalisation mathématique et le test statistique des hypothèses » (Bailly et Racine 1979, 290).

Toutefois, la fin de l'article de Jean-Bernard Racine et d'Antoine Bailly insiste sur les qualités de l'ouvrage d'Hildebert Isnard, notamment la présence des nuances « grâce au choix des exemples et à la finesse des observations géographiques ». Les auteurs concluent alors : « Même si nous parcourons actuellement d'autres horizons, nous ne dirons jamais assez à quel point nous avons encore besoin de ces qualités-là. L'affirmer, n'est-ce pas avouer que nous les avons peut-être perdues, faute de les pratiquer ? » (Bailly et Racine 1979, 291). Cette reconnaissance finale remet partiellement en question l'« incommensurabilité » des deux approches. Il est en effet difficile d'imaginer qu'un nouveau paradigme reconnaisse si ouvertement les qualités de ce qui est censé fonder le paradigme précédent. Imagine-t-on un physicien einsteinien dire qu'il regrette les bases perdues de la physique newtonienne ?

Sur les autres auteurs utilisant le plus « espace » en 1972-1975 dans les *Annales de géographie*, soulignons que Wolgang Hartke et André-Louis Sanguin enseignent alors hors

de France. Une remarque similaire peut être réalisée avec les auteurs utilisant le plus le lemme « spatial » dans cette revue à la même époque. Au-delà de ces auteurs étrangers ou enseignants à l'étranger, les utilisations de ce lemme restent assez faibles, ce qui relativise le poids de l'évolution sémantique précédemment observée. Un début de changement intervient à la période suivante, notamment avec le cas assez isolé d'André Fischer (géographe économiste). Les années 1983-1988 se caractérisent par un renouvellement avec l'arrivée en tête de classement de nouveaux auteurs comme Guy Di Méo (promoteur important d'une géographie sociale) et Violette Rey. Ces acteurs ne font plus partie de la génération 1930 (Bataillon, 2009) qui a porté et revendiqué le changement. L'idée d'un apaisement des conflits avec le début d'une nouvelle période telle que la dessine Olivier Orain peut ici se retrouver.

Ces analyses mobilisant les textes mêmes montrent que les résultats précédemment obtenus doivent être relativisés. Les changements sémantiques du terme « espace » sur la période 1961-1965 et du lemme « spatial » sur la période 1966-1971 dans les *Annales de géographie* reposent sur assez peu d'auteurs et de textes. Une hypothèse très probable, mais qui n'est pas démontrée dans cette thèse, est que cette observation provient des corpus utilisés. Les évolutions mises en avant s'expliquent alors plus largement par des dynamiques marquant la géographie française des années 1960 (Claval et Sanguin 1996). Parmi ces dernières, il est possible de citer :

- Un renouvellement fort des thèses régionales avec notamment celles de Michel Rochefort en 1960 et de Raymond Dugrand en 1963.
- L'essor de l'« économie spatiale » avec le travail par exemple de Paul Claval sur la *Géographie générale des marchés* en 1962.
- Le développement des pensées aménagementistes durant l'ensemble des Trente Glorieuses (1945-1975).

Une petite partie de ces évolutions a percolé dans les *Annales de Géographie.* Malgré cette percolation réduite, la méthode utilisée est assez efficace pour rendre compte tout de même des changements sémantiques liés à ces évolutions. Par conséquent, même si en remontant aux textes, il peut être objecté que les évolutions sémantiques mises en avant ne tiennent qu'à quelques auteurs, l'hypothèse avancée (et étayée à partir des grandes dynamiques précédemment citées) est que ces évolutions sémantiques sont représentatives d'évolutions importantes qui ne sont qu'en partie restituées par les *Annales de Géographie*.

Face à cette hypothèse, il est probable qu'un partisan d'une vision kuhnienne de la géographie française ne manquerait pas de faire remarquer que l'idée de percolation réduite

peut être remplacée par celle d'un « filtre » rapprochable d'une résistance de l'« ancien paradigme ». Sur ce point, il y a eu sans conteste des résistances qui ont pu jouer un tel rôle de « filtre », mais cette reconnaissance ne valide pas pour autant une lecture kuhnienne de la géographie française. De plus, il est possible d'objecter que le « filtre » a tout de même laissé passer suffisamment d'informations permettant à la méthode utilisée de détecter des changements sémantiques.

Par rapport à un tel échange d'arguments[232], la conclusion à laquelle je suis arrivé est qu'il n'est possible de venir à bout de la controverse initiale à partir des analyses quantitatives réalisées.

Le problème rencontré est formalisable sous la forme d'une triple inférence :

- La première repose sur le passage de l'observation des termes ayant les vecteurs les plus proches d'un terme donné, à la mise en évidence de continuités ou de ruptures sémantiques. Cette inférence a été très peu discutée précédemment. Elle s'appuie tout d'abord sur une pratique commune qui consiste à induire d'une proximité entre les vecteurs issus d'un même plongement de mots, une proximité sémantique des termes représentés par ces vecteurs. Le terme « induction » est ici utilisé, car il n'existe pas de démonstration d'une telle relation. Ensuite, la mise en évidence de continuités ou de ruptures sémantiques repose sur l'observation d'un changement ou d'un maintien des termes obtenus d'une époque sur l'autre. Cette approche a été formalisée par Gonen *et al.* (2020) en termes d'intersection des k plus proches voisins, mais à la différence de leur travail, cette thèse a jusqu'ici utilisé une approche très qualitative. La raison expliquant cette approche qualitative tient à l'existence fréquente de termes proches sémantiquement, mais qui ne sont pas exactement les mêmes. L'analyse qualitative permet alors de faire face plus facilement à ce problème. Par exemple, des termes comme « urbain », « quartier » et « immeuble » peuvent être ainsi regroupés dans un ensemble commun « lié aux espaces bâtis ». Si à l'époque suivante, un groupe sémantique est composé de termes différents, mais reste globalement lié aux espaces bâtis, l'analyse qualitative permet de facilement établir une continuité en mettant en avant le changement de vocabulaire. Une question qui peut être légitimement posée est celle du remplacement de ces analyses qualitatives par des approches quantitatives permettant de se passer de

[232] Pouvant vite tourner autour du problème du verre à moitié vide ou à moitié plein.

la subjectivité de l'analyste et d'objectiver encore plus les continuités et les discontinuités. Cet objectif est abordé dans les sections suivantes (*cf.* sections Chap7.III.2 et 3).

- La deuxième inférence consiste à partir des évolutions sémantiques précédentes (continuité ou rupture) à affirmer des évolutions cognitives. Cette inférence a été discutée dans les analyses déjà réalisées des résultats, notamment avec la reconnaissance que l'hypothèse sémantique pouvait être invalidée sans forcément discréditer les évolutions cognitives mises en avant par Olivier Orain. Il y a là un problème majeur, car si l'évolution du sens des mots n'est pas sans lien avec celle des idées, il n'existe pas, d'un point de vue formel, de déduction possible permettant d'affirmer à partir de l'ampleur d'un changement sémantique, l'importance d'un changement cognitif (simple continuité, évolution, rupture…). Il y a toujours discussion pour induire à partir des changements sémantiques des évolutions cognitives probables en essayant de comparer et de revenir aux textes.
- La troisième inférence consiste à passer des évolutions cognitives inférées à la validation du schéma kuhnien. Cette inférence a fait l'objet d'un scepticisme et d'une critique de fond dans ce travail. S'il y a indéniablement certains points de la thèse d'Olivier Orain qui peuvent être soutenus à partir des résultats obtenus, les utiliser pour valider une lecture kuhnienne reste toujours problématique. Par exemple, l'importance de la rencontre avec la scène aménagementiste pour comprendre l'évolution sémantique du terme « espace » est facilement soutenable à partir du texte d'Etienne Juillard précédemment analysé. *A contrario*, l'idée de rapprocher cette rencontre d'une « anomalie » kuhnienne est tout à fait discutable et indémontrable.

Cette triple inférence permet de comprendre la reconnaissance d'indéterminations. Les résultats précédents permettent d'appréhender tout de même plusieurs évolutions sémantiques différenciées (« espaces », « espace » et « spatial ») qui ne concordent pas et ne valident pas les dynamiques sémantiques mises en avant par Olivier Orain. Toutefois, ces résultats quantitatifs ne permettent pas non plus de conclure directement à la validité ou à l'invalidité de la lecture kuhnienne.

La section suivante s'appuie sur les limites précédemment développées de la première inférence. L'objectif poursuivi est celui de l'objectivation de la reconnaissance des groupes sémantiques et de leurs évolutions. Cette formulation résulte d'un long processus qui mérite d'être détaillé, car il permet de comprendre la suite de la recherche effectuée.

III. Approche réticulaire et évolution de groupes sémantiques

Une évolution majeure dans la forme des résultats produits pour ce doctorat est liée à un travail[233] mené au laboratoire ERIC pour visualiser les plongements de mots sous forme réticulaire et interactive.

1. Première exploration privilégiant une forme réticulaire

En m'inspirant de ce travail, j'ai construit un module dans l'application permettant de choisir les résultats d'un plongement, un terme et un paramètre n. Les n-termes (ou n-lemmes suivant le corpus) dont les vecteurs sont les plus proches du terme choisi dans les résultats sélectionnés sont affichés sous forme d'un réseau avec au centre le terme de référence et un lien pour chacun des n-termes. Par exemple, à partir des résultats issus du corpus « ArticleLem »[234] et de la méthode *GloVe* avec les paramètres précédemment utilisés[235], en choisissant la période 1950-1961 des *Annales de Géographie,* le lemme « espace » et une valeur de 10 pour le paramètre n, la représentation obtenue est la suivante.

[233] Il s'agit d'un stage réalisé par Masmoudi Abderrahemen sous la direction d'Adrien Guille. Son travail est disponible à cette adresse : https://github.com/Hoshun112/PRe (consulté le 18/09/2023).

[234] Malgré le fait d'avoir affirmé précédemment qu'il était préférable d'utiliser le corpus « ArticleNonLem » pour étudier la dynamique sémantique d'« espace », le corpus « Article Lem » a été privilégié pour cette présentation. En effet, en prenant le corpus « ArticleNonLem », des articles ou des prépositions rendent les résultats moins évidents. L'objectif étant dans cette partie de présenter le parcours de recherche et non d'analyser des résultats, j'ai privilégié ici le corpus « ArticleLem ».

[235] 15 pour le minimum d'occurrence, 100 pour le nombre d'itérations, 10 pour la taille du contexte et 100 pour la taille du plongement.

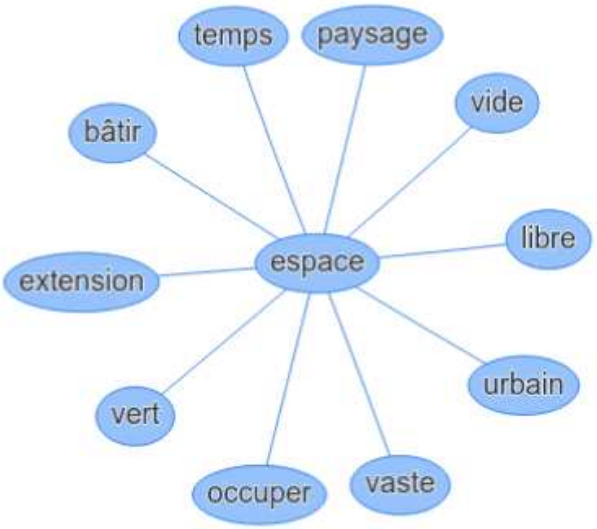

Figure n°12 : Exemple de représentation réticulaire des lemmes dont les vecteurs sont les plus proches du vecteur du lemme « espace ».

L'intérêt est que cette représentation peut être répétée pour chacun des nouveaux lemmes. Par exemple, si à partir de cette Figure n°12, la représentation est développée pour les lemmes « bâtir », « urbain », « occuper » et « vaste », la figure suivante est obtenue.

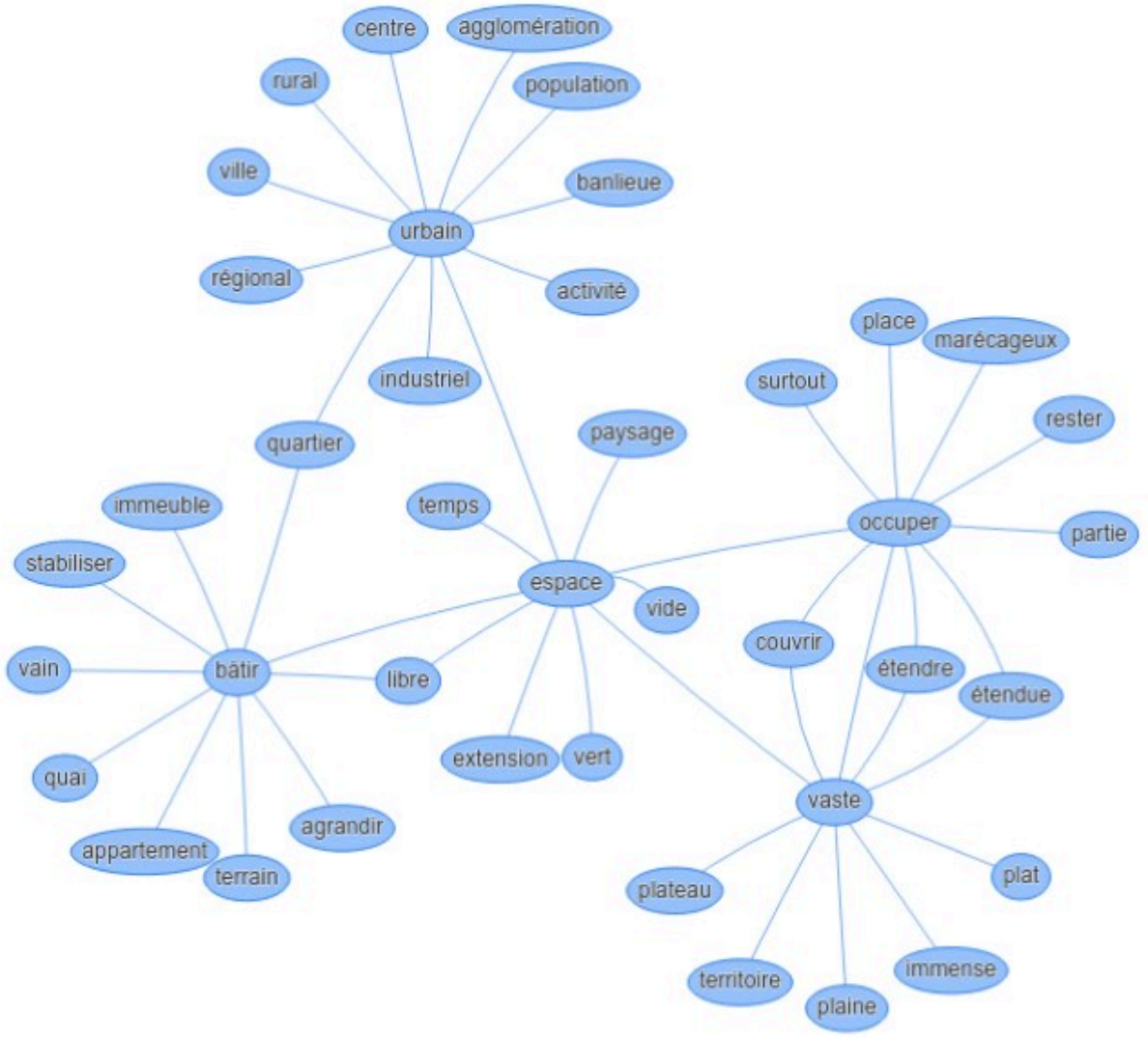

Figure n°13 : Exemple de représentation de lemmes dont les vecteurs sont les plus proches des vecteurs des lemmes « espace », « urbain », « bâtir », « occuper » et « vaste ».

Sur cette figure, deux ensembles sémantiques reliant les termes ouverts peuvent être identifiés. D'un côté, « quartier » reliant « bâtir » et « urbain ». De l'autre, « étendre », « couvrir », « étendue » reliant « occuper » et « vaste ». Le premier ensemble est directement lié à l'espace bâti. Le deuxième peut être rapproché de l'espace « naturel » en observant les termes périphériques (notamment « plaine », « plateau », « marécageux »). Si cette observation est en adéquation avec ce qui a été précédemment affirmé, elle est scientifiquement faible, car rien ne justifie en amont l'ouverture de ces lemmes spécifiques.

Il faut souligner qu'une telle exploration est particulièrement interactive puisqu'il est possible en cliquant sur des lemmes ouverts de les refermer.

Deux raisons m'ont conduit à reprendre ce travail pour l'adapter à l'application construite pour cette thèse :

- La première est d'offrir au lecteur la possibilité de choisir son propre terme de départ et d'explorer les réseaux à sa guise. L'exemple précédemment développé est accessible avec le lien suivant : **@ExFigureRetiDev**
- La seconde raison est qu'Olivier Orain utilise dans sa thèse la formule de « constellations terminologiques » (Orain 2003, 355). Il est évident que ces formes réticulaires se rapprochent beaucoup plus de cette idée que les précédents tableaux. Cette réflexion m'a conduit à essayer d'approfondir cette direction de recherche. Une piste que j'ai suivie est d'ouvrir tous les premiers lemmes obtenus et de comparer les graphiques à différentes époques pour les deux revues. Par exemple, en partant de la Figure n°13, la représentation alors obtenue est la suivante : **@ExFigureRetiDevPlus**

Mais, j'ai abandonné cette piste pour plusieurs raisons. Tout d'abord, d'un point de vue méthodologique, il n'était pas facile d'analyser et de comparer les figures alors obtenues à différentes époques pour les deux revues. Ensuite, le mode de construction (qui valorise les n plus proches voisins, puis les n plus proches voisins de ces n plus proches voisins) ne m'a pas semblé optimal et difficilement justifiable. Ce mode de construction reposant sur un motif de base[236] et l'objectif étant de mettre à jour des structures sémantiques, il existe un risque de biais, car rien n'indique que ce motif de base soit pertinent. Cette réflexion m'a conduit à m'orienter plutôt vers des méthodes de clustering (regroupement) appliquées aux résultats seulement de premier ordre (les n plus proches voisins) et non de second ordre (les n plus proches voisins des n plus proches voisins)[237].

L'objectif, en réalisant un clustering sur les vecteurs (issus d'une méthode de plongement de mots) des n plus proches voisins d'un terme choisi, est d'obtenir sans l'intermédiaire d'analyses qualitatives, différents groupes sémantiques caractéristiques du terme choisi. Par exemple, si les 6 termes obtenus les plus proches d'« espace » sont « urbain », « rural », « immeuble », « champ », « quartier », « agriculture », l'objectif est d'obtenir deux groupes : le premier rassemblant «les termes « urbain », « immeuble » et « quartier » ; le second regroupant « rural », « champ » et « agriculture ».

[236] Ce motif de base est un réseau basique en étoile à n branches.

[237] Si les méthodes faisant appel à cet ordre de rang 2 se justifient tout à fait dans le cadre de l'utilisation de méthodes de cooccurrence pour saisir des rapprochements sur l'axe paradigmatique (*cf.* section Chap3.IV.1), elles sont beaucoup moins justifiables dans le cas de plongements de mots qui prennent déjà en compte cet axe.

2. Seconde phase d'explorations sur les méthodes de clustering

Il existe un très grand nombre de méthodes de clustering. Mes recherches pour déterminer la méthode la plus adéquate de clustering m'ont reconduit au double problème de l'évaluation et de l'absence de consensus dans la littérature. De la même manière que pour le choix de la méthode de plongement, il m'est impossible de pouvoir revendiquer l'obtention des meilleures « constellations sémantiques », si tant est que ces dernières existent[238]. Ce travail reconnaît par conséquent un travail incomplet sur l'optimisation des méthodes de clustering. Les expérimentations ici présentées s'appuient sur une des méthodes les plus basiques de clustering : la Classification Ascendante Hiérarchique (CAH).

Au départ d'une CAH, il y autant de clusters[239] que de termes. À chaque itération, les deux clusters les plus proches sont agglomérés dans un cluster composite. À la fin de l'algorithme, il n'existe plus qu'un seul cluster avec tous les termes. Cette méthode nécessite donc à chaque itération de calculer la distance entre les clusters, qu'ils soient composés d'un seul terme ou de plusieurs. Le calcul ici utilisé pour les distances est celui des similarité-cosinus entre les vecteurs des termes.

[238] Dans des sémantiques différentielles comme celle de François Rastier, il y a la reconnaissance d'une pluralité de configurations sémantiques. Il n'y a pas un ou plusieurs sens préexistants qu'il suffirait de trouver pour expliquer totalement un texte. *A contrario*, ces sémantiques insistent sur la possibilité de trouver toujours des nouveaux sens en mettant à jour des configurations encore peu remarquées.

[239] Le terme de « cluster » est équivalent à celui de « classe ».

Quand au moins un des deux clusters comparés est composé de plusieurs termes, plusieurs stratégies sont possibles pour définir la distance entre les deux clusters :

- Celle du « saut minimal » privilégie la distance entre les deux termes les plus proches et peut être représentée ainsi :

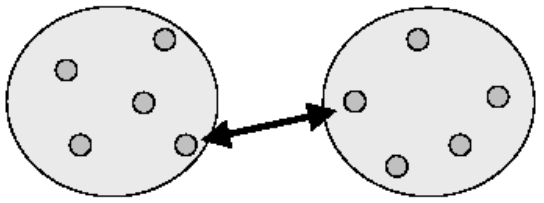

Figure n°14 : Représentation simplifiée du « saut minimal ».

- Celle du « saut maximal » privilégie la distance entre les deux termes les plus éloignés :

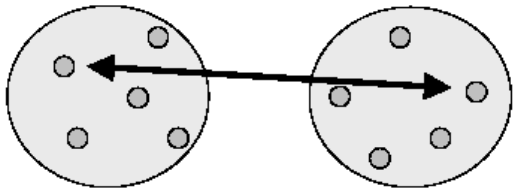

Figure n°15 : Représentation simplifiée du « saut maximal ».

- Celle du « saut moyen » privilégie la distance moyenne entre les deux clusters :

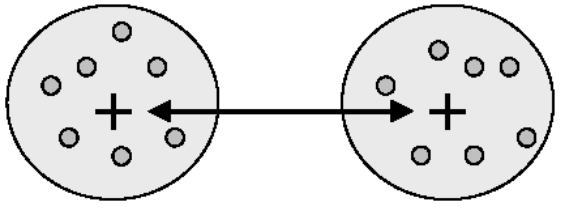

Figure n°16 : Représentation simplifiée du « saut moyen ».

Le résultat des agglomérations successives peut être présenté sous la forme d'un dendrogramme. En fonction du nombre de clusters souhaité par l'utilisateur, l'arbre du dendrogramme peut être « coupé ». Par exemple, la figure suivante présente le résultat en prenant comme base les 10 premiers lemmes les plus proches d'« espace » avec le corpus « ArticleLem » sur les *Annales de Géographie* 1950-1960 (obtenus avec la méthode *GloVe* et les paramètres précédemment privilégiés), en utilisant la stratégie du « saut maximal » et en demandant 3 clusters en sortie.

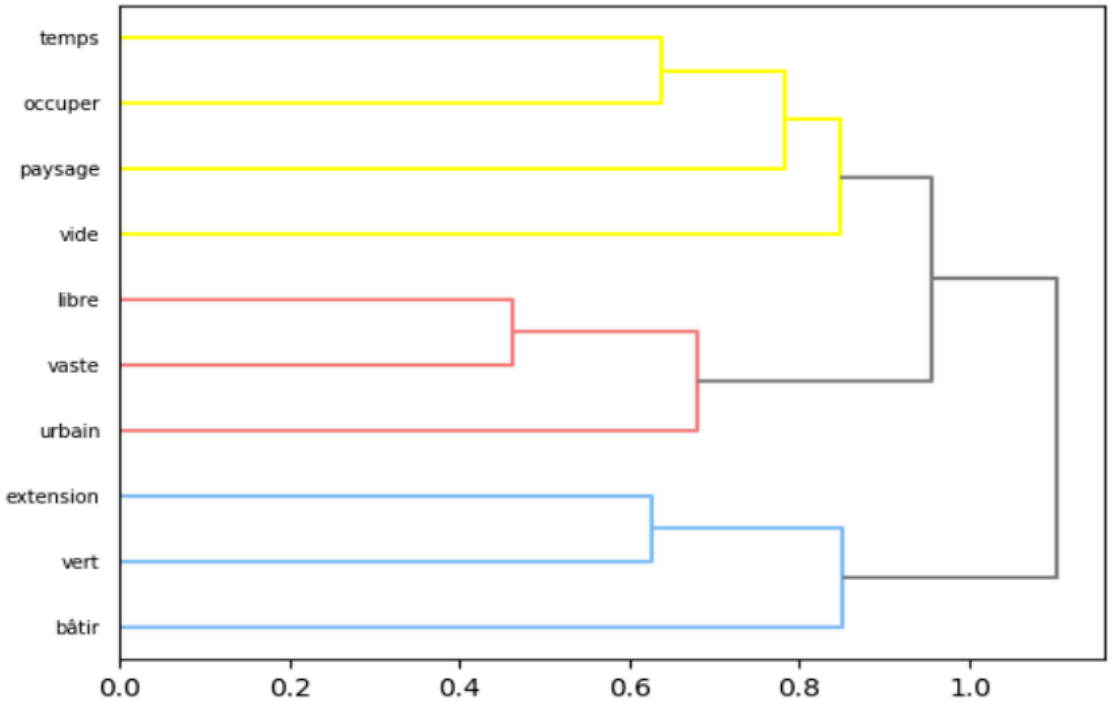

Figure n°17 : Partition en 3 clusters des 10 premiers lemmes dont les vecteurs sont les plus proches d' « espace » en utilisant le « saut maximal ».

(sur les *Annales de Géographie* 1950-1960 avec la méthode *GloVe* avec nombre minimum d'occurrences : 15, nombre d'itérations : 100, taille du contexte : 10 et taille des plongements : 100)

Dans cette figure[240], la distinction entre un pôle lié aux « espaces bâtis » et un pôle lié aux « espaces naturels ou agricoles » précédemment mis en avant (*cf.* section Chap7.III.1) ne se retrouve pas. Un retour dans les textes permet de comprendre cette situation. Les études rurales traitent souvent du « bâti ». De plus, les études urbaines peuvent utiliser les adjectifs « libre » et « vaste » pour désigner certaines zones. Ainsi, les deux pôles précédemment construits sont moins sémantiquement disjoints que cela pouvait être initialement pensé. Il

[240] Consultable aussi à partir de ce lien (en bas de page) : **@PresentDendoSautMax**.

est possible de trouver une spécificité sémantique au cluster formé par les lemmes « temps », « occuper », « paysage » et « vide ». Le rapprochement notamment des lemmes « temps » et « paysage » est intéressant, au sens où, comme l'affirme Jean-Charles Filleron (2008), les deux concepts sont intimement liés. Toutefois, là encore, le registre sémantique de ce cluster est loin d'être totalement disjoint des deux autres.

Cette figure permet également d'observer ce qui aurait été obtenu si un nombre supérieur ou inférieur de clusters avait été demandé[241]. Faut-il par exemple augmenter le nombre de classes pour séparer le lemme « bâtir » du cluster « extension » et « vert » ? Cette réflexion pose la question du nombre optimal de classes. Il existe sur ce sujet une pluralité de méthodes (*Cubic Clustering Criteria* , *pseudo-t², pseudo-F...*). Elles produisent dans certains cas des résultats concordants ; dans d'autres cas, elles produisent des résultats différents reposant la difficulté de l'évaluation. Le problème est qu'il existe au niveau des significations sémantiques de nombreuses zones d'indéterminations et d'interpénétrations où l'établissement de classes est particulièrement complexe et peut donner lieu à des propositions multiples. Face à cette situation, j'ai décidé de ne pas accorder trop d'importance aux résultats d'une détection automatique du nombre optimal de clusters, mais plutôt de laisser la prise de décision à l'utilisateur avec un accès au dendrogramme de construction pour qu'il puisse disposer d'un outil de réflexion.

L'objectif de cette section étant de justifier le choix réalisé dans cette thèse par rapport aux trois stratégies possibles d'agrégation (saut « maximal », « moyen » et « minimal »), les liens suivants permettent de consulter les dendrogrammes obtenus pour 30 lemmes et 4 clusters suivant ces trois stratégies : **@SautMin30term4clust**, **@SautMoy30term4clust** et **@SautMax30term4clust**. Pour des raisons de lisibilité et les conclusions étant les mêmes, la figure suivante présente le résultat obtenu pour 10 lemmes et 3 clusters en utilisant cette fois-ci la stratégie du « saut moyen ».

[241] Il faut pour cela se déplacer dans le dendrogramme à partir de l'endroit où la « coupe » a été effectuée. En allant vers la droite, il est possible de savoir quels clusters sont agrégés si une demande inférieure en nombre est réalisée. À l'inverse, en allant vers la gauche, il est possible de savoir quels clusters sont divisés si une demande supérieure en nombre est réalisée.

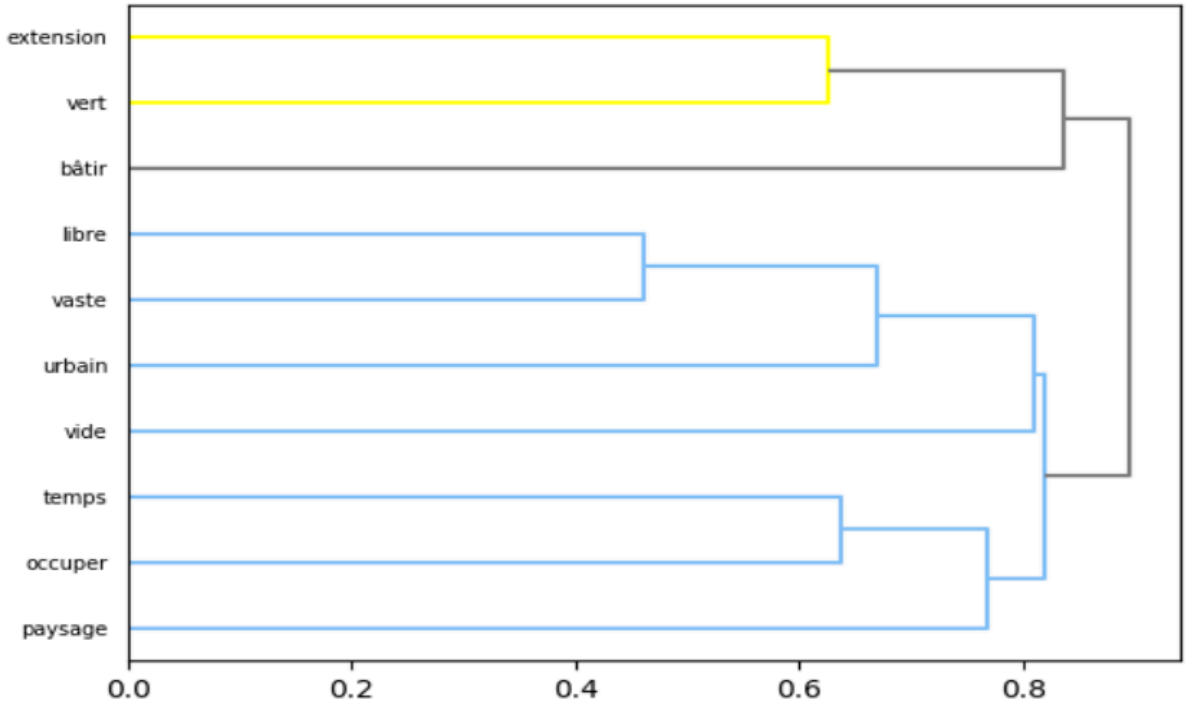

Figure n°18 : Partition en 3 clusters des 10 premiers lemmes dont les vecteurs sont les plus proches d' « espace » en utilisant le « saut moyen ».

(toujours sur les *Annales de Géographie* 1950-1960 avec la méthode *GloVe* avec minimum d'occurrences : 15, nombre d'itérations : 100, taille du contexte : 10 et taille des plongements : 100)

Dans ce découpage[242], le lemme « bâtir » forme un cluster à lui tout seul. Il faut souligner que la valeur expliquant la différenciation des clusters « extension » et « vert » d'un côté, et « bâtir » de l'autre, est très proche de celle qui aurait divisé ensuite « libre », « vaste », « urbain » et « vide » d'un côté, et « temps », « occuper », « paysage » de l'autre. Du fait de cette proximité, ce découpage peut être considéré comme plus fragile que celui effectué précédemment.

Les expérimentations menées m'ont conduit à avancer une autre idée qui repose sur les profils des clusters obtenus. En effet, l'utilisation du « saut maximal » tend à produire des clusters de taille proche (par exemple, dans le résultat de la Figure n°17, 3 à 4 lemmes par clusters) alors que le « saut moyen » tend à produire des configurations plus déséquilibrées (par exemple, dans la Figure n°18, un cluster composé de 7 lemmes et deux petits clusters

[242] Consultable aussi à partir de ce lien (en bas de page) : **@PresentDendoSautMoyen**.

composés respectivement de deux et d'un seul lemme). Le « saut minimal » tend à produire des configurations encore plus déséquilibrées[243].

Un tel profil est problématique pour la section suivante, à savoir la comparaison des clusters dans le temps. En effet, ce profil déséquilibré a de grandes chances de se répéter sur les autres époques. Le cluster composé d'un maximum de termes d'une époque donnée a alors une probabilité importante de partager un lexique commun avec le cluster composé d'un maximum de termes d'une époque de l'époque suivante. Pour les clusters composés d'un seul terme : soit ce dernier se retrouve à l'époque suivante, la continuité est alors totale ; soit ce dernier n'apparaît pas à l'époque suivante, la discontinuité est alors totale. L'avantage de clusters plus équilibrés est de répartir plus équitablement les probabilités d'existence de lexique en commun entre clusters d'époques successives. Ceci explique le choix de la distance maximale pour le travail ultérieur de suivi des clusters dans le temps.

3. Exploration diachronique de suivi des clusters

Cette exploration est basée sur un outil existant : *Diachronic'Explorer*[244]. Mes expérimentations m'ont toutefois conduit à plusieurs changements au niveau du modèle théorique et de la visualisation. Pour comprendre les modifications apportées, il est nécessaire de remonter à la formulation originelle de cet outil.

a. Le modèle originel et ses variantes

La théorie sur laquelle est fondé cet outil a été développée par Jean-Charles Lamirel et Shadi Al Shehabi (2006). L'objectif est d'utiliser le théorème de Bayes pour suivre l'évolution de clusters dans le temps. Pour comprendre le fonctionnement, imaginons un cluster C1 à une époque n défini par plusieurs mots (m_i, m_{i+1}…) avec pour chaque mot m_i

[243] Le résultat initialement obtenu **@PresentDendoSautMin** a permis de s'apercevoir d'un défaut de l'application. Il est en effet possible de penser en regardant le dendrogramme qu'il n'y a que deux clusters. Or, cela provient d'une mauvaise coloration des groupes n'ayant qu'un lemme. En réalité, « vide » et « paysage » forment deux clusters séparés. Une recherche a été réalisée pour supprimer cet effet induit. Elle a donné lieu à un résultat (**@PresentDendoSautMinColorSingleton**) qu'à moitié satisfaisant. En effet, les branches des clusters constitués d'un seul terme sont bien colorisées, mais un effet induit est facilement observable avec la colorisation (non souhaitée) de la liaison entre ces clusters. Il aurait fallu un travail beaucoup plus important pour résoudre ce problème. Il n'a pas été jugé prioritaire pour cette thèse, car l'objectif était avant tout d'étudier des continuités et des discontinuités sémantiques dans le temps. La réflexion et le travail réalisés expliquent la présence de l'option « ColorSingleton » dans l'application que l'utilisateur peut choisir s'il souhaite colorer les branches qu'avec un seul terme tout en ayant connaissance de l'effet induit précédemment mentionné.

[244] https://github.com/nicolasdugue/istex-demonstrateur (consulté le 18/09/2023).

une probabilité associée p_i. Dans l'implémentation de *Diachronic'Explore*r réalisée par Nicolas Dugué, ces probabilités correspondent à des moyennes harmoniques des mesures de rappel et des mesures de prédominance (Dugué, Lamirel, et Cuxac 2016a). Ces calculs n'ayant pas été utilisés dans cette thèse, ils ne sont pas ici détaillés.

À l'époque n+1, un cluster C2 est caractérisé par les mots (m_j, mot_{j+1}…) et par les probabilités suivantes : (p_j, p_{j+1}…). La perspective bayésienne vise à considérer une activation Act qui est réalisée par un cluster (qui est alors appelé la source) sur un autre (qui devient la cible). Le théorème de Bayes permet d'écrire :

$$P(Act|C) = \frac{P(Act \cap C)}{P(C)}$$

Intuitivement, il est plus facile de penser à une activation de C2 par C1 (sens de la temporalité), mais le modèle théorique s'autorise à réaliser cette activation dans les deux sens. En choisissant C2 comme cluster activé, le numérateur P(Act ∩ C2) est égal à la somme des probabilités dans C2 des mots présents à la fois dans C2 et C1. Le dénominateur est égal à P(C2), c'est-à-dire la somme des probabilités de tous les mots présents dans C2.

Le premier problème est que cette formulation bayésienne originelle n'a pas été strictement reprise pour concevoir et implémenter *Diachronic'Explorer*. Il y a une reformulation de P(Act | C2) en P (t | s), t étant le cluster cible et s le cluster source (Dugué, Lamirel, et Cuxac 2016b). Une telle reformulation est certes possible. Il existe toutefois un second problème qui provient du sens de la formule utilisée pas les auteurs.

$$P(t|s) = \frac{P(t \cap s)}{P(t)}$$

En effet, une application stricte du théorème de Bayes devrait aboutir à la formule suivante :

$$P(t|s) = \frac{P(t \cap s)}{P(s)}$$

Mes explorations m'ont conduit à trouver une interprétation au calcul réalisé, mais à sortir d'une approche bayésienne. En effet, la formule $\frac{P(t \cap s)}{P(t)}$ peut être lue comme le calcul d'un

indicateur s'interprétant comme le ratio du poids des mots en commun entre les clusters source et cible par rapport au poids de l'ensemble des mots du cluster cible.

La reformulation suivante en termes non plus de probabilités, mais de poids peut alors être effectuée :

$$I(s) = \frac{\sum_{mi \in s \cap t} w_i}{\sum_{mi \in s} w_i}$$

$$I(t) = \frac{\sum_{mi \in s \cap t} w_i}{\sum_{mi \in t} w_i}$$

La notation w_i a été utilisée pour bien marquer qu'il s'agit du poids du m_i dans le cluster et non plus la probabilité du mot dans le cluster. Le cluster source s fait dans cette formulation uniquement référence à un des clusters de l'époque n et le cluster cible à un des clusters de l'époque n+1. J'illustre ci-dessous le calcul des indicateurs I(s) et I(t) en utilisant un exemple tiré de l'article (Beligné, Lefort, et Loudcher 2020a).

élevage (0,18), bétail (0,16), culture (0,16), **animal (0,14)**	⇨	**élevage (0,15), animal (0,14),** machine (0,13), pré (0,11)
Un exemple de cluster source		*Un exemple de cluster cible*

En gras, les mots en commun permettent de comprendre les calculs suivants :

I(s) = (0,18 +0,14) / (0,18 +0,16 +0,16 +0,14)= 0,32 /0,64 = 0,5

I(t) = (0,15 + 0,14) / (0,15 + 0, 14 + 0,13 + 0,11) = 0,29 / 0,53 = 0,56

Les résultats sont compris théoriquement entre 0 et 1. Il faut souligner que dans la pratique, à partir les exemples sur lesquels j'ai travaillé, les valeurs obtenues pour I(s) et I(t) sont statistiquement le plus souvent faibles. Un score global de 0,53 (moyenne de I(s) et de I(t)) témoigne d'une continuité importante entre deux clusters.

Diachronic'Explorer ajoute ensuite des règles pour sélectionner les liens les plus importants[245]. Ces règles n'ayant pas été retenues dans ma thèse, elles ne sont pas ici détaillées. Le processus de sélection des liens les plus importants que j'ai implémenté repose

[245] Ces règles sont dans la fonction « getTargetClusterMatching » du fichier « LabelDiachronism.java » hébergé sur https://github.com/nicolasdugue/istex/blob/master/JAVA/CSR/src/model/diachronism/ (consulté le 18/09/2023).

sur la possibilité laissée à l'utilisateur au moment de la visualisation de choisir un seuil de manière interactive. Si l'utilisateur baisse ou augmente ce seuil de sélection des liens, il peut voir immédiatement le résultat généré. Ce choix a été initialement réalisé, car j'ai pu constater que la méthode de sélection de *Diachronic'Explorer* était très restrictive (*cf.* section Chap7.III.3.c). Il a pour conséquence de ne pas faire reposer la sélection des liens sur une technique statistique, mais de faire appel à l'analyse de l'utilisateur.

Ce qui a donc été gardé par rapport au modèle originel est uniquement le calcul des liens reformulé sans les dimensions probabilistes et bayésiennes. Tout le processus de sélection des liens a été abandonné ce qui simplifie nettement le modèle. Ce dernier peut à la suite de ces modifications être résumé ainsi : pour un cluster source d'une époque n et un cluster cible d'une époque n+1, deux indicateurs sont calculés. Le premier I(s) correspond à la proportion de la somme des poids dans le cluster source des termes en commun entre le cluster source et cible par rapport à la somme de l'ensemble des poids des termes du cluster source. Le deuxième I(t) correspond à la proportion de la somme des poids dans le cluster cible des termes en commun entre le cluster source et cible par rapport à la somme de l'ensemble des poids des termes du cluster cible.

À ces premières modifications, il faut en ajouter plusieurs autres pour la visualisation. Comme précédemment, il s'agit d'expliquer ce qui existe initialement pour comprendre les raisons des modifications apportées.

b. Changements dans la visualisation

La figure suivante présente une visualisation issue de *Diachronic'Explorer* :

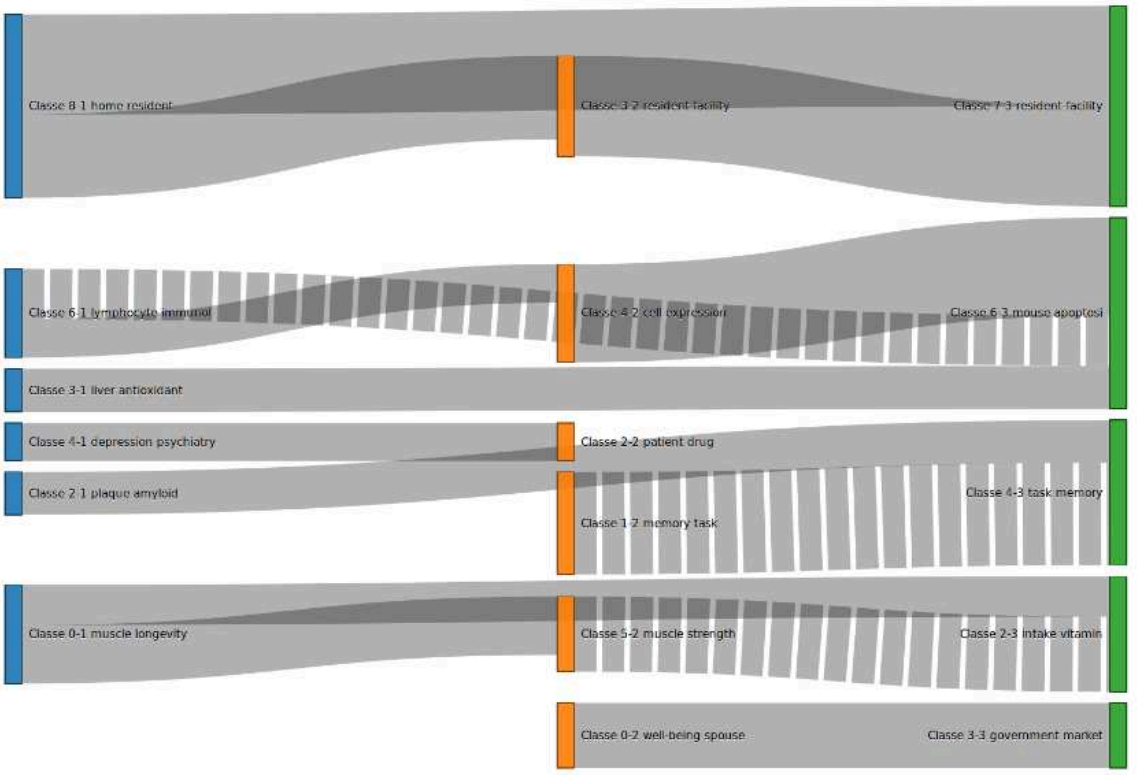

Figure n°19 : Exemple d'une représentation obtenue avec *Diachronic'Explorer.*

Ce graphique est un diagramme de Sankey. La figure est peu lisible. Son contenu détaillé n'a aucune importance pour cette thèse. L'objectif est seulement de faire comprendre comment une telle figure se lit et pourquoi certains choix présents initialement dans *Diachronic'Explorer* n'ont pas été retenus.

Chaque colonne (bleu, orange et vert) représente une période différente. Au sein d'une période, chaque rectangle représente un cluster. Il est inutile de détailler dans cette thèse comment ont été effectués en amont ces clusters et à quoi ils correspondent exactement. Deux clusters sont reliés entre eux en grisé[246] si des liens ont été établis avec la méthode précédemment détaillée (*cf.* section Chap7.III.3.a). Quand deux liens sont superposés, le grisé est plus foncé. Une grande partie de ces superpositions provient du fait que des liens ont été également cherchés dans ce cas pour des périodes non successives comme le montre la présence de liens entre la première et la troisième période. La largeur d'un lien dépend de sa force calculée comme une moyenne entre P(Act | s) et P(Act | t).

[246] Le fait qu'il existe des liens en grisé continu et en hachuré n'est pas ici détaillé, car il faudrait rentrer dans le processus de sélection des liens les plus importants réalisé par *Diachronic'Explorer* pour une compréhension complète. Cela aurait complexifié nettement la partie précédente sans présenter un grand intérêt pour cette thèse.

La première modification rapidement adoptée a été de ne pas reprendre cette idée de liens entre époques non successives. Elle complique à mon avis grandement la lecture des graphiques. Une étape importante a été ensuite marquée par la découverte de représentations similaires issues de la plateforme *CorText*[247] et jugées plus pertinentes. La force des liens n'est plus proportionnelle à leur taille. Cette force est plutôt marquée par des niveaux de gris. Un réglage dynamique permet d'afficher plus ou moins de liens en fonction de leur importance. Cette modification permet de faire dépendre la hauteur du rectangle qui représente un cluster d'une autre variable (le nombre de mots qui les compose ou le nombre d'occurrences de ces mots dans le corpus). Dans la figure suivante, nous présentons une de ces représentations dont la version dynamique est accessible avec le lien suivant : **@Diachro1**

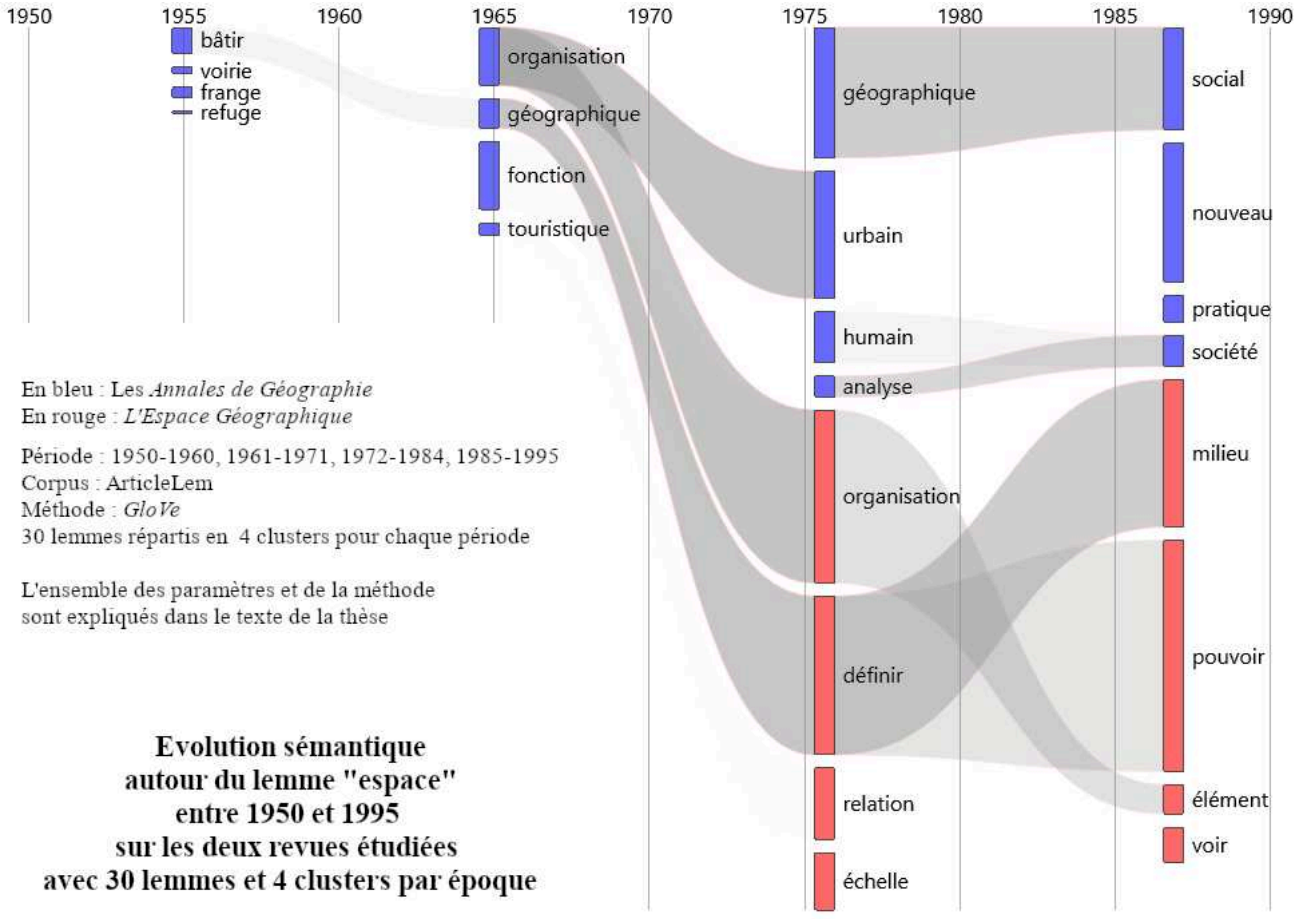

Figure n°20 : Exemple d'une représentation obtenue avec l'application créée.

[247] https://www.cortext.net/ (consulté le 18/09/2023).

Dans la version dynamique, il est possible de choisir d'afficher un seul ou deux labels[248] pour chaque cluster. L'option « Un seul label » a ici été privilégiée, car en choisissant deux labels, il existe des superpositions pour la période 1950-1960 (située en dessous de 1955). La version dynamique permet aussi de régler le seuil de sélection des liens. Sur la version ici présentée, le seuil est réglé à 0,3. Si une valeur nulle est choisie pour ce dernier, tous les liens apparaissent, mais il y a une saturation de la figure qui devient difficilement analysable. *A contrario*, si la valeur de 0,5 est privilégiée pour ce seuil, plus aucun lien n'apparaît. Pour cette raison, le seuil a été réglé à une valeur de 0,3 pour enlever les liens dont les moyennes de I(s) et de I(t) sont les plus faibles. Enfin, dans la version dynamique, il est possible en situant la souris sur chaque cluster de faire apparaître plusieurs informations concernant le cluster en question, notamment les termes composant le cluster et les auteurs ayant le plus utilisé ces termes dans la revue et la période donnée. En cliquant sur un cluster, il est possible d'accéder aux liens des documents dans le portail *Persée* qui présentent le maximum d'occurrences des termes constituant ce cluster (toujours dans la revue et sur la période concernée)[249]. Un lien vers le dendrogramme permettant de comprendre la partition effectuée (*cf.* section Chap7.III.2) est également accessible. Pour finir, en passant sur les liens entre deux clusters, il est possible de voir sa force (moyenne de I(s) et de I(t)), les termes partagés entre le cluster source et cible, les termes spécifiques du cluster source et les termes spécifiques du cluster cible.

Ces fonctionnalités et ces choix de visualisation étant détaillés, il s'agit maintenant de comprendre quels paramètres ont été utilisés pour construire cette figure.

c. Explication des paramètres

Pour réaliser la Figure n°20, le corpus « ArticleLem » a été utilisé. La méthode *GloVe* a été privilégiée avec les paramètres précédemment spécifiés[250.] Contrairement à la section précédente où seulement 10 lemmes avaient été retenus pour présenter des dendrogrammes simples (*cf.* section Chap7.III.2), le seuil de 30 lemmes a été choisi pour avoir une couverture sémantique plus large. Le nombre de groupes demandés a été augmenté à 4 en utilisant la stratégie du « saut maximum ». Par rapport à ces valeurs de 30 lemmes et de 4 groupes, il n'y a pas eu dans un premier temps de recherche d'optimisation.

---

[248] La problématique de l'attribution du label est traitée dans la section suivante sur les paramètres.

[249] Pour faire disparaître ces liens vers les documents et le dendrogramme, il suffit de cliquer une nouvelle fois sur le cluster en question.

[250] 15 pour le minimum d'occurrence, 100 pour le nombre d'itérations, 10 pour la taille du contexte et 100 pour la taille du plongement.

Une question s'est posée à propos du poids à attribuer à chaque terme au sein de chaque cluster. Au départ, la mesure de similarité-cosinus calculée par rapport au vecteur du terme choisi (ici le lemme « espace ») a été utilisée. Toutefois, ce choix peut être contesté. En effet, cette mesure de similarité-cosinus n'est pas représentative de l'importance du terme à l'intérieur de son cluster. Pour faire face à ce problème, une proposition a été implémentée en calculant pour chaque terme la somme des similarités-cosinus par rapport aux autres termes du groupe auquel il appartient. Le résultat maximum obtenu détermine le label du groupe. Par exemple, pour un cluster constitué des lemmes « vert », « quartier » et « urbain », trois sommes sont effectuées :

- Pour le lemme « vert », la somme des similarités-cosinus entre les vecteurs des lemmes « vert » et « quartier » et entre les vecteurs des lemmes « vert » et « urbain ».
- Pour le lemme « quartier », entre « quartier » et « vert » et entre « quartier » et « urbain ».
- Enfin, pour le lemme « urbain », entre « urbain » et « quartier » et entre « urbain » et « vert ».

Du fait d'une plus grande proximité sémantique entre « urbain » et « quartier » qu'entre « urbain » et « vert » (ou « quartier » et « vert »), les lemmes « urbain » et « quartier » renvoyant plus directement à une sémantique urbaine obtiennent des sommes des similarités-cosinus plus élevées. Cette méthode permet ainsi d'exclure « vert » et de sélectionner une étiquette représentant l'orientation sémantique globale du cluster (« urbain » ou « quartier »). Cette méthode fonctionne bien dans certains cas, mais les expérimentations menées montrent qu'elle peut aussi conduire à mettre en avant des lemmes ou termes très génériques sémantiquement. C'est par exemple le cas des adverbes ou des articles. En effet, ces derniers se retrouvent proches de beaucoup de termes et sont donc favorisés par cette méthode quand ils n'ont pas été écartés lors de la construction du corpus. Ceci explique que la création de ce paramètre « Calcul PoidsLabel »[251] a été accompagnée par la mise en place d'un autre paramètre appelé « stop mots ».

Ce paramètre permet d'enlever certains termes des résultats obtenus. Deux listes, une pour le corpus « ArticleLem »[252] et une pour le corpus « ArticleNonLem »[253], ont été

[251] Ce paramètre offre trois possibilités : « Similarité cosinus par rapport à terme initial et label max », « 1 pour tous les termes et maintien label max similarité cosinus » et « 1 pour tous les termes et label max somme similarité cosinus intra groupe » Dans cette thèse, l'option intitulée « 1 pour tous les termes et label max somme similarité cosinus intra groupe » a été privilégiée du fait des réflexions réalisées.

[252] Liste constituée par les lemmes : tant, autant, ainsi, aussi, quelques, parfois, surtout, toujours, encore, plus, moins, bien, peu, très, toute, plupart, certaines, alors, également, ici, notamment, enfin, particulièrement, seulement, pas.

[253] Liste constituée par les termes : autant, tant, ainsi, aussi, de, et, son, l, un, c, si, est, que, qu, tout, mais, même, donc, il, qui, dans, comme, elle, cet, sa, dont, celui-ci, celui, leur, (ainsi que les signes de ponctuation point, virgule et point virgule).

utilisées pour la génération de tous les résultats. Cette méthode permet de ne pas créer un filtre spécifique pour chaque résultat. Il s'agit d'éviter une sélection arrangeante des résultats grâce à un filtrage des « stop-mots » adapté à chaque expérimentation.

Au niveau des autres paramètres, une option existe dans l'application pour définir la taille des clusters avec deux modalités possibles : « Nombre de termes constitutifs du cluster » ou le « Nombre d'occurrences des termes constitutifs du cluster ». Dans la première modalité, si un cluster est constitué de 3 termes, la valeur de 3 est retenue comme base pour calculer la hauteur du rectangle représentatif de ce cluster proportionnellement aux valeurs retenues pour les autres clusters. Dans la seconde modalité, en reprenant l'exemple d'un cluster constitué par les lemmes « urbain », « quartier » et « vert », l'application créée sélectionne tout d'abord les textes où apparaît au moins une fois le terme originellement étudié (« espace » par exemple pour la Figure n°20) dans la revue et sur la période du cluster. Ensuite, des comptages dans les textes retenus du nombre d'occurrences pour chaque terme constituant le cluster sont effectués. Enfin, la somme de ces nombres d'occurrences est calculée et retenue comme résultat. C'est cette dernière option qui a été privilégiée dans cette thèse, car elle donne une idée approximative de l'importance de l'usage des termes constitutifs d'un cluster (en comparant la taille des différents clusters).

En ce qui concerne les liens entre clusters, un paramètre fait référence aux différences précédemment explicitées entre la méthodologie liée à la théorie initiale, celle correspondant à l'implémentation de Nicolas Dugué dans *Diachronic'Explorer* et celle retenue pour cette thèse (*cf.* section Chap7.III.3.a).

Enfin, il existe deux dernières options. La première est purement visuelle. Elle permet de choisir la couleur des clusters représentant les revues. Ici, bleu pour les *Annales de Géographie* et rouge pour *L'Espace Géographique*. La seconde appelée « CompareJustNew Revue » gère les époques comparées quand plusieurs revues sont présentes. Quand cette option est retenue, la comparaison entre les deux revues s'effectue seulement lors de l'apparition de la nouvelle revue. Par exemple, dans le cas de la Figure n°20, cette option a été retenue. L'algorithme n'a effectué comme comparaison inter-revues que celle entre les *Annales de Géographie* 1960-1971 et *L'Espace Géographique* 1972-1984. Si cette option n'avait pas été retenue, la comparaison inter-revues aurait concernée également les *Annales de Géographie* 1972-1984 et *L'Espace Géographique* 1985-1995 ainsi que *L'Espace Géographique* 1972-1984 et les *Annales de Géographie* 1985-1995. Par rapport à la

problématique étudiée, il m'a semblé que la comparaison inter-revues était surtout pertinente lors de l'apparition de *L'Espace Géographique*.

Tous ces paramètres sont présentés dans l'application en amont de chaque figure. Pour la Figure n°20, cette présentation est la suivante.

Figure n°21 : Ensemble des paramètres tels que présentés dans l'application ayant permis la construction de la Figure n°20.

Tous les paramètres utilisés ayant été explicités, la section suivante présente une analyse de deux résultats obtenus sur les termes les plus cruciaux par rapport à la problématique de cette thèse : « espace » et « milieu » (*cf.* section Chap2.II.3.c).

IV. Évolutions sémantiques d' « espace » et de « milieu »

1. Évolution sémantique d' « espace »

De premières analyses peuvent bien entendu être réalisées à partir de la Figure n°20. Afin de ne pas surinterpréter les résultats d'une seule figure, une seconde est présentée (*cf.* Figure n°22). La construction de cette seconde figure a fait l'objet d'une réflexion plus approfondie sur les paramètres utilisés. Les différences par rapport à la Figure n°20 sont les suivantes :

- Le corpus « ArticleNonLem » a été privilégié pour prendre en compte l'observation précédente[254] de dynamiques sémantiques différentes entre « espace » et « espaces ». Les mots éliminés (« stop words ») ont été adaptés pour prendre en compte la non-lemmatisation du corpus.

[254] *Cf.* section Chap7.II.3

- Pour les termes retenus, il m'a semblé intéressant de différencier leur nombre en fonction des revues et des époques. S'il semble pertinent de retenir les 50 premiers termes (dont les vecteurs sont les plus proches du vecteur du terme « espace ») pour *L'Espace Géographique* 1972-1984, la question se pose plutôt après 30 termes pour les *Annales de Géographie* 1950-1960. Cette observation empirique m'a conduit à implémenter la possibilité de définir en entrée un nombre de termes différents selon les époques et les revues. Les observations précédentes m'ont conduit à retenir 30 termes pour les *Annales de Géographie* 1950-1960, 50 termes pour les *Annales de Géographie* 1961-1971 et 60 termes pour toutes les autres périodes et revues.
- Enfin, il a paru logique d'adapter le nombre de clusters demandés en fonction du nombre de termes retenus. J'ai donc implémenté la possibilité de définir en entrée un nombre de clusters demandés différents selon les époques et les revues. Proportionnellement aux nombres de termes retenus précédemment, j'ai demandé 3 clusters pour les *Annales de Géographie* 1950-1960, 5 pour les *Annales de Géographie* 1961-1971 et 6 pour toutes les autres périodes et revues.

La représentation obtenue est alors la suivante.

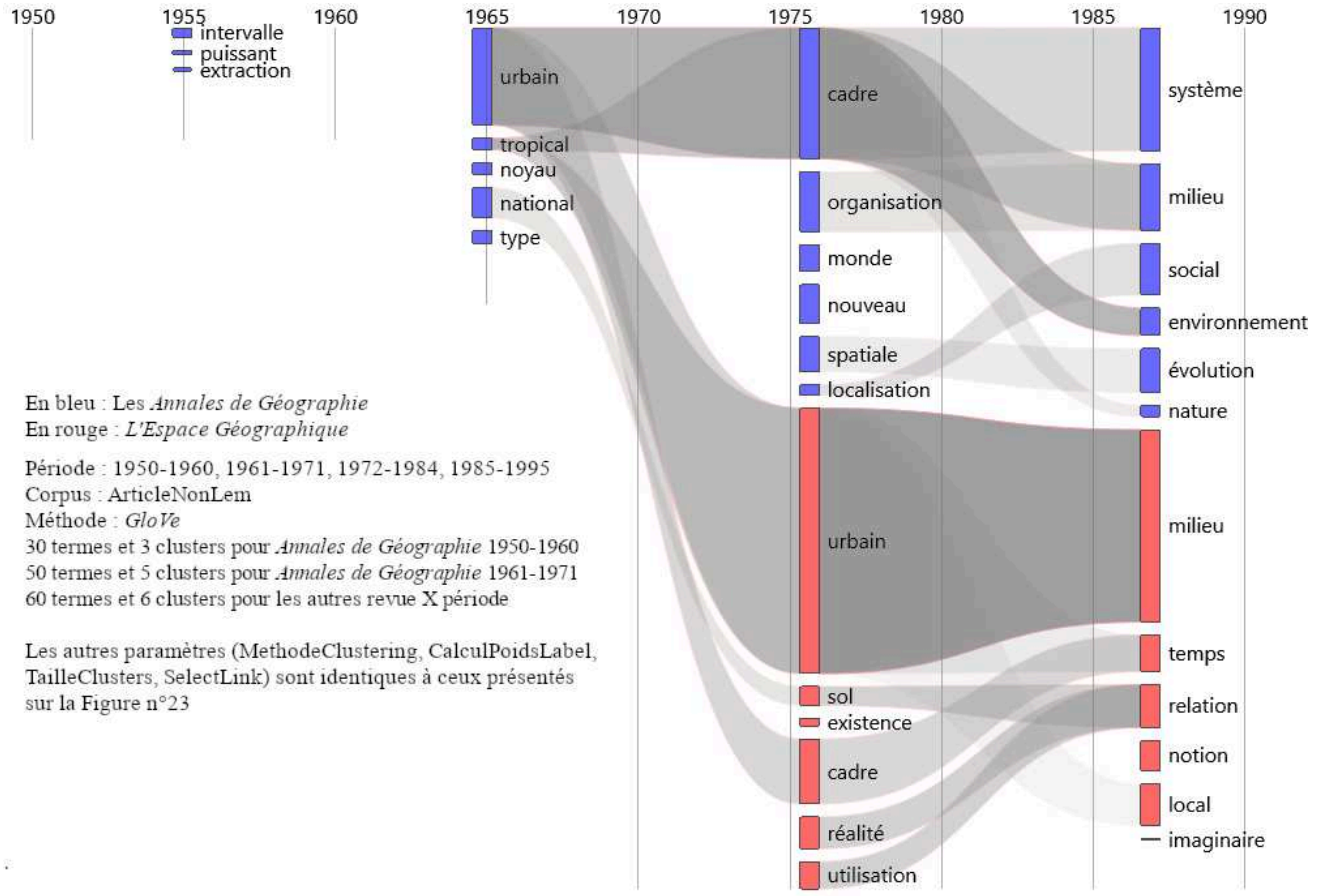

Figure n°22 : Évolution sémantique autour du terme « espace » entre 1950 et 1995 sur le corpus ArticleNonLem avec des nombres différenciés de termes et de clusters.

La version dynamique est accessible avec le lien suivant : **@Diachro2**

Le constat précédemment réalisé (*cf.* section Chap7.II.1) d'une discontinuité sémantique autour de ce lemme « espace » dans la revue les *Annales de Géographie* entre 1950-1960 et 1961-1971 se retrouve sur cette figure. Une transition entre les *Annales de Géographie* 1961-1971 et *L'Espace Géographique* 1972-1984 est marquée par le lien entre deux clusters partageant la même étiquette « urbain ». Les termes partagés entre ces deux clusters, en plus d' « urbain », sont : « organisation », « paysage », « milieu », « géographique ». Toujours sur ces deux clusters, les termes non partagés sont :

- Du côté des *Annales de Géographie* 1961-1971 : « économique », « aménagement » et « caractère ».
- Du côté de *L'Espace Géographique* 1972-1984 : « système », « environnement », « social », « vécu » et « objet ».

- « cadre » et « régional » apparaissent aussi comme non partagé à l'échelle de ces deux clusters, mais ils sont partagés par l'intermédiaire du cluster ayant pour étiquette « cadre » sur *L'Espace Géographique* 1972-1984.

Ces observations illustrent l'hypothèse d'une continuité entre les dynamiques scientifiques des années 1960 précédemment mentionnées[255] (*cf.* section Chap7.II.5) et le mouvement de l'analyse spatiale promu par *L'Espace Géographique*. Si de nouvelles dynamiques sont effectivement visibles, elles ne forment pas une discontinuité majeure de type « changement de paradigme » sur cette figure. Le cluster marqué par l'étiquette « réalité » peut être bien entendu mis en rapport avec la thèse développée par Olivier Orain, mais des interprétations opposées sont réalisables : d'un côté, il est possible d'insister sur la présence de ce cluster et sur sa discontinuité avec l'époque précédente ; de l'autre, l'émergence réduite du mot « réalité » ne signifie nullement une sortie du réalisme et peut même être vue et revendiquée comme étant dans une continuité de perspective.

Enfin, le cluster « milieu » de *L'Espace Géographique* 1985-1995 pose évidemment question par rapport à l'hypothèse sémantique d'Olivier Orain, d'autant plus que ce cluster se retrouve également dans la Figure n°20 sur la même revue et époque. Il me semble ici important de mentionner que j'ai obtenu par ailleurs de nombreuses autres représentations où cette étiquette de « milieu » n'apparait pas directement sur *L'Espace Géographique* 1985-1995. Par exemple, la représentation accessible via ce lien **@Diachro3** qui privilégie le corpus « ArticleLem », mais qui a été construit par rapport à la Figure n°20 en différenciant le nombre de lemmes et de clusters (comme sur la Figure n°22).

[255] Ces dynamiques sont le renouvellement des études régionales, l'essor de l'économie spatiale et le développement des réflexions aménagementistes.

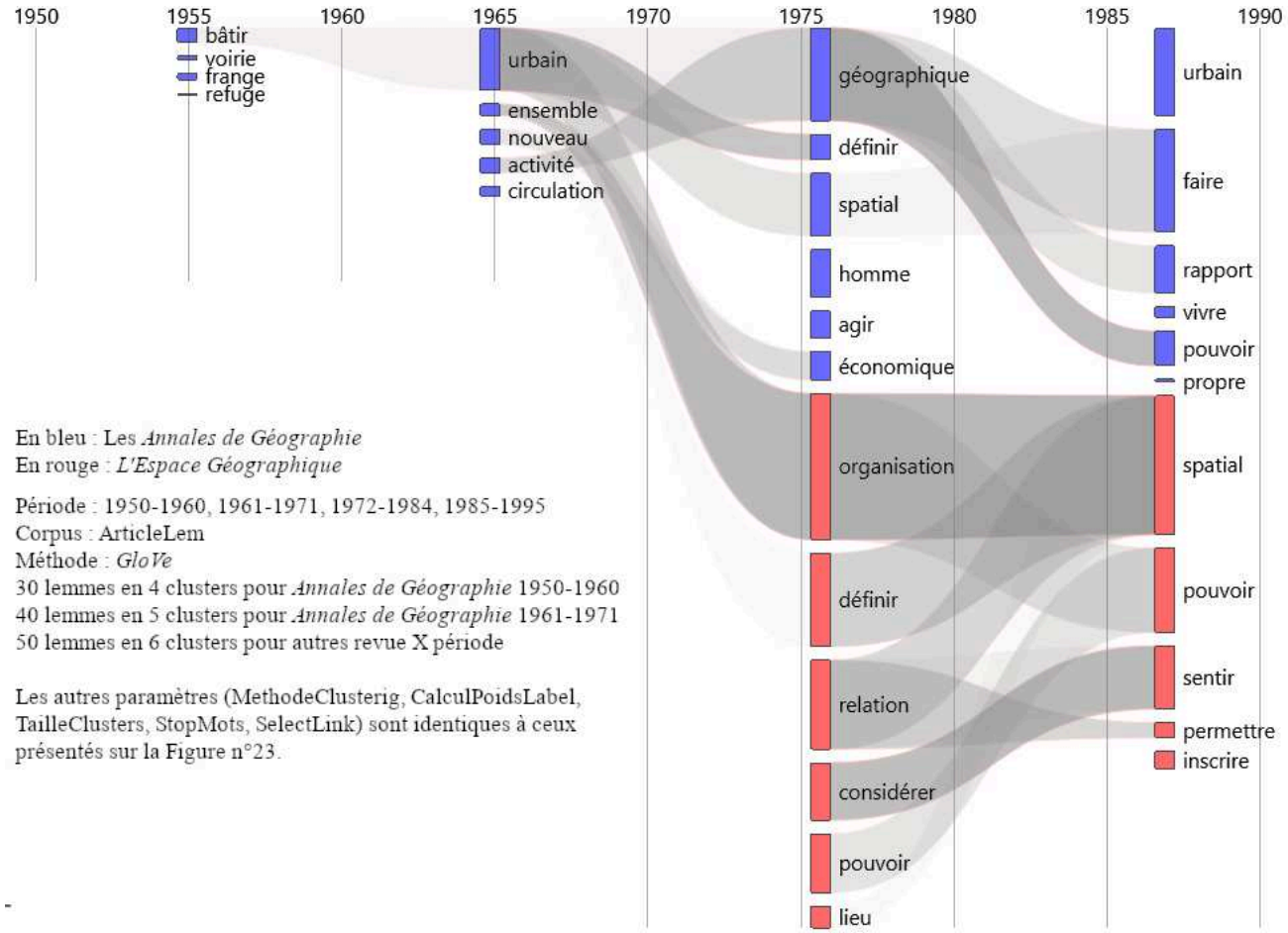

Figure n°23 : Évolution sémantique autour du terme « espace » entre 1950 et 1995 sur le corpus ArticleLem avec des nombres différenciés de termes et de clusters.

Si « milieu » n'apparait pas directement comme étiquette de cluster, il faut souligner que ce terme fait tout de même partie du cluster « spatial » de *L'Espace Géographique* 1985-1995. Il reste par conséquent proche sémantiquement du lemme « espace », ce qui interroge profondément l'hypothèse sémantique d'Olivier Orain qui repose sur une différenciation forte de ces deux termes.

L'ensemble des réflexions précédentes sur la construction du corpus, le choix de la méthode, la réalisation des prétraitements, l'établissement des paramètres et la possibilité des multiples visualisations me conduisent à être prudent et à considérer ces résultats, non pas comme une preuve de remise en cause, mais bien plutôt comme un élément questionnant fortement l'hypothèse socio-sémantique d'Olivier Orain. La reconnaissance de cette fragilité démonstrative interroge aussi directement la notion de paradigme. Si toute recherche en SHS dépend ainsi du corpus, de la méthode et de multiples choix, n'y a-t-il pas là une explication au constat fait par Thomas Kuhn d'une instabilité de ces sciences, le conduisant à forger la

notion de paradigme pour mieux comprendre par opposition la dynamique des sciences dites dures ?

Avant de creuser cette piste, comme la signification de « milieu » a également varié dans le temps, il est pertinent de l'étudier avec la méthodologie mise en place**.**

2. Évolution sémantique de « milieu »

Il n'existe pas de différence notable entre les dynamiques sémantiques de « milieu » et « milieux » telle que celle observée pour « espace » et « espaces » (*cf.* section Chap7.II.3). Cette observation justifie l'utilisation d'un corpus lemmatisé. Le tableau suivant présente le nombre d'occurrences du lemme « milieu » sur le temps long dans les deux revues étudiées.

	1890 -1910	**1911 -1931**	**1932 -1952**	**1961 -1971**	**1972 -1985**	**1986-2000**
Annales de Géographie	814	922	761	1104	1321	1619
L'Espace Géographique					1795	1479

Tableau n°26 : Nombre d'occurrences du lemme « milieu » dans le corpus ArticleLem.

En m'appuyant sur des appréciations qualitatives, le nombre de termes retenus a été fixé à 40 et le nombre de groupes à 4 pour les *Annales de Géographie* 1890-1910, 1911-1931, 1932-1952. Pour les autres revues et périodes, le nombre de termes retenus est de 50 et le nombre de groupes est de 5.

Concernant les autres paramètres, des choix identiques à ceux précédemment réalisés[256] ont été privilégiés. Le lien suivant permet d'accéder au résultat obtenu : **@Diachro4**. Le seuil de 0,25 pour la sélection des liens et le choix d'une seule étiquette pour chaque cluster ont été de nouveau retenus pour présenter la figure suivante.

[256] 15 pour le minimum d'occurrence, 100 pour le nombre d'itérations, 10 pour la taille du contexte et 100 pour la taille du plongement ; Saut maximal pour la méthode de clustering, ;1 pour tous les termes et label max somme similarité cosinus intra groupe ; Nombre d'occurrences des termes constitutifs du cluster.

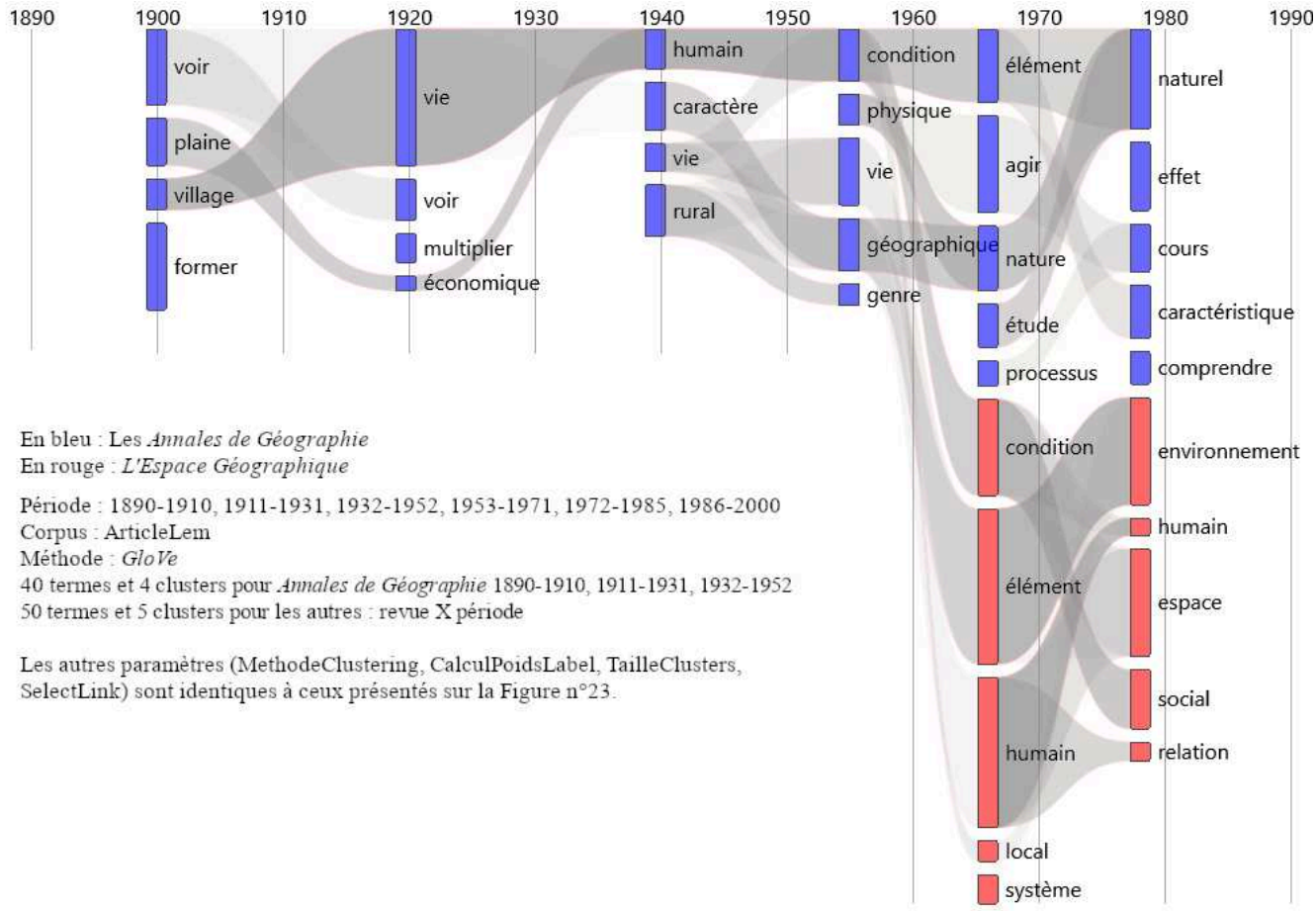

Figure n°24 : Évolution sémantique autour du lemme "milieu" entre 1890 et 2000 sur le corpus ArticleLem avec des nombres adaptés de termes et de clusters.

Cette figure permet d'appréhender à la fois des évolutions sémantiques du terme « milieu » et plusieurs continuités entre époques successives. Par rapport à la thèse d'Olivier Orain, il faut tout d'abord souligner que cette représentation ne permet pas de mettre en évidence une rupture sémantique autour du terme « milieu » entre la période vidalienne (1890-1910) et la suivante. Il y a plutôt une continuité entre plusieurs clusters. Un élément intéressant qui est mobilisable dans le sens de la thèse d'Olivier Orain est cette présence des deux clusters avec les étiquettes « voir » dans les *Annales de Géographie* 1890-1910 et 1911-1931. Avant de surinterpréter le fait que ce verbe disparaisse sur la période suivante, il faudrait effectuer des recherches plus approfondies dans les textes de ces époques pour mieux comprendre cette observation. Cela n'a pas été réalisé, mais l'idée que cela puisse révéler une évolution entre la première génération de vidaliens et les suivantes, est une hypothèse plausible.

Un élément à première vue intriguant est le fait que le cluster avec l'étiquette « vie » des *Annales de Géographie* 1911-1931 ne soit pas relié avec celui ayant la même étiquette sur

les *Annales de Géographie* 1932-1952. En rentrant dans le détail de ces deux clusters, cela s'explique facilement. Ces deux clusters ne partagent que ce lemme, ce qui aboutit à un score de lien peu élevé. En 1911-1931, les lemmes dans les *Annales de Géographie* reliés à ce cluster « vie » sont très généraux : « siècle », « plaine », « sol », « village », « condition », « homme »… Le vocabulaire relié à ce cluster « vie » change dans cette revue en 1932-1952 avec les lemmes : « élément », « véritable », « existence », « complexe », « combinaison », « adapter » et « vivant ». Le premier auteur[257] de ce cluster est Maximilien Sorre (1880-1962). Le fait que ce cluster se modifie ensuite et ne se poursuive pas dans le temps, correspond bien à la situation historique d'un auteur dont l'œuvre n'a pas été directement reprise et poursuivie.

À partir de 1972, des changements sont observables avec la montée des lemmes « élément », « système » et « environnement ». Globalement, l'idée d'une assiette sémantique différente d'« espace », plus proche pour « milieu » de l'étude des relations hommes / nature, est facilement repérable. Toutefois, l'hypothèse d'une dissociation nette des deux dynamiques sémantiques d'« espace » et de « milieu » à partir du milieu des années 1980, n'a été validée par aucun résultat obtenu lors de toutes mes expérimentations.

Si ces résultats interrogent sur les phénomènes de continuités/discontinuités sémantiques et scientifiques, ils ne permettent pas de répondre à la question de la pertinence d'une lecture kuhnienne. Une discussion élargie fait advenir et valoir des éléments complémentaires et plus décisifs de mon point de vue.

[257] Dans le sens sens où cet auteur comptabilise un maximum d'occurrences des lemmes de ce cluster dans les documents de cette revue sur cette période.

Chapitre 8 :
Discussion élargie à partir des résultats et du processus de leurs productions

Mon travail doctoral m'a fait expérimenter le passage d'un constructivisme faible, surtout méthodologique (avec la formulation d'une problématique, la construction d'un corpus de données, le choix d'une méthode et la production de résultats) à un constructivisme un peu plus avancé, ayant notamment une dimension ontologique (avec une reconnaissance que les résultats ne sont pas univoques, qu'ils comportent des incertitudes et qu'ils dépendent de problématiques, de corpus et de grilles de lecture appliquées). Mon processus de recherche a été marqué plus spécifiquement par une évolution importante quand j'ai compris et accepté que mes résultats quantitatifs ne seraient pas directement conclusifs. Cette reconnaissance et ce choix m'ont permis de rendre compte de manière plus détaillée des difficultés rencontrées. En effet, l'enjeu autour de la partie quantitative est devenu en quelque sorte moindre à partir du moment où je l'ai complétée avec une approche plus qualitative (*cf.* Part3). La reconnaissance d'une fragilité des résultats quantitatifs ne remettait plus en cause l'ensemble de la thèse.

Même avec cette trajectoire scientifique (qui diminue fortement l'attente et la pression sur les résultats quantitatifs obtenus), la présentation des difficultés rencontrées et des points de fragilité a continué de relever de décisions réfléchies de formulation. En effet, il existe toujours une tension avec l'objectif légitime de ne pas dévaloriser les résultats produits. Par conséquent, si le passage d'un constructivisme méthodologique à un constructivisme plus ontologique peut être mis en avant dans la dynamique de ma thèse, il est également nécessaire de prendre en compte la résistance forte d'un pôle que je qualifie volontiers de « réaliste ». En effet, cette tension s'explique par la volonté de montrer que les résultats ne sont pas de purs artifices. Il ne s'agit pas, bien entendu, d'affirmer que les résultats obtenus représentent ce qui est, de manière complète et fidèle. Toutefois, l'objectif reste tout de même de réaliser une construction qui conduit à penser que ces résultats sont au minimum vraisemblables, et au mieux qu'ils comportent un fort niveau de véridicité. Sur ce plan, ma recherche ne me semble pas être une exception par rapport à l'ensemble de la géographie théorico-quantitativiste qui est à mon sens marquée par cette recherche de résultats véridiques et ce pôle « réaliste ».

Cette tension rentrant en contradiction avec l'idée d'un paradigme réaliste révolu, l'objectif est dans un premier temps de situer la discussion à l'endroit théorique où Olivier Orain (2003) dans sa thèse place la « révolution scientifique » (Kuhn 2008, *1962*), c'est-à-dire sur le cœur de chauffe entre réalisme et constructivisme-nominalisme.

I. Sur le cœur de chauffe : entre réalisme et constructivisme-nominalisme

Quand Ian Hacking (2008) développe ce qu'il y a derrière le pôle cognitif qu'Olivier Orain qualifie de « réaliste », il préfère le nommer « structurisme-inhérent » (*Ibid* 2008, 119). Il le décrit ainsi : « des scientifiques croient que le monde vient à l'existence avec une structure inhérente qu'il est de leur tâche de découvrir » (*Ibid* 2008, 118). Il s'agit là d'une définition ontologique du réalisme à laquelle la « nouvelle géographie » est originellement rattachée puisqu'elle a été initialement marquée par l'importance du structuralisme[258].

Il existe donc sur ce critère plutôt une continuité avec la géographie post-vidalienne qui est problématique vis-à-vis de la lecture développée par Olivier Orain. Face à cette remarque, il est possible de faire valoir que les formes de réalisme de ces deux géographies sont différentes. De plus, il est bien entendu possible de rappeler qu'Olivier Orain reconnaît que les principaux théoriciens de la « nouvelle géographie » (Roger Brunet, Philippe Pinchemel et Henri Reymond) sont des réalistes (Orain 2003, 292). Enfin, il est également vrai que la thèse du plain-pied n'affirme pas complètement l'existence d'un nouveau paradigme. Toutes ces remarques ne me semblent toutefois pas lever l'ambiguïté ici soulignée : le développement d'une lecture kuhnienne, alors que la principale caractéristique du pré-supposé paradigme − le réalisme − n'est pas vraiment révolue et que, par conséquent, l'existence de deux mondes incommensurables est très discutable.

Dans sa réponse à Isabelle Lefort et Caroline Leininger-Frézal (2011), Olivier Orain reconnaît l'importance du maintien d'un pôle « réaliste » dans la géographie française et ajoute : « Il me semble que les géographes contemporains ne démarquent pas encore suffisamment la question du « réel » de celle de la construction de l'objectivité, question connexe plus fréquente chez les sociologues. » (Orain 2011, 241). S'il est possible en effet de reconnaître facilement qu'Émile Durkheim était bien plus avancé que Vidal de la Blache sur la question de la construction d'une problématique avec des données et des traitements statistiques, la réalisation de cette thèse me conduit à penser que sous la construction de l'objectivité, la question du réalisme n'est jamais bien loin.

Les analyses de Gilles-Gaston Granger dans *Science et réalité* (2001) permettent de développer cette réflexion. Cet auteur montre qu'il faut plutôt parler de réalités scientifiques au pluriel, car, suivant les domaines, les critères de validité des connaissances sont variables.

---

[258] La proposition la plus célèbre est celle des chorèmes comme structures élémentaires de l'espace géographique. Le terme est construit en 1980 par Roger Brunet par analogie avec les phonèmes de la linguistique structurale. Sur un autre plan, l'emploi du prisme kuhnien pour lire l'histoire de la géographie s'inscrit aussi dans cette importance de la perspective structuraliste (*cf.* section Chap2.II.1.a).

Ceci conduit à distinguer le concept de réel (dont Gilles Gaston Granger montre une partie des évolutions dans l'histoire de la philosophie et fait remonter sa conception moderne à Leibniz) de celui de réalité. Un des caractères du réel est son « unicité » et son « autonomie » (Granger 2001, 42). Cette affirmation se retrouve également chez Clément Rosset : *Le réel, Traité de l'idiotie* (2012). Les réalités ne sont que des saisies particulières, toujours partielles de ce réel. Les réalités scientifiques ne sont alors que des types de réalité parmi d'autres.

Cette distinction s'inscrit dans une orientation qui peut être qualifiée de constructiviste puisque ce n'est plus le réel qui est visé, mais des réalités, des constructions parmi d'autres. Toutefois, l'objectif global de la science reste dans cette conception de viser des réalités. Il s'agit d'obtenir des représentations qui font sens, mais aussi qui proposent, ou *a minima* qui laissent penser, à des adéquations avec un ou des éléments du monde. Cette conception conduit Gilles-Gaston Granger à plaider « pour un réalisme bien tempéré » (Granger 2001) dans la conclusion de son ouvrage. Cette expression permet de penser les changements mis en avant par Olivier Orain dans sa thèse plutôt sous la forme d'une « tempérance » que d'une révolution. Cela étant dit, il n'est pas question d'affirmer ici que le réalisme de la géographie post-vidalienne est le même que celui de la « nouvelle géographie ». Au-delà de l'indéniable diversité du pôle réaliste suivant les auteurs, il faut reconnaitre qu'il est possible de faire des rapprochements et d'analyser des évolutions. Toutefois, l'hypothèse que je défends ici, est que la lecture en termes de changement de paradigme n'est pas efficiente, car elle conduit de manière erronée à imaginer une fin et un remplacement du réalisme.

La conclusion d'Olivier Orain repousse habilement la rupture à plus tard :

> « Demeure un invariant : quand une volonté d'articulation théorie/empirie est manifestée, le résultat est trop souvent incertain, préliminaire, essentiellement programmatique. Cela signifie peut-être que le constructivisme géographique reste pour partie à imaginer ? » (Orain 2009, 375).

Après avoir beaucoup approfondi ce sujet, Jean-Claude Passeron dans *Le raisonnement sociologique* (2006, *1991*) montre bien comment la grille de lecture kuhnienne projetée sur les sciences sociales tend à les lire au travers de modes de scientificités qui ne sont pas les leurs. Cette perspective conduit alors souvent à penser les SHS par l'intermédiaire du motif récurrent d'une science encore imparfaite, mais qui va advenir. Olivier Orain ne peut pas être suspecté de projeter le mode de scientificité des sciences dites dures sur l'ensemble de la géographie française, mais il est particulièrement intéressant de remarquer que son travail fin d'adaptation du modèle kuhnien en vient à repousser la discontinuité à plus tard et à une formulation semblable à ce motif de la science à venir.

La fixation d'un point de rupture si éloigné pose dès lors un problème important. Pour prendre une analogie imparfaite, c'est comme si les continuités et discontinuités de l'histoire

informatique du XX^{ème} siècle étaient jugées à l'aune du développement généralisé de l'ordinateur quantique. Il n'est pas alors étonnant que ce qui est analysé avant semble continu. Cette analogie permet de cerner une partie importante du problème, mais elle est à mon avis très trompeuse, car elle entretient l'illusion que la révolution scientifique est à venir. Il me semble important ici de développer les réflexions de Jean-Claude Passeron pour bien saisir cette illusion tout en répondant à cette question quelque peu embarrassante : pourquoi plus d'un siècle après l'institutionnalisation de la géographie et plus de 40 ans après l'avènement de la « nouvelle géographie », ce constat de « résultats souvent incertains, préliminaires, essentiellement programmatiques » (Orain 2009, 375) demeure de la part d'un penseur averti de cette discipline ?

II. Jean-Claude Passeron : « un espace non poppérien de l'argumentation »[259]

Pour Jean-Claude Passeron, les constructions théoriques dans les SHS qui tentent de monter en généralité doivent toujours composer à partir d'empiries spécifiques et hétérogènes. Il est tentant pour un chercheur qui a essayé de construire ses propres définitions et l'ensemble d'un système conceptuel de croire « toucher à la terre promise d'une langue protocolarisée qui aurait enfin les vertus d'un paradigme durable » (Passeron 2006, 564). Toutefois, les SHS reposent avant tout sur l'élasticité du « langage naturel » qui leur est indispensable pour rapprocher des situations empiriques hétérogènes et créer des intelligibilités du réel. Dans ces disciplines, « ce ne sont pas les liens logiques qui nouent l'essentiel de la connaissance, mais les liens typologiques » (Passeron 2006, 589) qui rapprochent des cas de manière toujours imparfaite.

Il existe pour cet auteur des intelligibilités multiples et dans beaucoup de cas, à la différence des sciences expérimentales, il n'y a pas de réfutabilité possible au sens poppérien du terme. Sur ce sujet, les analyses par exemple de Bernard Lepetit (1993) sur le modèle de « l'économie-monde » montrent l'impossibilité d'une telle réfutation. Cette situation provient de plusieurs raisons. Tout d'abord, à la différence des sciences expérimentales, la formule qui consiste à raisonner « toute chose égale par ailleurs » est encore plus difficilement applicable dans les SHS. Ensuite, à la différence des mathématiques, il existe très rarement des formalisations complètes reposant sur la seule logique.

[259] Ce titre reprend la formulation du titre secondaire de l'ouvrage de Jean-Claude Passeron (2006) : *Le raisonnement sociologique : un espace non-poppérien de l'argumentation*. Par rapport à la première édition, il s'agit d'une nouvelle formulation, car initialement, l'ouvrage avait été intitulé : *Le raisonnement sociologique : un espace non-poppérien du raisonnement naturel*. Cette évolution est notable, car elle donne moins d'importance à la notion discutable de « raisonnement naturel ».

Après avoir rappelé qu'une démonstration ne tient que par la valeur de son maillon le plus faible, il est facilement compréhensible que la présence d'une partie formalisée et/ou quantitative (aussi intéressante soit-elle) ne suffise que très rarement pour juger de la pertinence globale d'un travail de SHS. Quand une méthode quantitative est mobilisée, il faut rappeler qu'en amont, il existe toujours une transformation discutable du monde en données. En aval, dans l'interprétation des résultats, il existe aussi toujours des processus de retraduction dans un langage non formalisé. Un raisonnement statistique n'énonce rien en lui-même comme le souligne Jean-Claude Passeron. Dès qu'il est traduit en mots, du sens et du contexte sont ajoutés.

Dans cette perspective, toute tentative d'axiomatisation ou de chorématisation[260] du monde, par son caractère universel (c'est-à-dire dénué de nécessité de contextualisation), est condamnée à être remise en cause. Il est possible d'objecter que cette situation est finalement identique à celle des sciences « dures » puisque, selon le modèle kuhnien, tout paradigme est appelé à être remplacé par un autre. La différence essentielle réside dans l'espace non poppérien qui caractérise les SHS pour Jean-Claude Passeron. C'est bien plus la véridicité d'un énoncé étayé par des exemples qui est atteint plutôt que la vérité d'une proposition. Certes, la loi de la chute des corps, considérée un moment comme universelle, se trouve après un changement de paradigme n'être qu'un cas local non généralisable. Toutefois, elle est remplacée par une loi, celle de la relativité générale, jugée plus universelle. *A contrario*, les SHS ne reposent pas fondamentalement sur cette dynamique de connaissances caractérisée par l'énonciation de lois jugées de plus en plus universelles.

Leurs modes de scientificités sont situés pour Jean-Claude Passeron entre deux pôles. D'un côté, le pôle historique et herméneutique qui renvoie à l'objectif de compréhension fine de chaque empirie spécifique. De l'autre, la recherche de régularités, de types, de structures transverses à plusieurs empiries, renvoyant à un pôle plus expérimental et nomologique. Aucun de ces deux pôles ne peut être suffisant à lui seul. Par exemple, l'apport constitué uniquement par des monographies (des « idiographies »[261]) est jugé comme minime par Jean-Claude Passeron. Les rapprochements de cas qui apportent de nouvelles compréhensions[262] ont bien plus de valeur. Symétriquement, une théorie qui ne ferait pas l'objet de mises à l'épreuve empirique, aurait également un intérêt minime. Il est bien entendu alors nécessaire de considérer les résultats acquis dans une perspective critique,

[260] La chorématisation renvoie à la proposition de Roger Brunet des chorèmes, décalque des phonèmes de la linguistique saussurienne. Les chorèmes sont ainsi pensés comme des structures élémentaires de l'espace géographique qu'il suffit de combiner pour décrire et expliquer toute organisation spatiale.

[261] C'est bien entendu l'unicité de chaque cas, l'« idiotie du réel » (Rosset 2012), qui est en toile de fond dans cette expression.

[262] Les compréhensions sont nouvelles si elles n'existaient pas avant que les cas utilisés soient rapprochés.

notamment en prenant en compte toutes leurs limites intrinsèquement liées à leurs productions.

Ainsi, l'affirmation précédente d'Olivier Orain, à savoir le constat que les résultats sont « souvent incertains, préliminaires, essentiellement programmatiques dès qu'il y a une volonté d'articulation théorie/empirie », se comprend aisément. En effet, la composition d'une théorie à partir d'ancrages empiriques hétérogènes est toujours une opération complexe. De plus, la reprise d'une théorie déjà construite pour l'appliquer à de nouveaux éléments est rarement simple du fait des différences entre les situations empiriques originelles (celles qui ont servi à élaborer la théorie) et celles nouvellement étudiées. Il est certes possible d'objecter qu'une théorie peut être conçue en dehors de toute référence à un ou plusieurs terrains. Dans la pratique, si de telles théories « hors sol » existent, il est facilement compréhensible qu'elles rencontrent par la suite également des difficultés dans leurs mises en correspondance avec des situations empiriques spécifiques et hétérogènes.

Cette conception conduit Jean-Claude Passeron à développer une forte opposition à toute lecture paradigmatique des SHS du fait de leurs modes de scientificité qui diffèrent en partie de ceux des sciences expérimentales et formelles. Il ne s'agit pas de nier les expérimentations et les modélisations dans les SHS, mais leurs espaces sont *in fine* non poppériens car basés en dernière instance sur l'importance du langage naturel avec des agencements argumentatifs pouvant toujours être discutés (relevant plus de la typologie que de la logique). Il arrive bien entendu que dans un domaine donné, à une époque donnée, un type d'approche domine, mais ce phénomène est plus conjoncturel que structurel. Du fait de cette dimension non poppérienne, il existe souvent des réactivations intellectuelles d'héritages anciens sous des formes renouvelées. De telles réactivations n'ont pas lieu dans les changements de paradigme tels que Thomas Kuhn les a conçus et présentés. En effet, dans son schéma, les paradigmes révolus sont abandonnés.

Il est nécessaire de préciser ici que Jean-Claude Passeron, en tant que sociologue, développe surtout son argumentaire autour de ce qu'il appelle le « raisonnement sociologique ». C'est par conséquent un transfert de ses réflexions qui est ici effectué au-delà de sa discipline d'origine. Pour la géographie, le transfert peut être directement justifié par le fait que les empiries spécifiques et hétérogènes qui sont à la base de l'élaboration théorique de Jean-Claude Passeron se caractérisent, entre autres, par des références spatio-temporelles. Cette référence au spatial permet de faire un lien évident avec la géographie.

En continuant ce transfert disciplinaire, il est certain que la « nouvelle géographie » a promu et développé le pôle expérimental et nomologique, mais elle n'a pas aboli le pôle historique et herméneutique. Les avancées du pôle expérimental n'ont pas mis fin à

l'utilisation du langage naturel. La mise en avant de constructions logiques n'a pas fait rentrer la géographie française dans un espace poppérien. Il est par conséquent ici pertinent et nécessaire de briser l'illusion du motif d'une science à venir[263]. Il semble bien plus adéquat de prendre acte que les modes de scientificités de la géographie ne peuvent pas et n'ont pas, à se conformer, ou même à devoir ressembler, à ceux des sciences dites dures.

Comme les réflexions de Jean-Claude Passeron ont été publiées initialement en 1991, une question légitime est celle de leur prise en charge par la thèse d'Olivier Orain. La section suivante développe l'idée d'une esquive stratégique.

III. Une esquive stratégique du « plain-pied »

Jean Claude Passeron n'est pas cité une seule fois dans la thèse d'Olivier Orain : ni dans le texte, ni dans la bibliographie. Il est utile ici de présenter un autre texte présenté par Olivier Orain pour le colloque *Géopoin*t 2004, c'est-à-dire un an après sa soutenance de thèse. Il le présente dans son blog :

> « Sur un autre plan, le thème proposé pour le colloque Géopoint 2004, *la forme en géographie*, m'a suggéré un réemploi de certaines réflexions que je n'avais pas voulu développer dans ma thèse. J'en ai tiré une contribution qui essaie de sortir des schèmes épistémologiques traditionnels (mais aussi du cadre discontinuiste de mes recherches antérieures) et de montrer, dans une conceptualisation proche de celle de Jean-Claude Passeron, que l'une des opérations les plus usitées en géographie serait la *clinique*, opération visant à identifier des cas en combinant des références universelles, une attention à la particularité et une indispensable (mais souvent peu consciente) interprétation » (Orain 2007b).

Par rapport à cette présentation, il m'importe de souligner, tout d'abord, que l'approche de Jean-Claude Passeron met bien en valeur une approche par cas, mais que la mention « en combinant des références universelles » est impropre[264]. Cette précision est importante, car elle explique une distance des réflexions de Jean-Claude Passeron par rapport aux postulats structuralistes de la « nouvelle géographie », beaucoup plus grande que ce que cette présentation d'Olivier Orain le laisse penser. Une lecture attentive de Jean-Claude Passeron me conduit plutôt à penser qu'il se rapproche d'un positionnement proche des sémantiques différentielles, notamment quand il affirme : « Or, le " contexte " pertinent d'une assertion de science sociale ne peut jamais être défini *a priori* : c'est l'ensemble potentiellement infini de ses pertinences explicatives que ne peut épuiser aucune liste finie de propriétés

[263] Ce motif a été entretenu par plusieurs géographes post-vidaliens comme André Cholley, mais aussi par les « nouveaux » géographes (Ferrier, Racine et Raffestin 1978).

[264] « Les "faits" sociaux élémentaires ne peuvent être définis univoquement à partir de "monèmes" comme on peut le faire pour les phonèmes d'une langue naturelle par la "commutation" » (Passeron 2006, 573).

susceptibles de " descriptions entièrement définies " » (Passeron 2020). Derrière la complexité de la formulation, une proximité forte peut être établie avec la « non-clôture du sens » mis en avant par François Rastier (2001). Faire le saut d'une approche structuraliste à un tel positionnement n'a rien d'évident. Cette tension est ici occultée par Olivier Orain.

Cela étant précisé, cette citation de blog montre explicitement qu'Olivier Orain connaissait les réflexions de Jean-Claude Passeron et les a délibérément non mentionnées lors de sa rédaction de thèse. Cette esquive ne veut pas dire qu'il ne les a pas prises en compte. Une utilisation indirecte est à mon avis formulée à la page 137 de sa thèse :

> « Alors que le couple expérimentation/prédiction fonde en théorie la légitimation des sciences naturelles, il n'y a que très peu de champs des sciences sociales (et « humaines » en général, fussent-elles quantitativistes) qui se prêtent complètement à ce type de protocoles, ne serait-ce qu'en raison de leur impossibilité structurelle (notamment lorsqu'on raisonne sur le passé) » (Orain 2003, 137).

Cet extrait est indéniablement marqué par une approche passeronienne. La parenthèse « (notamment lorsqu'on raisonne sur le passé) » renvoie au pôle historique et herméneutique. Toutefois, l'argumentaire d'Olivier Orain ne développe pas plus cette réflexion alors que ses implications sur la pertinence d'une lecture kuhnienne des sciences sociales (et *a fortiori* de la géographie française) sont majeures. La thèse du plain-pied revendique plutôt la nécessité de prendre en compte et de développer un facteur externe d'explication (la révolution de mai 68) comme conséquence de ce passage reconnaissant la spécificité des SHS. Cette construction interroge d'autant plus que le prisme kuhnien accepte originellement des explications externes sans avoir à recourir à un tel détour par les SHS. Toute cette réflexion m'a conduit *in fine* à formuler et approfondir un questionnement : alors qu'une grande partie de la postérité du schéma kuhnien est due au fait qu'il a ouvert la voie des lectures externalistes des sciences dites dures, Olivier Orain ne se sert-il pas de cet outillage pour internaliser une grande partie de l'explication du changement scientifique ?

La très faible importance des explications externalistes dans la lecture d'Olivier Orain (mise à part cette exception de mai 68) est en effet à mon avis problématique. Par exemple, le peu de poids accordé à l'importance de l'économie spatiale ou à la géographie anglophone est flagrant alors que l'article de Paul Claval (1972) inaugurant *L'Espace Géographique* ne laisse pas beaucoup de doutes à ce propos. Sur ce sujet, dans sa justification du choix du modèle kuhnien, Olivier Orain défend dans sa thèse une prééminence des explications internalistes tout en reconnaissant la nécessité de « se doter de références hybrides si l'on veut sortir du cognitif et avancer des explications sociologiques » (Orain 2003, 115). Si les noms de Bruno Latour et Michel Callon sont alors cités, leurs travaux ne sont pas développés. L'hybridation est par conséquent réduite dans la thèse du plain-pied. Elle passe indirectement que par quelques propositions sans être véritablement analysée.

Par exemple, la proposition suivante correspond à mon sens à une hybridation : « Et cette idée d'une matrice implicite[265], pour peu qu'on reprenne à frais nouveaux la question de son *incarnation*, est capitale si l'on veut comprendre la pérennité de l'école vidalienne » (Orain 2009, 108). Cette idée de l'incarnation introduit en effet un élément externe non négligeable par rapport au modèle kuhnien. Elle est pertinente, à mon avis, autant pour l'époque vidalienne que pour la « nouvelle géographie » avec la figure de Roger Brunet qui a joué un rôle central (Massardier 1996). Cet ajout est important puisque c'est autant, ou même davantage[266], une personne qui rassemble, qu'une matrice intellectuelle. Il faut toutefois ici souligner que cette optique aggrave le problème de l'identification de la rupture entre les vidaliens et les post-vidaliens puisque la même figure, celle de Paul Vidal de la Blache, est restée centrale. De plus, comme le met en avant Olivier Soubeyran dans une visée plus contemporaine, la relation des géographes à Paul Vidal de la Blache est « moins distante que, par exemple, celle des physiciens à l'égard des Ptolémée, Kepler, Galilée… » (Soubeyran 1997, 24). Dans les facteurs d'explication de cette situation, Olivier Soubeyran met en avant « l'évolution fantastique des connaissances » permettant aux physiciens de prendre plus de distance par rapport à ces illustres prédécesseurs. L'idée d'une géographie ayant évoluée sans véritable révolution est sous-jacente à de telles réflexions.

Pour poursuivre cet objectif de prise en compte des réflexions de Jean-Claude Passeron, il me semble important de proposer plusieurs relectures de différents travaux à l'aune de ses propositions. Il s'agit d'avoir une approche moins théorique que la première approche qui lui a été consacrée (*cf.* section Chap8.II) et, par rapport à ce qui vient d'être développé dans cette section, d'avoir une perspective moins axée sur des points de détails. L'objectif spécifique de la section suivante est ainsi de comprendre, plus globalement, ce qu'apporte ce prisme passseronien à travers la relecture de plusieurs travaux : ceux de Thomas Kuhn, de la thèse d'Olivier Orain, mais aussi de mon propre travail de doctorat.

IV. Plusieurs relectures avec le prisme passeronien

Tout d'abord, soulignons que les réflexions de Jean-Claude Passeron permettent de comprendre une grande partie de la construction kuhnienne. En effet, les études de cas sont déterminantes dans son élaboration théorique. La plasticité du langage naturel a

[265] Il est fait référence à la matrice définissant pour Thomas Kuhn un paradigme et étant composée des quatre éléments suivants : généralisations symboliques, la métaphysique, les valeurs et les exemples-types (Kuhn 2008, postface 1969).

[266] Cette distinction et question entre « autant » ou « davantage » n'est pas abordée dans la thèse du plain-pied.

incontestablement servi pour rapprocher les différents cas (Masterman 1970). Il n'existe pas de critère déterminant permettant de discriminer si un courant intellectuel forme un paradigme (ou non) et si une évolution cognitive constitue un changement de paradigme (ou non). Il est pertinent de penser que Jean-Claude Passeron apporte alors des éléments complémentaires à ceux de Thomas Kuhn pour comprendre l'instabilité plus grande des SHS par rapport à la relative stabilité des sciences naturelles[267].

La thèse d'Olivier Orain multiplie les études de cas et tente de les faire correspondre du mieux possible avec le système sémantique créé par Thomas Kuhn. Il utilise également la plasticité du langage naturel. Par exemple, plusieurs éléments de la matrice kuhnienne comme les généralisations symboliques ou les exemples-types n'ont pas le même sens chez Thomas Kuhn et Olivier Orain (*cf.* Chap10). De plus, le besoin de compléter le pôle historique et herméneutique par un pôle expérimental est patent dans la thèse du plain-pied. Tout d'abord avec la petite partie quantitative réalisée et présentée par Olivier Orain. Ensuite, par le chantier empirique ouvert sur la base de la conjecture socio-linguistique pris en charge par ma thèse.

Dans cette perspective, le travail jusqu'ici réalisé, avec la mise en évidence d'une insuffisance du seul pôle expérimental pour administrer la preuve, est en concordance avec les réflexions de Jean-Claude Passeron. Si l'absence de paradigme est justifiée ici par une réflexion épistémologique, il faut mentionner qu'historiquement elle a été intuitée et formulée par quelques tenants de la « nouvelle géographie », comme en témoigne cette citation d'Antoine Bailly et Jean-Bernard Racine :

> « Comment ne pas remarquer, à travers l'analyse des pratiques, l'imprécision des raisonnements, l'absence d'épistémologie propre ? En fait, pragmatique par essence, notre discipline ne s'est pas encore développée comme une activité scientifique "normale", puisque, après les phases empiriques et inductives, elle n'est pas arrivée aux phases théoriques, déductives et axiomatiques. Cet arrêt dans l'évolution est-il inévitable, fondé sur la nature de la discipline ? Ce serait alors à la réflexion épistémologique d'en rendre compte, et de le légitimer » (Bailly et Racine 1978, 7-8).

Par rapport au prologue de ce doctorat où cette citation avait déjà été utilisée (*cf.* section Chap2.II.2.b), l'épistémologie développée par Jean-Claude Passeron permet de mieux comprendre ces constats et questionnements historiques. L'essentiel des raisonnements des SHS ne peut « aligner ses opérations sur des modèles discursifs qui lui sont inaccessibles, comme ceux de la déduction et de l'induction » (Passeron 2006, 365). Or, ce qui a été revendiqué historiquement par les acteurs de la « nouvelle géographie, ce n'est pas tant la dichotomie mise en avant par Olivier Orain (réalisme / constructivisme-nominalisme) que

[267] Ce qui, pour rappel, est présenté comme point de départ de *La Structure des révolutions scientifiques* par Thomas Kuhn (2008, *1962*).

celle entre approches inductives et déductives. En rendant cette dichotomie partielle et assez caduque dans les SHS, c'est l'argument central et initial de ceux qui ont revendiqué une « révolution scientifique » dans la géographie française des années 1970 qui est battu en brèche.

Concrètement, le fait que la géographie post-vidalienne n'ait pas explicité de manière forte ses cadres théoriques ne veut pas dire qu'elle ait été effectuée sans préconception théorique. À l'inverse, le fait que la « nouvelle géographie » ait revendiqué comme injonction l'explicitation de ses cadres théoriques, ne veut pas dire que les formulations de ses problématiques étaient exemptes de connaissances issues d'observations de terrain. Ces affirmations peuvent être rapprochées de celle du cercle herméneutique. Cette notion a été développée principalement par des philologues (Friedrich Ast…) et des philosophes (Martin Heidegger…). Elle permet de mettre en avant que la compréhension d'un texte passe par celles d'ensembles plus grands (œuvre entière de l'auteur, courant littéraire) et, dans le même temps, la compréhension de ces ensembles plus grands passe par celle des textes. Il n'est pas possible de sortir du cercle, de la même manière que les problématiques, les théories et les terrains sont intimement intriqués.

Sur ce sujet, il est possible de se demander si la situation est fondamentalement différente dans les sciences dites dures. Dans *Qu'est-ce que le travail scientifique des données,* Christine L Borgman (2020) répond par la négative quand elle reprend l'analogie rapprochant la construction d'un article d'astronomie de celle d'une maison :

> « Le point de départ est rarement clair : les habitants rénovent et agrandissent une maison sur des années, voire des décennies. Même quand une famille part d'un terrain vague, il peut y avoir eu une structure à cet endroit par le passé. Si on remonte encore le temps, quelqu'un a décidé de comment le terrain serait loti, ce qui a déterminé les possibilités pour la taille et l'orientation de la maison, et ainsi de suite » (Borgman 2020).

Cette perspective affirme *in fine* l'imbrication des théories et des terrains ainsi que l'importance du contexte dans les productions des sciences expérimentales. Elle est congruente avec le modèle kuhnien. Les paradigmes sont ces matrices qui déterminent les possibilités d'une science à un moment donné. *A contrario*, par rapport aux développements de Jean-Claude Passeron, ce témoignage est moins convergent et plus porteur d'une interrogation. En effet, la forte dichotomie entre sciences formelles et expérimentales d'un côté et SHS de l'autre, peut être bousculée par ce témoignage montrant que les sciences expérimentales sont formées aussi par des rapprochements (des liens typologiques et pas uniquement des liens logiques). Cette reconnaissance m'a conduit à réfléchir et à proposer

une approche moins dichotomique, mais ne remettant pas en cause fondamentalement les réflexions de Jean-Claude Passeron.

V. Pour un gradient épistémologique entre sciences formelles, expérimentales et SHS

Même si les exemples pris par Thomas Kuhn relèvent pour la plupart de la physique et non directement du champ mathématique, il me semble, en me basant sur les réflexions précédemment développées, qu'une science formelle reposant sur des liens logiques, est le mode de scientificité le plus susceptible de faire naître des paradigmes profonds. Une nuance doit être pourtant précisée. Quand Kurt Gödel prouve le théorème d'incomplétude, il met certes un coup fatal au programme de Hilbert qui peut être considéré comme le paradigme sur lequel reposaient alors les mathématiques. Toutefois, cela marque aussi à mon sens la fin d'une forme idéale de paradigme. En effet, les mathématiques reconnaissent alors que tout système axiomatique cohérent est nécessairement incomplet. Le mirage d'un système complètement formalisé et totalement logique tombe (Ladrière 1957).

Sans répondre à cette forme idéale inatteignable, l'idéal type du paradigme, tel que défini par Thomas Kuhn repose sur un niveau de formalisation important. Les généralisations symboliques sont un pilier de la matrice disciplinaire définie dans la postface de la *Structure des révolutions scientifiques*. Cette formalisation est importante, car elle permet de sortir d'une partie des ambiguïtés du langage naturel tel que les décrit Thomas Kuhn. Par exemple, le terme « planète » peut avoir facilement des significations différentes selon les paradigmes alors que les concepts de poids et de masse définis à travers la formule P=mg sont plus difficiles à interpréter de manière ambigüe. L'universalité présumée d'une loi permet une forte prééminence du pôle expérimental. Les résultats obtenus doivent sembler en adéquation avec la nature jusqu'à la survenue d'une anomalie kuhnienne[268].

Par rapport à cet idéal type, d'autres situations existent. Il est possible que les sciences expérimentales construisent des paradigmes à partir d'études de cas. Par exemple, pour la théorie de la tectonique des plaques, différentes recherches sur la topographie des fonds marins ont joué un rôle crucial dans son élaboration. Il est vrai qu'il existe maintenant des « modèles numériques complexes du fonctionnement du manteau »[269]. Toutefois, avant cette formalisation avancée, le paradigme avait déjà été accepté sur la base de modèles plus

[268] C'est une des définitions de l'anomalie donnée par Thomas Kuhn : « l'impression que la nature, d'une manière ou d'une autre, contredit les résultats attendus dans le cadre du paradigme qui gouverne la science normale » (Kuhn 2008, 83).

[269] https://planet-terre.ens-lyon.fr/ressource/histoire-tectonique-plaques.xml (consulté le 18/09/2023).

simples et faisant appel à des degrés de généralisations symboliques bien moindres. Le rapprochement entre les études de cas de différents fonds marins a été d'autant plus fort qu'il s'est appuyé sur un modèle commun et des mesures ont montré la validité de ce dernier.

Il peut exister en SHS des modèles ayant un pouvoir explicatif, et même dans certains cas, donnant lieu à des prédictions valables. Toutefois, le pouvoir de ces modèles par rapport aux lois de la physique ou aux grandes théories de la biologie est réduit, car toute société est toujours marquée par de l'imprévisibilité. Pour approfondir ce sujet, soulignons qu'il existe en SHS une grande diversité de modèles avec des degrés de formalisation très variables. Plus le modèle repose sur le langage naturel, plus il est bien souvent difficile de déterminer vraiment un critère de réfutabilité. En outre, il est fréquent que plusieurs schèmes d'intelligibilité (pas forcément formalisé sous forme de modèle) co-existent, donnant lieu à des situations fondamentalement non paradigmatiques.

À l'autre bout du spectre, il me semble nécessaire de distinguer plus fortement que ne le fait Jean-Claude Passeron, les monographies des recherches ne faisant aucunement l'effort de chercher des points d'appui dans une empirie. En effet, dans le cas des monographies, la volonté d'élaborer un système explicatif cohérent et donnant à comprendre les situations observées, amène déjà une présence implicite du pôle expérimental. Même si elle ne prend pas la forme d'un modèle explicité, cette volonté de correspondance avec les situations observées relève déjà à mon avis d'une forme minimale d'expérimentation. Il est évident que ce pôle expérimental est plus facile à identifier quand il mobilise des outils statistiques.

Avant de conclure, il est nécessaire de souligner que le risque de cette approche spectrale est de retomber dans une forme de hiérarchie et de penser de nouveau les SHS par rapport à des modèles de scientificité extérieurs à son fonctionnement. Il est certain que l'approche plus dichotomique de Jean-Claude Passeron permet d'éviter ce risque. Toutefois, une fois ce point conscientisé et énoncé, il me semble que cette approche spectrale permet de mieux représenter une diversité de pratiques autant en SHS qu'en sciences expérimentales. Elle est à mon sens moins problématique, notamment vis-à-vis d'une dichotomie qui prend toujours le risque d'être fantasmée et fétichisante.

VI. Réflexions conclusives

Le texte et l'utilisation de l'application Web ne se prolongeant pas au-delà de la conclusion de cette partie, une dimension plus épistémologique de cette application est ici présentée. Le point de départ de cette réflexion est un colloque sur le goût de l'archive à l'ère numérique dans lequel Sean Takats a proposé un projet qui est pour lui un idéal. Il s'agit de

la création d'une application qui pourrait être utilisée par de nombreux projets de SHS et qui permettrait de rendre compte de la subjectivité de ce qui a été construit. À la fin de son exposé, Sean Takats prend le soin de préciser que « l'objectif n'est pas de transformer les sciences sociales en sciences dures » (Farge et Takats 2018). Nul doute qu'il anticipe ainsi des critiques pouvant être exprimées par certains praticiens de SHS à propos d'une telle application. Certains esprits peuvent en effet être heurtés par cette idée d'une subjectivité captée et rendue par une application.

Par rapport à l'idéal exprimé par ce chercheur, il est certain que l'application ici créée est très limitée en raison de sa transférabilité réduite. Il n'en reste pas moins qu'il est particulièrement intéressant de la réfléchir dans la perspective développée par ce chercheur. L'idée notamment qu'une telle application puisse être perçue comme se rapprochant de méthodes de sciences dures mérite d'être développée dans le cadre de la problématique de mon travail doctoral. En effet, si l'expérimentation menée s'avère concluante (avec un rapprochement des méthodes de sciences dures), utiliser la notion de paradigme dans les sciences sociales y gagnerait en légitimité. Ce qui est en jeu, est l'effort de réduction et de captation de subjectivité des choix successifs opérés. Si l'application a pu en effet me conduire à préciser et à reconnaitre des zones de subjectivité, je pense qu'il n'existe aucun automatisme dans ce processus. Ce dernier me semble bien plus tenir à une volonté personnelle avec, dans ce cas, une possible aide de l'application pour accompagner cette démarche.

Sean Takas ne détaille malheureusement pas les différences entre son projet d'application pour les SHS et les méthodes de sciences dures. Il me semble ici intéressant de développer sur un sujet connexe quelques réflexions à partir du livre de Bruno Latour intitulé *Aramis ou l'amour des techniques* (2020). Sans rentrer dans les détails techniques, *Aramis* était un projet de mini-métro automatique prévu au sud de Paris dans les années 1970-1980. Après l'échec de ce projet, Bruno Latour a réalisé une étude sociologique autour de l'ensemble des réseaux constitués pour réaliser *Aramis*. Il montre notamment comme ce grand programme de construction était divisé en sous-programmes (véhicule, automatisme, réseau électrique, voie…), eux-mêmes divisés en sous-sous-programmes. Une fois qu'un sous-programme marche, il n'est plus nécessaire de s'intéresser vraiment à toute son architecture. Seules les sorties obtenues à partir des entrées données peuvent être considérées sans rentrer dans plus de détail. Un programme informatique est divisé et fonctionne de la même manière. Ce n'est que si un problème apparait, qu'il faut re-rentrer dans la chaîne des fonctions et des sous-fonctions pour trouver d'où vient l'erreur. Ce qui est la base d'une telle construction est la

logique formelle avec des structures du type « si telle action est réalisée ou telle variable est égale à telle valeur, alors telle fonction doit être effectuée ».

En considérant maintenant un projet en SHS comme l'exercice doctoral que je restitue ici, il est évidemment possible de diviser le projet global en sous-projets (élaboration de la problématique, constitution du corpus, choix de la méthode…) et en sous-sous-projets (par exemple pour la constitution du corpus : choix des revues, amélioration de l'OCR, choix des textes et partie de textes…). La différence majeure est que ce n'est pas une logique formelle qui organise et régit cet ensemble, mais un discours argumentatif. Par rapport à cette affirmation, il existe des sous-projets qui sont de mon point de vue de nature différente. Par exemple, par rapport à l'amélioration de l'OCR, il est possible, comme cela a été réalisé, de calculer des scores qui déterminent la réussite plus ou moins grande du projet. Ce sous-projet technique est de mon point de vue achevable, au sens où un score de 100 % (ou très proche) peut être obtenu et reconnu par tous. *A contrario*, le sous-projet global de la constitution du corpus ne peut être clôturé. D'autres recherches sur la même question pourraient mettre en avant des corpus différents en utilisant des argumentations valables et obtenir des résultats distincts. Il existe ici un inachèvement fondamental qui, à la différence d'*Aramis,* n'aboutit pas à un échec.

Pour finir et conclure cette deuxième partie, il est utile de synthétiser les résultats obtenus. Les représentations du changement sémantique autour des termes « espace » et « milieu » obtenues à l'aide des méthodes de plongements de mots ont mis en évidence des difficultés pour valider l'hypothèse sémantique affirmée par Olivier Orain. Ces difficultés ont conduit à émettre des premières critiques importantes à l'encontre de la lecture kuhnienne de la géographie française. Pour autant, les résultats obtenus n'ont pas été directement déterminants, donnant lieu à plusieurs discussions serrées entre arguments favorables et défavorables à la thèse d'Olivier Orain.

Ce qui a été finalement déterminant est tout d'abord une réflexion sur le processus de construction des résultats, puis une discussion approfondie qui a montré la pertinence d'une prise en compte des réflexions de Jean-Claude Passeron sur cette problématique. Ce processus a abouti à un approfondissement notable de la critique des propositions d'Olivier Orain. En effet, Jean-Claude Passeron développe une opposition ferme et argumentée à toute lecture paradigmatique des SHS. Ses réflexions théoriques ont permis de mieux comprendre plusieurs travaux (ceux de Thomas Kuhn, de la thèse d'Olivier Orain et les miens) et ont éclairé tout autant des réflexions historiques des tenants de la « nouvelle géographie » que l'état épistémologique de la géographie contemporaine.

Les développements passeroniens n'ont pas été pris pour autant pour argent comptant et une proposition épistémologique a été réalisée pour tenter une présentation moins dichotomique et plus spectrale de plusieurs modes de scientificité.

À ce stade, il est nécessaire d'expliquer pourquoi cette thèse ne se termine pas sur les résultats et les discussions qui viennent d'être développées. En effet, par rapport à un type de plan nommé IMReD (Introduction, Méthode, Résultat et Discussion), très utilisé pour rendre compte de travaux quantitatifs, il serait pertinent de s'acheminer vers la conclusion de cette thèse. En miroir de la thèse d'Olivier Orain qui se conclut sur l'annonce d'un projet de recherche quantitatif beaucoup plus ample que celui qu'il a mené, il serait possible d'ouvrir ce travail sur un projet qualitatif de réévaluation globale des arguments avancés dans la thèse du plain-pied. Parallèlement, le développement d'une lecture alternative de l'histoire de la géographie, basée sur une approche passeronienne serait mis en avant comme un chantier de première importance restant à réaliser.

Mais, je dois avouer ne pas avoir souhaité remettre à plus tard l'effectuation de ce travail. Cette prise en charge s'explique non seulement par la forme d'indétermination éprouvée lors des analyses de mes résultats quantitatifs (*cf.* section Chap7.II), mais aussi par l'envie de ne pas reporter le développement des pistes de travail ayant été identifiées comme majeures une fois cette partie quantitative achevée. C'est pourquoi le projet qui vient d'être développé - une analyse plus en détail des arguments de la thèse du plain-pied et le développement d'une lecture passeronienne - constitue le cœur de la partie suivante.

Troisième partie

Analyses non quantitatives de la lecture kuhnienne d'Olivier Orain et développement en contrepoint d'une lecture passeronienne

Le titre de cette partie a fait l'objet de nombreuses réflexions avant d'aboutir à l'expression « analyses non quantitatives ». Nul doute que certains penseront que l'utilisation du syntagme « analyses qualitatives » aurait été plus simple et donc plus claire. Mais, ce choix n'a pas été retenu pour souligner la présence déjà existante d'appréciations qualitatives[270] intriquées dans la construction des analyses quantitatives (*cf.* Part1 et 2). Ces constatations rejoignent celle de François Rastier, qui dans l'ouvrage *La mesure et le grain*, renverse une conception habituelle en affirmant : « Ce n'est pas l'instrumentation qui permet l'interprétation, mais l'inverse » (2011, 51). En effet, l'interprétation joue un rôle majeur dès la constitution du corpus et l'élaboration de la méthodologie[271]. Cette affirmation remet en cause la dualité classique « quantitatif / qualitatif » que le pas de côté – avec l'utilisation de l'expression « analyses non quantitatives » – permet de ne pas endosser frontalement.

Il faut souligner que la complexité de ce choix de dénomination provient en grande partie du problème crucial de l'objectivation et du fait que la quantification soit devenue parfois « synonyme de respectabilité scientifique » (Rastier 2017, 10). La proposition retenue n'évite pas l'écueil de l'opposition (quantitatif / non quantitatif) et reste insatisfaisante. Toutefois, elle permet de rappeler l'existence des réflexions effectuées sur ce sujet ainsi que leur inaboutissement.

Dans un premier temps, ces deux approches du réel (quantitatif / non quantitatif) peuvent être exprimées assez simplement :

- Dans la première, quantitative, le comptage d'évènements (pouvant se résumer dans notre cas à l'apparition ou non d'un terme dans le contexte[272] d'un autre terme) est central. Cette approche oblige à une préparation des données en amont et à des interprétations de résultats en aval.
- Dans la seconde, non quantitative, la construction de sens passe plutôt par des analyses de l'usage des mots et de la littérarité. La dimension historique est alors essentielle, car c'est elle qui permet de remettre les textes dans leurs contextes[273] et de mieux les comprendre.

---

[270] En effet, nommer cette partie « analyses qualitatives » aurait pu en creux faire penser à tort qu'il n'existait pas de telles analyses dans les parties précédentes.

[271] Les développements réalisés par Gaston Bachelard à ce sujet méritent ici d'être mentionnés : « Alors il faut que le phénomène soit trié, filtré, épuré, coulé dans le moule des instruments, produit sur le plan des instruments. Or les instruments ne sont que des théories matérialisées. Il en sort des phénomènes qui portent de toutes parts la marque théorique » (Bachelard 2020, *1934*, 16).

[272] Le contexte est ici une fenêtre de n mots à droite et à gauche de chaque terme.

[273] Le contexte ne se limite pas alors évidemment à une fenêtre spatiale de n mots à droite et à gauche. Il déborde le texte lui-même en intégrant la dimension historique, notamment les contextes de production et de réception.

Ces descriptions synthétiques sont loin d'aborder et de régler la problématique de fond qui porte sur l'articulation (ou non) de ces deux approches. Cette problématique dépasse le cadre de la géographie comme le montrent les réflexions du sociologue Lionel-H Groulx :

> « Quiconque ouvre un livre de méthodologie de recherche publié dans les années 90 ne manquera pas de trouver un chapitre sur la recherche qualitative. L'auteur présente l'opposition entre la recherche qualitative et la recherche quantitative en traçant leur différence ou leur incompatibilité quant à la connaissance de l'objet, du devis mis en œuvre, des instruments de cueillette de données et des modes d'analyse. On aboutit, dans la plupart des cas, à l'élaboration d'un tableau où sont opposées termes à termes les diverses dimensions de chacune des recherches qui renvoient à des paradigmes méthodologiques incommensurables, pour reprendre la formule de T. Kuhn » (Groulx 1997).

Cette citation permet de comprendre en quoi cette réflexion amorcée peut être mise en rapport avec la problématique étudiée par Olivier Orain et par moi-même. Elle permet aussi de mettre en avant une différence générationnelle entre la décennie 1990 – moment où Olivier Orain fait ses études et où cette dualité « quantitatif / qualitatif » est fortement affirmée – et la période actuelle où *a contrario*, l'enseignement[274] et la recherche mobilisent, voire revendiquent, fréquemment des postures méthodologiques valorisant une articulation méthodologique.

Il faut également souligner qu'une autre difficulté réside dans le fait que ces deux démarches, quantitative et non quantitative, peuvent *in fine* se référer à un même cadre théorique, celui d'une herméneutique matérielle. Ce cadre défini par Peter Szondi (1975) peut être positionné par rapport à deux pôles : d'une part, une démarche herméneutique philosophique (Gadamer 1996) qui développe des réflexions avant tout théoriques ; d'autre part, un positionnement positiviste prétendant résoudre objectivement les problèmes. Comme l'analyse Rossana De Angelis (2020), le cadre de l'herméneutique matérielle de Peter Szondi est distinct de ces deux pôles tout en se situant dans un entre-deux. En effet, il s'agit d'« articuler une méthode historique et une méthode systématique pour mettre en place une nouvelle méthode, elle aussi ancrée dans l'histoire » (De Angelis 2020, 3). Cette définition est indéniablement proche du positionnement passeronien (*cf.* section Chap8.II).

La question du « systématisme »[275] mentionnée par Rossana De Angelis est une difficulté majeure de la construction qui suit. En effet, il est impossible de prendre en charge toutes les réflexions d'Olivier Orain du fait de leur abondance et leur complexité. La recherche

[274] Les Travaux Dirigés (TD) que j'ai pu donner en L1 et L3 de sociologie à l'Université Grenoble-Alpes confirment complètement cette affirmation avec à chaque fois des TD pensés et effectués en articulation : enquête ethnographique / observation quantifiée du monde social, enquête par entretien / par questionnaire et analyse qualitative de documents / analyse quantitative de corpus.

[275] Question qui se pose, rappelons-le, aussi pour les études non quantitatives.

d'un « systématisme » s'est donc jouée à un autre niveau, dans le développement d'un plan de recherche permettant de faire le tour des points considérés comme les plus importants. L'approche retenue privilégie trois axes :

- Le premier prend comme point d'entrée les grandes étapes du schéma explicatif détaillées par Thomas Kuhn et reprises en partie par Olivier Orain : de la proto-science à la science, du paradigme à l'anomalie, de l'anomalie à la crise, de la crise au nouveau paradigme.
- Le deuxième axe revient sur les éléments de la matrice disciplinaire qui sont également développés par Thomas Kuhn et réinterprétés par Olivier Orain : les généralisations symboliques, la métaphysique, les valeurs et les exercices-types.
- Enfin, le dernier axe détaille les positions de plusieurs chercheurs sur lesquels s'appuie Olivier Orain : Ian Hacking, Hilary Putnam et Jean-Michel Berthelot.

L'objectif a été de détailler plusieurs difficultés générées par la lecture kuhnienne du plain-pied et de les aborder en contrepoint par l'épistémologie passeronienne. Les outils utilisés sont en grande partie les mêmes que ceux mobilisés par Olivier Orain dans sa thèse : analyse critique et commentaire de texte. J'ai également parfois développé de manière significative quelques éléments extérieurs (ne provenant ni de la thèse d'Olivier Orain, ni des réflexions de Jean-Claude Passeron), quand j'ai jugé que ces développements additionnels apportaient une plus-value dans la compréhension du propos.

Enfin, il faut ici rappeler que cette analyse détaillée de la thèse d'Olivier Orain se justifie par le fait que son travail est, dans un contexte historiographique francophone, celui qui a élaboré de la façon le plus systématique, le transfert, l'adaptation et la justification d'une lecture kuhnienne de l'histoire de la géographie. Soulignons qu'une difficulté rencontrée dans cette partie est due à la succession des analyses. Quand les propos d'un auteur sont discutés et interprétés par Olivier Orain, et qu'ensuite, je propose de nouvelles analyses, les discours se déploient sur trois niveaux (celui de l'auteur initial, d'Olivier Orain et le mien) avec des risques de confusion. J'ai essayé de préciser autant que possible ces différents niveaux, car il existe indéniablement des complexités dans ces successions d'analyses.

Le chapitre suivant développe le premier axe retenu en détaillant les différentes étapes du schéma kuhnien appliquées à la géographie française.

Chapitre 9 :
Une lecture historique par les étapes du schéma kuhnien

Ce chapitre examine les étapes du schéma kuhnien (de la proto-science à la science, du paradigme à l'anomalie, de l'anomalie à la crise, de la crise au nouveau paradigme) en reprenant les analyses effectuées par Olivier Orain dans sa thèse. En contrepoint, je propose des développements construits à partir de réflexions issues du livre *Le Raisonnement sociologique : Un espace non poppérien de l'argumentation* (Passeron 2006). Plusieurs points (*bataille des Annales*, analyses du *détail du monde* de Romain Bertrand…) ont été ajoutés par rapport à la thèse du plain-pied, car ils m'ont paru apporter des éléments complémentaires intéressants.

La première étape est celle du passage de la géographie vidalienne aux géographies post-vidaliennes (qui correspond logiquement en suivant le schéma kuhnien au passage d'un état pré-paradigmatique au premier paradigme, de la proto-science à la science).

I. De la géographie vidalienne aux géographies post-vidaliennes

Une question, peu prise en charge par Olivier Orain, est celle de la rupture pouvant justifier l'existence d'un paradigme post-vidalien par rapport à la période vidalienne. Dans son archéologie du réalisme géographique, il cite, en premier et à juste titre, Lucien Gallois comme prometteur d'un réalisme fort dans la géographie française du début du XX^{ème} siècle (Orain 2003, 39). Lucien Gallois, qui a été l'élève, puis le fidèle lieutenant de Paul Vidal de la Blache, a eu une place particulièrement centrale dans la géographie française de son époque. Il a co-dirigé les *Annales de géographie* à partir de 1895 avec Paul Vidal de la Blache. Il l'a remplacé ensuite comme maître de conférences à l'École Normale Supérieure en 1898.

Dans le détail, les évènements historiques qui ont donné une place centrale à ce géographe ne sont toutefois pas développés par Olivier Orain. Or, ces évènements peuvent être considérés comme cruciaux, car ils constituent la transition entre la géographie vidalienne et la géographie post-vidalienne. Cette situation justifie l'explicitation de ces évènements historiques fréquemment désignés sous l'expression de « bataille des Annales » (Soubeyran 1989).

1. Analyse de *la bataille des Annales*

Dans les premières années de la revue les *Annales de géographie* (1892-1894), deux conceptions de la géographie s'opposent. La première est portée par Marcel Dubois (1856-1916), alors chef de file de la géographie coloniale. Pour lui, le poids des facteurs naturels n'est pas déterminant pour comprendre les sociétés. *A contrario*, pour Lucien

Gallois (1857-1941), la nature, et en particulier la géologie, doit être étudiée en premier comme principal facteur d'explication. À cette divergence scientifique, s'ajoute une opposition politique : Marcel Dubois est anti-dreyfusard tandis que Lucien Gallois et Vidal de la Blache sont dreyfusards. Tous les détails de cette *bataille des Annales* ne sont pas bien connus, notamment l'importance des facteurs politiques et scientifiques par rapport à son issue, mais elle se solde par le départ de Marcel Dubois.

C'est donc le choix d'une géographie centrée sur les « régions naturelles » plutôt que sur une perspective aménagementiste, qui est alors privilégié. Ce choix accentue sans conteste le réalisme de la géographie française. En effet, la compréhension du monde passe alors avant tout par l'explicitation des facteurs naturels qui sont plus stables et moins relatifs que des situations dépendant des actions humaines. Toutefois, il faut souligner qu'il existe dans les travaux d'Olivier Soubeyran (1997), spécialiste de cette *bataille des Annales*, une opposition manifeste à une perspective kuhnienne :

> « Ainsi, on reconnaît à Kuhn l'avancement considérable qu'il a permis en montrant que la science n'est pas seulement une pratique cognitive, mais aussi sociale. Seulement, on se rend compte aussi que sa modélisation passe à côté (et donc ne la modélise pas) de la complexité et surtout de la nature des dimensions mobilisées dans la transformation de notre pensée géographique. Or, cette critique que P. Claval, V. Berdoulay et d'autres ont formulée, est doublement convaincante. D'une part, elle s'appuie sur l'étude de notre corpus disciplinaire et d'autre part sur une critique parallèle qui s'effectuait dans l'histoire des sciences dures elle-même » (Soubeyran 1997, 27).

Les analyses d'Olivier Soubeyran se fondent, entre autres, sur une étude approfondie de la géographie coloniale dans les *Annales de Géographie* et sont reprises en partie par Pascal Clerc qui montre bien l'absence de changement de paradigme :

> « La géographie coloniale ne disparaît pas avec la mise à l'écart (ou le départ volontaire ; on ne sait pas trop) de M. Dubois et la géographie appliquée, comme l'engagement des géographes dans la cité, non plus. La simple comptabilité des écrits relatifs aux colonies dans les *Annales de Géographie* pourrait suffire à installer un sérieux doute [Deprest 2009]. Quasiment jusqu'à la Seconde Guerre mondiale, les *Annales* – le lieu du « rejet » duboisien – publient de nombreux articles relatifs aux colonies. Et la vision de L. Gallois n'est pas celle qui domine ; la plupart des articles privilégient une approche mixte, associant des approches descriptives des genres de vie à des réflexions plus engagées sur les possibilités d'aménagement des espaces [Clerc 2006] » (Clerc 2020, 23).

Si la rupture entre géographie vidalienne et post-vidalienne ne peut donc pas être située à ce niveau, est-il possible de l'identifier par la suite en s'appuyant sur les arguments qu'Olivier Orain développe dans sa thèse ?

2. Une discontinuité réduite et problématique

Un premier argument se trouvant dans la thèse du plain-pied pour justifier de la rupture entre vidaliens et post-vidaliens est celui de l'existence d'une réduction de la littérarité entre le *Tableau de la géographie de la France* de Vidal de la Blache (1903) et *La plaine picarde* d'Albert Demangeon (1905). Un élément problématique par rapport à cet argument provient d'un chapitre écrit par Marie-Claire Robic[276] et consacré à une lecture du *Tableau de la géographie de la France* comme un exemple-type kuhnien (2003). En effet, l'auteure insiste sur la filiation entre *La Plaine picarde* de Demangeon et le *Tableau de la géographie de la France* de Vidal de la Blache. Elle précise que cette filiation n'est pas une reconstruction épistémologique *a posteriori* mais a été « reconnue d'emblée par la critique » (Robic 2003, 93). Or, cette filiation est justement très peu explicitée par Olivier Orain : le fait même que ce qu'il considère comme un paradigme post-vidalien se construise sur un exemple-type vidalien est profondément problématique. En effet, le schéma kuhnien nécessite en amont l'explicitation d'une rupture épistémologique.

Pour nourrir cette idée de « rupture », Olivier Orain argue du fait qu'en diminuant la présence de la rhétorique, les post-vidaliens auraient abandonné l'exposition de la « fêlure entre l'écriture et son référent » (Orain 2003, 45) provoquant l'installation d'un paradigme reposant sur un réalisme naïf. Cet argument subtil reste infalsifiable, car il relève de suppositions psychologiques[277] sur des pratiques de recherche d'un autre siècle. Toutefois, par rapport au passage d'une proto-science à une science qui marque la première étape du schéma construit par Thomas Kuhn, l'idée même d'un réalisme naïf des post-vidaliens pose problème. En effet, l'établissement d'une démarche scientifique, notamment par rapport au modèle des sciences dures qui sert de référence à Thomas Kuhn, suppose dans un premier temps le dépassement de ce même réalisme naïf.

Un autre argument d'Olivier Orain par rapport à cette pré-supposée "rupture épistémologique" repose d'une part sur la multiplicité des pistes de travail ouvertes par l'œuvre de Paul Vidal de la Blache et d'autre part sur une réduction ultérieure effectuée par les post-vidaliens. Si cet argument est exact, il ne permet pas, à mon avis, d'instituer une rupture épistémologique équivalente au passage d'un état pré-paradigmatique au premier paradigme dans le descriptif kuhnien. Olivier Orain a d'ailleurs conscience que la discontinuité qu'il cherche à mettre en avant est réduite. C'est ainsi que dans sa conclusion, il revient sur les « différentes questions historiographiques qui demeurent en suspens » (*Ibid*

[276] Directrice de thèse d'Olivier Orain.

[277] Le degré de conscience de la rupture entre les « mots » et les « choses » et la volonté sous-jacente à l'utilisation des figures de style.

2003, 352). Le premier point problématique qu'il reconnaît concerne les « (dis)continuités entre P. Vidal de la Blache et ses divers "élèves" » (*Ibid* 2003, 352). Cette formulation « (dis)continuités » exprime clairement et typographiquement l'absence d'une rupture épistémologique telle que celles formalisées par Thomas Kuhn (2008, *1962*).

Sur le fond de ce questionnement −un changement dans l'appréhension et le rendu du « réel » entre la géographie vidalienne et post-vidalienne − la lecture de l'ouvrage de Romain Bertrand (2019) intitulé *Le détail du monde* m'a conduit à effectuer plusieurs rapprochements intéressants permettant de mieux comprendre une partie des idées et arguments avancés par Olivier Orain, rapprochement que développe la section suivante.

3. Mises en parallèle à partir du *détail du monde*

Romain Bertrand enquête dans cet ouvrage sur les conceptions et les pratiques des premiers naturalistes. En reprenant ces développements, je suis parfaitement conscient du pas de côté que j'effectue par rapport à l'argumentation d'Olivier Orain : en effet, ce dernier rapproche Vidal de la Blache des écrivains naturalistes (Emile Zola, Guy de Maupassant…) alors que les naturalistes présentés par Romain Bertrand sont Alfred Russel Wallace, Charles Darwin, Alexander Von Humboldt, Tom Harisson… La piste développée ici semble donc, dans un premier temps, bien différente de celle d'Olivier Orain.

Toutefois, un premier point particulièrement intéressant, largement souligné par Romain Bertrand, renvoie aux catégories mentales (et pratiques) de l'exhaustivité et de l'objectivité, présentes chez Alexander Von Humboldt (1769-1859), à la fois naturaliste et géographe :

> « Si d'aucuns ont pu voir en Humboldt le "dernier savant universel", c'est qu'à travers lui se réitère un très ancien défi : celui d'une "histoire naturelle" de plain-pied avec le monde, capable de prendre dans les rets de son récit la totalité des êtres et des faits, quelle que soit la taille de leurs causes ou de l'envergure de leurs conséquences » (Bertrand 2019, 38).

L'objectif de réalisme passe par la mesure, la recherche de structures, mais aussi par le style littéraire pour saisir les choses « sans tomber dans la sécheresse de la science pure » (Humboldt cité par Bertrand, 2009, p. 50). Ensuite, Romain Bertrand analyse l'œuvre de Alfred Russel Wallace, *The Malay Archipelago* (1869) qui ne recouvre pas, selon lui, la même "histoire naturelle" que celle d'Humboldt :

> « Quelque chose y fait défaut, qui interdit la totalisation de la description. En lieu et place d'un tableau général de la nature, *The Malay Archipelago* donne à voir, par le jeu de la découpe en chapitres étanches, une succession de contributions spécialisées – l'une dédiée aux oiseaux, l'autre aux fougères, une autre encore aux humains. Il y manque un liant entre les êtres – quelque chose comme "le vent qui

> murmure à travers les buissons de bambous" dans les pastorales de Bernadin de Saint-Pierre, c'est-à-dire l'écho de la rime du monde » (Bertrand 2019, 88).

Il y a là un parallèle évident avec ce que veut montrer Olivier Orain dans sa différenciation géographie vidalienne / post-vidalienne quand il met en avant une diminution de la littérarité et une progression du plan à tiroirs.

Il est ici particulièrement intéressant de remarquer que pour Romain Bertrand le « plain-pied » est du côté de Alexander Von Humboldt, c'est-à-dire, en suivant cette analogie, plutôt du côté de la géographie vidalienne. Pour comprendre cela, il faut préciser que, pour Romain Bertrand, l'objectif originel du travail naturaliste est de combler la faille première et inéluctable entre les mots et les choses. Les démarches les plus avancées dans ce domaine sont à chercher pour lui du côté des poètes, et plus généralement des artistes. En effet, les saisies de la texture du monde, des phénomènes, des expériences sensibles sont toujours singulières. Les réflexions d'Henri Bergson (1934) ou de Maurice Merleau-Ponty (1964) à propos de l'art, permettent de facilement comprendre pourquoi les démarches artistiques sont particulièrement aptes à saisir ces singularités[278]. Le processus est toutefois toujours incomplet… et le plain-pied, donc, jamais complètement atteint.

Dans cette perspective, il existe toujours une tension résultant des écarts entre les mots et les choses, même si des tentatives multiples essayent de réduire ces écarts. Par exemple, Romain Bertrand développe le cas d'une gravure du volcan Chimborazo constituée d'une double partie : à gauche, un dessin rendant compte du relief et de l'étagement de la végétation ; à droite, un ensemble de mots correspondant à cette végétation. Romain Bertrand analyse cet agencement :

> « Loin de combler le fossé entre les mots et les choses, la gravure le rend plus visible, plus sensible que jamais. Mais à l'instar de l'art japonais du *kintsugi*, qui use de pâte d'or pour restaurer les porcelaines cassées, il s'agit moins d'abolir la brisure que de la conjurer, de déjouer son péril en l'exposant. Le portrait du Chimborazo d'Alexandre de Humboldt est donc un échec glorieusement exhibé : non la preuve d'une victoire, mais le vestige d'une audace » (Bertrand 2019, 49).

Il me semble possible et signifiant de rapprocher cette citation de ce qu'Olivier Orain qualifie de « conscience de l'effectuation littéraire » chez Paul Vidal de la Blache (*cf.* section Chap9.I.2). La rhétorique, le style de cet auteur n'aurait pas servi à « abolir la brisure »[279] entre les mots et choses, mais à « déjouer son péril en l'exposant » pour reprendre les termes

[278] Les pratiques contemporaines de ces quinze dernières années mettent largement en avant ces approches artistiques posant la question de leur articulation avec des régimes épistémologiques plus marqués par la rationalité scientifique (Volvey 2014).

[279] Il faut ici souligner que le terme « brisure » a l'inconvénient de pouvoir faire penser qu'il a existé auparavant un état où cette brisure n'existait pas, ce qui n'est pas le propos de Romain Bertrand pour qui un écart entre les mots et les choses est toujours présent.

de Romain Bertrand. L'argument, même s'il est soutenu ici par cette gravure plus explicite, reste au demeurant infalsifiable et nourri par une rhétorique importante.

Pour finir ce détour, il me semble intéressant de revenir sur la critique effectuée par Romain Bertrand à propos du livre d'Elisée Reclus l'*Histoire d'un ruisseau* (1869) :

> « Alors qu'avec Humboldt, on est toujours *là* et pas ailleurs – puisque les choses et les êtres, lestés de leurs coordonnées, sont saisis dans la singularité de leur surgissement, l'*Histoire d'un ruisseau* n'est l'histoire d'aucun ruisseau en particulier. Plutôt une série de digressions sur l'imaginaire de la source et du méandre, entrelardées de réminiscences sentimentales qui dégouttent sur les horizons comme une gouache sur un pastel. Encombrée d'elle-même, l'idée bave sur le phénomène, jusqu'à émousser sa silhouette, pèse sur ses contours jusqu'à les ébrécher. C'est la nature, oui, mais en pointillé » (Bertrand 2019, 89).

Contextualisons cette critique : Romain Bertrand précise que son livre *Le détail du monde* résulte plus d'une « rêverie » (*Ibid* 2019, 241) que d'une écriture guidée par les règles ordinaires de la démonstration. Ce statut lui permet de développer un style faisant la part belle aux images évocatrices et d'enjoliver certaines différences. Cette critique de l'abstraction chez Elisée Reclus n'est pas à prendre au premier degré, car les analyses de Romain Bertrand oublient de préciser que le « ici et maintenant » est une modalité importante de la géographie reclusienne, notamment par rapport à ses revendications anarchistes. Il faudrait aussi rappeler en amont que *l'Histoire d'un ruisseau* (1869) est un livre qui a une vocation plus pédagogique que scientifique. Les analyses stylistiquement recherchées et d'une prose soutenue de Romain Bertrand renvoient dès lors plutôt à une manière d'exprimer et de mettre en valeur une nostalgie à l'endroit d'une valorisation scientifique de la singularité[280].

Ces développements à partir du *Détail du monde* permettent à la fois d'alimenter la thèse d'Olivier Orain (2003) et de la mettre à distance. Par rapport à la complexité de la notion de plain-pied[281], il est certes possible de défendre que son utilisation ne renvoie qu'à une image, mais cette perspective n'est pas sans conséquence. S'il est évident que les géographes vidaliens et les post-vidaliens ont revendiqué une légitimité venant du « terrain » (Calbérac 2010), il ne suffisait pas pour eux de se rendre sur le terrain pour produire de la bonne géographie. Même s'ils ont occulté une partie des médiations existantes entre le terrain et le rendu scientifique, ces médiations existaient et il est dommageable que l'épistémologie

[280] Cet aspect trouve son apogée dans la conclusion de l'ouvrage : « Avec Bernadin de Saint-Pierre, avec Humboldt, avec Wallace, avec Ponge encore, le pluriel du monde est de retour : l'infinie variété des êtres vivants à nouveau se donne à voir. L'arbre compte plus que la forêt, l'oiseau plus que la nuée. Ainsi nous est-il donné de prendre soin du monde. Car les êtres naturels sont comme les êtres chers : il n'est possible, pour les aimer tous, que de les aimer un par un » (Bertrand 2019, 240).

[281] Rappelons que dans l'optique de Romain Bertrand, la géographie vidalienne est plus de plain-pied que la géographie post-vidalienne contrairement à l'optique d'Olivier Orain.

entretienne la fiction de leur inexistence. Le simple fait d'avoir observé ne suffisait pas. Toutes les données recueillies n'avaient pas la même validité. Le terrain était une condition nécessaire, mais non suffisante. De plus, tous les énoncés, dits ou écrits, n'avaient pas la même véridicité. La qualité des descriptions et la plausibilité des interprétations dépendaient de la qualité de l'écriture reliant avec plus ou moins de finesse, de justesse et de cohérence des éléments empiriques.

Ces remarques s'inscrivent totalement dans les réflexions théoriques de Jean-Claude Passeron (*cf.* section Chap8.II). La section suivante développe quelques éléments connus de la géographie vidalienne pour approfondir une lecture passeronienne de cette dernière.

4. Une géographie vidalienne analysée sous un prisme passeronien

Les analyses d'Olivier Soubeyran et Vincent Berdoulay (1991), en soulignant l'importance du néo-lamarckisme dans la géographie vidalienne, ont mis en avant plusieurs éléments correspondants au descriptif développé par Jean-Claude Passeron dans *Le raisonnement sociologique* (2006). L'attachement de Paul Vidal de La Blache au néo-lamarckisme s'explique par l'importance donnée par cette théorie au milieu dans l'évolution. Toutefois, alors que les néo-lamarckiens valorisent fortement la méthode expérimentale, les limites de cette dernière sont reconnues et même utilisées comme argument par Vidal de la Blache comme le soulignent Olivier Soubeyran et Vincent Berdoulay :

> « En ne se concentrant pas sur la recherche exclusive de conditions de laboratoire pour mieux contrôler son objet, la géographie humaine selon Vidal peut rester fidèle à la perspective néo-lamarckienne des relations homme-milieu. Elle s'attache alors à l'initiative du sujet et au caractère composite du milieu. En somme, c'est le « toutes choses étant égales par ailleurs » que permettent de négliger les conditions de laboratoire, que Vidal se propose justement d'analyser pour en montrer l'importance et la complexité » (Berdoulay et Soubeyran 1991, 627).

Cet argument est similaire à celui mis en avant par Jean-Claude Passeron pour expliquer pourquoi les SHS se distinguent des sciences expérimentales et ne connaissent pas des successions de paradigmes telles que Thomas Kuhn les décrit. La faiblesse d'une formalisation théorique de la géographie vidalienne peut alors se lire comme un positionnement stratégique par rapport à son projet scientifique. Il s'agit de garder une souplesse d'approche par rapport à des situations hétérogènes et évolutives. Le fait de vouloir « couvrir le monde » et d'en rendre compte correspond pour Vidal de la Blache et les post-vidaliens à un enjeu fort, qui n'est pas réductible à l'affirmation d'un réalisme intégral, mais qui renvoie à une recherche compréhensible de véridicité. La recherche de systématicité explique, à l'époque post-vidalienne, des parcours et des exposés les plus

complets possibles. Si la géographie s'attache pour Vidal de la Blache « au repérage de ce qui est fixe et permanent » (1903, 351), il est également reconnu que : « ce n'est pas une minime différence, que de considérer les faits comme des entités fixes, ou que d'y voir l'aboutissement de développements antérieurs, si lentement que marche l'horloge » (*Ibid* 1899, 107). Cette reconnaissance rattache la géographie à une dimension historique. Il existe une proximité avec l'affirmation (bien ultérieure) de Jean-Claude Passeron montrant que le raisonnement sociologique est fondamentalement historique.

Toutefois, les arguments développés par Jean-Claude Passeron sont évidemment loin d'être ceux de Vidal de la Blache. L'approche de l'entre-deux défendue par le sociologue « mêlant la sémantique du récit historique à la grammaire du modèle expérimental » (Passeron 2006, 162) laisse dans le cas de la géographie vidalienne un manque majeur du côté de la « grammaire du modèle expérimental ». Des exemples de terrains nourrissent les développements, mais il n'existe pas de grammaire spécifique, du moins telle que la pense Jean-Claude Passeron qui prend ici en compte l'apport des enquêtes quantitatives. De plus, à la différence de Vidal de la Blache, la société n'est pas un organisme pour Jean-Claude Passeron. Ces limites marquent des différences notables entre deux chercheurs appartenant à des périodes scientifiques totalement distinctes (Vidal de la Blache s'inscrit en partie dans l'héritage positiviste d'Auguste Comte[282] mêlé à un vitalisme alors que les réflexions de Jean-Claude Passeron prennent en compte à la fois les apports et les limites des démarches structuralistes et quantitatives dans les SHS).

Une fois ces différences explicitées, une lecture passeronienne permet de mieux comprendre une géographie classique prise dans un entre-deux fortement dominé par le pôle de la « sémantique du récit historique » (Passeron 2006, 162) tout en n'échappant pas à des tensions avec le pôle expérimental. Une lettre de Marcel Dubois à Ernest Lavisse[283] illustre parfaitement ce positionnement.

> « Nous avons de francs ennemis, c'est déjà quelque chose, et je sais d'où viennent les coups. Dans un certain monde, on ne veut admettre comme géographes que des fidèles du Muséum, de l'Observatoire ou du laboratoire de géologie du Collège de France. Or tant que nous ne serons rien dans aucun corps enseignant ou délibérant, tant qu'à défaut d'un patron convaincu comme vous, nous serons réduits à osciller de l'histoire aux sciences, nous serons désarmés. Je regrette de n'avoir pas dix ans de plus pour dire avec autorité ce que cette condition d'hybrides a de lamentable pour les géographes » (correspondance de Dubois à Lavisse, 13 mai 1892, citée par Ginsburger 2017).

[282] *Cf.* Bibliographie : (Comte, 1830).

[283] Figure majeure de l'historiographie française à la charnière des XIX et XX[èmes] siècles (1842-1922) qui a écrit notamment une histoire de France dont le *Tableau de la géographie de la France* (1903) de Paul Vidal de la Blache est le premier tome.

L'inconfort de la situation d'entre-deux de la géographie est, on le voit, fortement exprimé. Contrairement au vœu qu'exprime Marcel Dubois dans cette lettre, la tension avec le pôle expérimental n'a fait que grandir : d'autres disciplines que les naturalistes ont pris le relai du questionnement « scientifique » de la géographie, notamment la sociologie avec François Simiand (1909). Si la référence à Paul Vidal de la Blache a été unanimement partagée, il a existé une multiplicité de positionnements scientifiques, multiplicité permise par l'œuvre même de Paul Vidal de la Blache qui laissait ouverte une diversité de pistes de recherche. S'il n'est pas possible d'explorer toute la complexité des géographies post-vidaliennes avec des convergences, des tensions, des reprises et des voies originales plus ou moins suivies, il convient cependant d'aborder et de développer quelques points par rapport à ce qui est développé dans la thèse du plain-pied.

II. Les géographies post-vidaliennes

1. Composition et décomposition des géographes « réalistes »

À la figure de Lucien Gallois comme premier représentant du réalisme (*cf.* section Chap9.I.1), Olivier Orain ajoute ensuite celle d'Emmanuel de Martonne (1873-1955). L'objectif n'est pas ici de remettre en cause le « réalisme » de De Martonne, mais de nuancer un point de la présentation effectuée par Olivier Orain. En effet, le travail d'Emmanuelle Boulineau (2001) sur De Martonne situe ce chef de file de la géographie française de l'époque bien loin de la présentation « moniste » qu'en a faite Olivier Orain dans sa thèse. Ce dernier souligne en effet l'« environnementalisme impitoyable de De Martonne [qui] ne reconnaît aucune césure entre ce qui est de l'ordre du naturel et ce qui est de l'ordre de l'humain » (Orain 2003, 85). Sur ce point, la présentation d'Emmanuelle Boulineau insiste à l'inverse sur la prise en compte de facteurs multiples et différenciés par ce géographe :

> « E. de Martonne ne parle pas ici en géomorphologue attaché aux arguments de la géographie physique pour tracer des frontières dites naturelles. Il sous-tend son discours d'une argumentation fondée sur les identités régionales, les solidarités économiques et les intérêts stratégiques » (Boulineau 2001, 363).

Emmanuelle Boulineau reconnaît que la notion de région alors utilisée par De Martonne, n'a pas abouti à un renouvellement conceptuel profond. Il n'empêche que son approche lors

de cet épisode me semble difficilement compatible avec une conception moniste telle que décrite dans la thèse du plain-pied[284].

Olivier Orain agrège ensuite Jean Brunhes (1869-1930) à ce groupe des « réalistes », formé par Lucien Gallois et Emmanuel de Martonne. Quelques différences sont reconnues, mais elles sont réduites à peu de chose par Olivier Orain :

> « La frontière entre les uns et les autres, à ce niveau, est faible, car bien évidemment L. Gallois ou E. de Martonne n'auraient en rien contesté l'idée d'une éducation au regard, mais il y va d'une sorte de clause de style : tous partagent la priorité donnée à l' "étude exacte, précise, de ce qui est aujourd'hui", mais en modulant différemment la visibilité des taxinomies servant à "observer, […], grouper […], classer", que J. Bruhnes revendique, tandis que les deux autres les occultent » (Orain 2003, 42).

Sur ce sujet, l'analyse menée par Marie-Claire Robic (1988) de la thèse de Jean Brunhes offre une perspective bien différente. Toute l'hétérodoxie de sa thèse par rapport au « modèle » de la monographie régionale[285] est développée : approche thématique, exposé de méthode fourni, démarche comparative, perspective critique… Ces spécificités de la thèse de Jean Brunhes ont sûrement eu plus d'impact sur son rapport au réel que la présentation d'Olivier Orain le laisse entendre. Cette idée peut être soutenue en revenant sur l'analyse réalisée par Olivier Orain d'une citation de Jean Brunhes :

> « Cette accoutumance à voir les réalités où elles sont et telles qu'elles sont, produit sur l'esprit cet effet de lui inculquer une juste défiance des simples étiquettes et de lui procurer un sens critique de la valeur variable des réalités géographiques » (Brunhes 1925, 876).

Par rapport à cette citation, l'analyse réalisée dans le plain-pied se révèle, me semble-t-il, partielle et partiale. En effet, Olivier Orain ne se sert que de la première partie de la citation pour mettre en avant le réalisme de Jean Brunhes. Pour lui, cette citation « achève de mettre en équivalence géographie et vision des "réalités" en tant qu'elles sont localisées » (Orain 2003, 42). Toutefois, en prenant appui sur la seconde partie de la citation, il est possible de mettre plutôt en avant un constructivo-nominalisme, notamment avec les expressions : « une juste défiance des étiquettes » et « un sens critique de la valeur variable des réalités géographiques ».

---

[284] Sur ce sujet du « monisme d'E de Martonne », il peut être ajouté qu'Olivier Orain laisse lui-même entrevoir des doutes à l'encontre de cet argument dans l'évolution de son travail. En effet, il existe une évolution notable entre la version écrite pour la thèse: « Ce faisant, on retrouve l'ontologie et ce que nous avons pu appeler le "monisme" d'un De Martonne » (Orain 2003, 127) et la version remaniée pour le livre : « Ce faisant, on retrouve une forme de monisme que j'avais cru déceler chez E de Martonne » (Orain 2009, 122). Cette évolution, avec notamment l'emploi des formules « d'une forme de » et « que j'avais cru déceler », montre l'existence de doutes qui ne sont malheureusement pas explicités par Olivier Orain.

[285] « Modèle » incarné par la thèse d'Albert Demangeon sur la Picardie.

Ici, le poids du modèle en amont, avec la volonté de créer une unité « réaliste » pour nourrir l'idée d'un paradigme post-vidalien, explique sûrement ces choix effectués par Olivier Orain pour présenter Jean Brunhes. Certes, l'auteur du plain-pied est prêt à reconnaître un cas limite à l'application du modèle kuhnien (Camille Vallaux), mais sa thèse nécessite tout de même de présenter un socle « réaliste » pour la période post-vidalienne. Ce socle est construit par Olivier Orain en associant Jean Brunhes. *A contrario*, les réflexions les plus constructivo-nominalistes de Camille Vallaux sont mises en valeur dans la thèse du plain-pied jusqu'à souligner tout de même que cet auteur reste *in fine* un réaliste (Orain 2003, 61), ce qui parachève l'unité du groupe. Il y a là une construction assez habile qui intègre un cas limite tout en renforçant ce groupe « réaliste » créé.

Il faut ici souligner que mon propos ne vise pas à remettre en question profondément le caractère réaliste de la géographie post-vidalienne. Il s'agit seulement d'éclairer la construction d'Olivier Orain pour mieux identifier les limites d'une écriture kuhnéiforme de la géographie. En effet, il existe un jeu qui tente de concilier ce qui semble pourtant difficilement conciliable : une approche sous forme de spectre (de Lucien Gallois à Camille Vallaux en passant par Jean Brunhes), d'opposition (Lucien Gallois et Jean Brunhes, tous deux réalistes opposés à Camille Vallaux) et d'unité (tous réalistes).

Pour développer plus en avant les géographies post-vidaliennes, il est possible de s'appuyer sur un travail de Marie-Claire Robic (1991) intitulé *La stratégie épistémologique du mixte*.

2. Le paradigme du mixte

Marie-Claire Robic souligne que dans la géographie française classique, la description et l'explication, ont été les deux modes principaux de connaissance, associés dans « des configurations tant diverses que contradictoires » (Robic 1991, 55). À partir de ce constat, elle propose la notion de « paradigme du mixte », qu'elle revendique comme « d'une efficacité certaine, puisqu'il accorde à la géographie moderne une identité scientifique symbolique tout en permettant un jeu extrême entre ses deux limites, qui sont aussi les horizons, ou les pôles du champ épistémologique, entre l'empirisme et le constructivisme : la description pure et l'explication » (*Ibid* 1991, 57). Olivier Orain affirme dans sa thèse que le paradigme du mixte est « congruent » (Orain 2003, 45) avec le paradigme réaliste qu'il défend. Il est toutefois possible de mettre en perspective quelques objections.

En effet, la proposition de Marie-Claire Robic s'appuie sur deux pôles, dont un constructivisme. Ce point est occulté par Olivier Orain. Or, expliquer une situation, c'est construire un cadre interprétatif. Ce faisant, la proposition d'Olivier Orain réduit fortement

la géographie à son pôle empiriste. Il est intéressant à mon avis ici de rentrer dans le détail d'une de ces monographies considérées souvent comme la simple transcription d'un terrain pour comprendre que leurs constructions étaient bien plus complexes.

L'exemple choisi est celui de la thèse de Théodore Lefebvre soutenue en 1933 et dont le rapport de soutenance est analysé dans le détail par Nicolas Ginsburger (2014). Le directeur de cette thèse a été Albert Demangeon. Le jury a été présidé par Emmanuel De Martonne. Le fait que ce soient souvent les mêmes géographes qui se retrouvent à cette époque dans de nombreux jurys de thèse est plutôt un élément en faveur d'une interprétation kuhnienne, reposant en partie sur l'entre-soi d'un petit groupe. De surcroît, Nicolas Ginsburger détaille ensuite les dessous de la nomination de Théodore Lefebvre en montrant qu'elle repose moins sur la qualité scientifique de la thèse que sur un ensemble de tractations au sein d'un petit groupe. Cela peut s'inscrire tout à fait dans une perspective kuhnienne avec la reconnaissance du poids effectif des relations institutionnelles et personnelles dans les dynamiques scientifiques.

Le travail de Nicolas Ginsburger rend compte de tensions dans ce petit groupe notamment entre les « deux écoles de géographie dominantes de l'époque, celle de Paris et celle de Grenoble » (Ginsburger 2014, 9). Le rapport de soutenance possède plusieurs commentaires qui sont plutôt en accord avec la thèse d'Olivier Orain :

> « Partout le contact direct avec la réalité s'entrevoit et les nombreuses photographies, les excellents dessins de l'auteur en sont le très éloquent témoignage. […] On peut lui reprocher quelque abondance à cet égard ou quelque excès de précision : on y découvre une foi dans la documentation statistique que beaucoup ont perdue, placés en face de semblables problèmes, mais enfin tout est ici minutieux, tout s'efforce d'y être précis… » (Ginsburger 2014, 11).

Ce qui est reconnu et valorisé est donc « le contact direct avec la réalité » avec néanmoins des nuances bien marquées par les reproches de « quelque abondance » ou « quelque excès ».

Toutefois, ce qui est attendu n'est pas un rendu totalement complet. Le caractère trop analytique de la thèse de Théodore Lefebvre est critiqué. Le rapport de soutenance indique : « L'auteur plaide la nécessité de tenir compte de toutes les nuances ; mais la majorité du jury garde l'impression d'un morcellement excessif, parfois déconcertant » (Ginsburger 2014, 12). Albert Demangeon critique un plan systématique qui sépare des phénomènes liés comme les modes de vie pastoraux et agricoles. L'importance donnée à la géographie physique est critiquée par certains pour un sujet dont le titre *Les modes de vie dans les Pyrénées atlantiques orientales* appelle une importance de la géographie humaine. Enfin, si l'abondance et la qualité de production cartographique sont louées, le rapport de soutenance indique : « Peut-être [le candidat] s'est-il laissé entraîner un peu loin par une sorte d'ivresse

de cartographie, jusqu'à voir dans la carte un but au lieu d'un moyen d'expression, une image directe de la réalité au lieu d'un schéma… » (Ginsburger 2014, 11). On ne saurait mieux douter qu'une géographie post-vidalienne unanimement et complètement de plain-pied ait pu exister de façon univoque.

Cet exemple donne à penser qu'il est facile d'extraire quelques phrases d'un rapport de soutenance de thèse alors même que cette trace n'est pas toujours centrale dans une mémoire disciplinaire[286]. Cette objection me conduit à présenter une seconde analyse, celle d'une thèse d'un géographe beaucoup plus célèbre : Roger Dion. Le compte-rendu réalisé par Albert Demangeon (1934) revient sur toutes les qualités de sa thèse. Parmi celles-ci, la mise en relation de la géographie physique et humaine est soulignée avec une critique du plan à tiroirs. L'illustration est louée pour son abondance et son originalité, mais aussi parce qu'elle est « strictement aménagée pour la démonstration ». Enfin, Albert Demangeon conclut en saluant « un style sobre, lumineux, plein d'évocation et de couleurs, qui manifeste un talent d'écrire, devenu assez rare dans nos thèses de géographie » (*Ibid* 1934, 319). L'expression de ces qualités montre que, ce qui était développé sur le ton de la critique pour la thèse de Théodore Lefebvre, se retrouve ici et excède donc le propos d'un cas spécifique. De plus, ce compte-rendu de la thèse de Roger Dion est publié dans les *Annales de Géographie*, c'est-à-dire dans la revue dominante et par un des "patrons" de la géographie post-vidalienne.

Ces quelques références me conduisent à remettre en question le modèle de plain-pied promu par Olivier Orain, se caractérisant par une écriture « transparente » (Orain 2003, 46) et l'usage du plan à tiroirs comme modèle. Sur ce point, Olivier Orain, m'a oralement fait remarquer que le positionnement d'Albert Demangeon était assez ironique au vu de sa propre thèse plutôt à tiroir et de son écriture accordant une place minimale aux effets de style. Toutefois, il me semble très significatif que, ni Albert Demangeon, ni un autre chef de file de cette époque, n'a jamais fait officiellement la promotion d'un changement de modèle par rapport aux orientations vidaliennes, en valorisant une approche analytique et une écriture transparente. L'objectif et la difficulté résidaient bien davantage dans la réalisation de l'idéal vidalien.

Il est vrai que l'historiographie a parlé de thèse prototypique, mais il ne s'agissait pas d'exercice-type tel que le décrit Thomas Kuhn. Cette affirmation est bien confirmée par les

[286] Théodore Lefebvre obtient un poste de maître de conférences en 1933, puis de professeur titulaire en 1937 à l'université de Poitiers. Engagé dans la résistance, il est arrêté et déporté en Allemagne où il est décapité en 1943. Il a une place périphérique dans la mémoire disciplinaire, au sens où il est actuellement très peu connu des géographes français.

réflexions de Roger Dion : « Je ne parvins que très tard à la possession de ma propre langue. Durant une suite d'années, rédiger fut pour moi une opération si difficile et pénible que je mis trois fois plus de temps à écrire ma thèse que n'en eût mis un travailleur ordinaire » (Broc 1982, 206). Cette situation est très éloignée de celle d'une science « normale » où l'étudiant a déjà appris le langage à utiliser par la réalisation des exercices-types. Roger Dion montre qu'il est nécessaire de s'adapter à la spécificité de son cas d'étude et de trouver sa forme pour le rendre au mieux. La thèse du plain-pied développée par Olivier Orain me semble dès lors assez mal rendre compte de cette tension et de sa complexité.

Une critique qu'Olivier Orain m'a adressée lors de nos échanges concerne ma compréhension de ce paradigme du mixte. Je n'aurais, selon lui, retenu qu'un seul aspect : l'entre-deux entre pôle empirique et pôle constructiviste alors que Marie-Claire Robic a développé d'autres entre-deux : langue vernaculaire / langue savante et approche sensible / intelligible. Effectivement, il est vrai que c'est ce premier entre-deux qui avait alors surtout retenu mon attention. Le fait qu'il soit développé dans une partie qui conduit Marie-Claire Robic à distinguer la géographie de la sociologie durkheimienne « visant une vocation d'universalité » et de l'histoire « historisante » (Robic 1991, 59) a contribué à un rapprochement fort avec les développements passeroniens. Ainsi, quand Marie-Claire Robic conclut son article sur l'idée d'une réévaluation du paradigme du mixte « tant sur le plan des principes que sur celui des résultats » (*Ibid* 1991, 66), il me paraît particulièrement intéressant et légitime de proposer que les réflexions de Jean-Claude Passeron servent de base à cette réévaluation avec une extension de celle-ci jusqu'à la géographie contemporaine.

Je ne suis toutefois aucunement opposé à prendre en compte les autres entre-deux rappelés par Olivier Orain. Ces entre-deux peuvent être d'ailleurs rapprochés, en acceptant quelques modifications, de certains proposés par Jean-Claude Passeron. Le couple « langue vernaculaire / langue savante » peut être vu comme une variante du binôme « langue naturelle / langue artificielle ». Le couple « sensible / intelligible » peut être rapproché du binôme « approches herméneutiques / modélisantes ». Toutefois, ces propositions déplacent quelque peu les problématiques. En effet, par exemple, l'« herméneutique » vise la création de sens, mais ne renvoie pas forcément à une approche sensible. C'est pourquoi il m'a semblé plus fécond heuristiquement de m'attacher à l'entre-deux « empirique / constructiviste » et de proposer une réévaluation du « paradigme du mixte » au prisme des propositions de Jean-Claude Passeron. Les réflexions de Marie-Claire Robic n'apparaissent plus simplement alors comme la matrice d'une géographie révolue, mais plus globalement, comme les conditions de possibilité d'une production de connaissances en géographie et dans l'ensemble des SHS.

Pour revenir plus directement aux géographies post-vidaliennes, la section suivante aborde quelques éléments de leur diversité, tout d'abord, par le prisme de leur structuration générale, puis, par l'étude d'un cas spécifique.

3. Diversité des géographies post-vidaliennes

a. Structuration générale

Un élément qui a fortement contribué à structurer la géographie post-vidalienne – peu développé par Olivier Orain, mais bien étudié par Numa Broc (2001) – est l'opposition entre l'école de Paris (sous la houlette d'Emmanuel De Martonne) et l'école de Grenoble (dirigé par Raoul Blanchard). Précisons, tout d'abord, que cette concurrence de différentes écoles est dans le descriptif kuhnien plutôt rattachée soit à une phase pré-paradigmatique, soit à une phase de « crise ». Or, Emmanuel De Martonne (1873-1955) et Raoul Blanchard (1877-1965) font partie de la première génération de post-vidaliens avec, en suivant la perspective d'Olivier Orain, un paradigme déjà théoriquement installé.

Pour tenter de dépasser ce paradoxe, il est possible de s'appuyer sur l'attaque de Raoul Blanchard envers André Cholley : « La thèse de Cholley est un livre jeune qui sent encore l'école et la théorie. Devenu maître, assoupli au contact incessant de la réalité, M Cholley nous donnera des œuvres plus brèves, mais plus réellement denses et utilisables » (cité par Broc 2001, 98).

Cette critique peut être interprétée de deux manières différentes par rapport à la thèse d'Olivier Orain :

- La première interprétation insiste sur l'importance du réalisme dans l'école de Grenoble. Cette interprétation renforce la thèse d'Olivier Orain. En effet, l'approche théorique est sévèrement critiquée par Raoul Blanchard et c'est le « contact incessant avec la réalité » qui est valorisé.
- La seconde interprétation insiste *a contrario* sur la présence de la théorie dans l'école de Paris alors que le caractère de « plain-pied » de cette école est un argument majeur de la thèse d'Olivier Orain. D'autres critiques de Raoul Blanchard permettent d'approfondir ce point, notamment celle-ci : « La morphologie alpestre est plus complexe que veulent nous le faire croire ceux qui y découpent des constructions d'allure géométrique » (cité par Broc 2001, 100). C'est bien entendu ici l'école de Paris qui est visée et cette critique rappelle l'importance des grilles de lectures dans

cette école. L'image d'une géographie de « plain-pied » est alors bien difficile à soutenir.

Au-delà de la structuration via la dualité Grenoble / Paris, il faut souligner une diversité des positionnements. Par exemple, dans une nécrologie rédigée par Paul Pélissier à propos de Pierre Gourou (1900-1999), la situation suivante est explicitée :

> « Comment ne pas exprimer ici le regret que durant près un quart de siècle l'enseignement de Pierre Gourou à Paris soit resté aussi confidentiel, totalement ignoré des étudiants de l'Institut de géographie qui auraient eu pourtant que quelques centaines de mètres à franchir pour aller l'entendre. Encore eût-il fallu qu'ils en soient informés… » (Pélissier 2000, 213).

En effet, Pierre Gourou a enseigné au Collège de France de 1947 à 1970 dans une orientation très vidalienne, plutôt critique par rapport à une partie de ses contemporains post-vidaliens occupant l'Institut de Géographie. De la même manière, Maurice Le Lannou qui est professeur au Collège de France de 1969 à 1976 se réclame fortement de l'héritage vidalien. Les différentes universités (Clermont-Ferrand, Montpellier, Lille, Bordeaux…) produisent des « formes » (Hacking 2008, 248) variées qui ne sont pas forcément réductibles à chaque fois à une école, mais dépendent aussi fortement des individus. Par rapport à ce tableau, il est vrai que l'École de Paris avait un certain pouvoir, notamment en jouant sur une partie des carrières, mais il a existé une hétérogénéité avec des manières très différentes de faire fonctionner le projet vidalien suivant les générations et les individus.

In fine, cette situation explique finalement le nombre réellement important d'exceptions qu'Olivier Orain doit prendre en charge : Jules Sion, Jean Brunhes, Camille Vallaux, Jean Gottmann… À ces exceptions développées, il faut ajouter celle de Maximilien Sorre. Ce dernier réactive un axe de travail vidalien plus centré sur l'écologie et la biologie, montrant bien que les diverses voies ouvertes par Paul Vidal de la Blache n'ont pas été vraiment refermées par l'établissement d'un paradigme contrairement à ce que la thèse du plain-pied laisse entendre. On pourrait aussi citer une autre exception non abordée, par Olivier Orain : celle d'Eric Dardel (1899-1967) avec le développement d'une géographie phénoménologique. Certes, cette œuvre est restée longtemps marginale et comme les autres auteurs ici cités, elle n'a pas véritablement fait école. Toutefois, faire de tous ces auteurs des hapax qui auraient été marginalisés et effacés par un paradigme en place demeure problématique[287]. Mis à part le cas d'Eric Dardel qui a été « redécouvert » par les géographes, les autres auteurs n'ont jamais été occultés.

[287] Face à cette problématique, Michel Bruneau (2000) a individualisé la proposition kuhnienne en développant l'idée d'un « paradigme de Pierre Gourou ». De la même manière, les paradigmes de Maximilien Sorre, de Jules Sion et de beaucoup d'autres géographes, peuvent être affirmés. Il est entendu que cette solution ne

Globalement, il est nécessaire de bien différencier une lecture qui oppose, d'une part, un centre et quelques positionnements périphériques marginaux et, d'autre part, une conception où des tensions théoriques, méthodologiques et des jeux de pouvoir ne sont pas parvenus à imposer une univocité disciplinaire. L'auteur ayant le plus développé cette conception est Vincent Berdoulay (1981; 1988). Le travail considérable qu'ouvre en détail une telle voie de recherche dépasse largement le cadre de ma thèse. Il faudrait examiner le positionnement des différentes écoles (Paris, Grenoble, Clermont-Ferrand, Bordeaux…), mais aussi ceux des individus tout en prenant en compte leurs interactions. C'est la raison pour laquelle la section suivante s'attache à un positionnement singulier analysé par Olivier Orain, celui d'Henri Baulig (1877-1962).

b. Une réévaluation du positionnement d'Henri Baulig

Olivier Orain a développé plusieurs analyses autour des réflexions de ce géographe et s'en sert notamment pour mettre en avant le « schisme » entre deux cultures : une littéraire et une scientifique (Orain 2003, 103). Si ce terme de « schisme » peut être questionné et discuté, c'est un autre plan que j'ai choisi de développer, en m'appuyant sur une remarque d'Etienne Juillard (1962a) à propos d'Henri Baulig :

> « Sa méthode de travail était classique, mais pratiquée avec une conscience et une intelligence exceptionnelle. Elle reposait d'abord sur la lecture attentive, incisivement critique et ponctuellement tenue à jour des travaux des deux seules écoles qui comptaient alors, l'américaine et l'allemande. Armé d'hypothèses de travail, de points de comparaison, il se rendait ensuite sur le terrain, et tous ceux qui ont eu le privilège de l'y accompagner savent comme cet homme, que l'on a dépeint parfois comme un géographe de cabinet, savait observer. Puis c'était la réflexion, l'interprétation. Chaque hypothèse était passée au crible » (Juillard 1962a, 562).

L'utilisation de méthodes hypothetico-déductives est ici décrite avec précision. Cette description montre que la remise en cause du plain-pied qui a été esquissée précédemment à partir des critiques de Raoul Blanchard, n'est pas un cas isolé. Une différence notable est que contrairement à Raoul Blanchard, la théorie est ici loin d'être critiquée, mais apparaît comme un préalable.

Etienne Juillard précise en outre qu'Henri Baulig considère que « les processus mis en jeu par la nature sont trop nombreux et trop complexes pour que leur observation et leur mesure puissent conduire à l'explication intégrale de la morphogénèse » (Juillard 1962a, 562). La reconnaissance d'un « réel » jamais complètement atteint est évidemment contraire

---

correspond en rien au modèle kuhnien puisqu'elle fait fi de toute la dimension collective qui fonde originellement le concept de paradigme.

à la thèse du plain-pied. De plus, tandis que la volonté d'établir un répertoire de structures est affirmée, Henri Baulig reconnait aussi la nécessité d'une démarche plus idiographique. Ce positionnement n'est pas éloigné d'une approche passeronienne dont la section suivante explore quelques apports et limites dans son application aux géographies post-vidaliennes.

c. Apports et difficulté d'une lecture passeronienne des géographies post-vidaliennes

De manière très globale et schématique, les géographies post-vidaliennes peuvent être caractérisées, en mobilisant les modalités épistémologiques passeroniennes, par la progression limitée du pôle expérimental sans aucune remise en cause de la prééminence du pôle herméneutique. La difficulté principale d'une telle caractérisation réside dans la nécessité de préciser ce qui est entendu sous les expressions « pôle expérimental » et « pôle herméneutique ». Ces précisions doivent d'ailleurs être aussi déclinées en différenciant les différentes branches de la géographie, mais aussi à un niveau plus fin, en distinguant les positions des différents chercheurs. Comme précédemment (*cf.* section Chap9.II.3.a), ce travail considérable, s'il était réalisé en détail, dépasserait malheureusement le cadre de cette thèse. Mais pour ébaucher tout de même cette voie de recherche, il est possible de s'appuyer sur les propos de Bernard Grandjean (1957) dans une présentation intitulée *Coup d'œil sur l'école française de géographie au milieu du XX*^{ème} *siècle*.

La géographie physique y apparaît comme de plus en plus marquée par un effort de scientificisation passant par l'utilisation d'un vocabulaire plus précis, une spécialisation des travaux, un approfondissement des théories et l'introduction du travail en laboratoire. Le pôle expérimental est bien entendu moins développé dans la géographie humaine même si Bernard Grandjean souligne des efforts de systématisation en géographie urbaine, de nouvelles méthodes, des travaux comparatifs pour dégager des traits communs ainsi qu'un l'élargissement des enquêtes.

Ce pôle expérimental, pour la géographie humaine post-vidalienne, se caractérise globalement par la non-utilisation des modèles théoriques formalisés : les réactions des géographes français à la présentation par Walter Christaller de son modèle[288] au congrès de l'*Union Géographique Internationale* en 1938 témoignent en effet de positionnements hétérogènes. Par rapport à René Musset qui reconnaît pleinement l'intérêt et la pertinence

[288] Le modèle de Christaller, appelé la théorie des lieux centraux, essaye d'expliquer la hiérarchie des villes, selon leur taille, leur localisation et leurs fonctions. Il sera ensuite très largement mobilisé dans le cadre de l'analyse spatiale.

de l'approche de Christaller tout en regrettant sa dimension trop théorique, l'opposition d'André Gibert est par exemple beaucoup plus manifeste :

> « Les organismes où se déroule l'activité de l'homme n'ont point la régularité géométrique. La volonté, la fantaisie, la régularité géométrique d'un passé historique très divers, les diversités des conditions physiques donnent au paysage humanisé, une variété telle qu'il paraît vain de l'enfermer dans des lois. C'est ce qui donne à la géographie humaine son sens et son intérêt » (*Union Géographique Internationale* 1938, 292).

Faut-il voir, dans ce positionnement, la défense et l'illustration d'un paradigme dont aurait résulté l'ignorance en France de cette voie de recherche, et ce, pendant plusieurs décennies ? Une telle affirmation serait pertinente si l'approche modélisatrice s'était ultérieurement totalement imposée. Or, ce n'est pas le cas. En effet, si l'heuristique des modèles a été par la suite valorisée, leur utilisation est loin de s'être généralisée comme unique pratique disciplinaire et les critiques initialement formulées n'ont pas été complètement dépassées. De plus, pour qu'une lecture en termes de paradigmes soit vraiment pertinente, il faudrait que la géographie soit marquée par une succession de modèles incommensurables. Or, le fonctionnement de la géographie française n'est pas marqué par une telle dynamique.

Il est par conséquent pertinent de ne pas constituer cet épisode − la présentation par Christaller de son modèle et les réactions françaises − en anomalie de type kuhnienne. Olivier Orain la place d'ailleurs à un autre endroit.

III. L'anomalie

1. Le critère non déterminant de l'utilité sociale

Olivier Orain situe l'anomalie dans la rencontre de la géographie française avec la scène aménagementiste dans les années 1950-1960. Avant tout, il faut préciser que la géographie française avait rencontré des perspectives aménagementistes bien avant cette période (*cf.* section Chap9.I.1 sur la *bataille des Annales*). Toutefois, l'idée qu'il existe dans les années 1950-1960 une montée en puissance des problématiques d'aménagement interrogeant, voire entrant en tension, avec la géographie, peut largement être soutenue. Les articles de Pierre George et Etienne Juillard, précédemment cités (*cf.* section Chap7.II.5), en rendent parfaitement compte.

Un temps fort de cette tension s'incarne la manière dont Pierre George critique en 1955 le travail de thèse de Michel Philipponneau : *«* on n'a pas à faire d'application, c'est du travail de journaliste *»* et *«* le géographe n'a pas à prévoir l'avenir *»* (cité par Gaudin 2015,

115). Ces appréciations privent Michel Philipponneau de la mention « très honorable ». Les conséquences de cette situation sont pour ce jeune chercheur « de ne pas être immédiatement nommé maître de conférences à Rennes et de ne pas voir sa thèse analysée dans les *Annales de Géographie* » (Gaudin 2015, 115). Ici, il faut reconnaitre que ces points illustrent bien un des aspects mis en exergue par Thomas Kuhn avec la notion de paradigme, celui de la résistance au changement au sein d'une communauté scientifique. Toutefois, ce critère ne saurait suffire à lui seul.

Cette première tension liée au travail de thèse de Michel Philipponneau, s'est ensuite étendue dans une controverse plus globale entre la « géographie appliquée » (portée par Michel Philipponneau) et la « géographie active » (défendue par Pierre George). Cette controverse ne porte pas vraiment sur une « anomalie » au sens kuhnien du terme, mais bien plutôt sur la problématique de la dépendance (ou non) de la recherche en géographie à des donneurs d'ordre. Par rapport à cette situation, il est intéressant de rappeler un élément développé par Thomas Kuhn dans sa postface à *La structure des révolutions scientifiques* :

> « (Je pense maintenant que c'est une faiblesse de mon texte original d'avoir accordé si peu d'attention à des valeurs comme la cohérence interne et externe, quand j'ai étudié les sources de crise et les facteurs présidant au choix d'une théorie.) Il existe d'autres genres de valeurs – par exemple : la science devrait-elle ou ne devrait-elle pas avoir une utilité sociale ? - mais ce qui précède doit faire comprendre ce que j'entends » (Kuhn 2008, 252).

Il peut avoir débat sur cette dernière formule, mais l'interprétation la plus probable est que cette valeur de l'utilité sociale n'est pas déterminante pour Thomas Kuhn dans son schéma explicatif. Ce point de vue est en effet cohérent avec sa remarque sur le fait que, dans un paradigme tel qu'il l'a pensé, « les membres d'un groupe scientifique donné constituent les seuls spectateurs et les seuls juges du travail de ce groupe » (Kuhn 2008, 283). Or, la reconnaissance de l'utilité sociale de la science tend à élargir le groupe puisqu'une instance de légitimation peut être alors extérieure à ce dernier. Quand la géographie française approfondit ses liens avec la sphère aménagementiste, il peut donc être considéré qu'elle s'éloigne d'autant plus d'une perspective kuhnienne[289] qu'elle ne dépend plus d'un petit groupe « seul spectateur et juge de son travail » (pour reprendre les termes de Thomas Kuhn sus-cités).

Cela étant souligné, pour approfondir la notion d'« anomalie » dans une perspective critique, il n'est pas inutile de développer ici une distinction sémantique relative à ce terme.

[289] Dans le cas, bien entendu, où une perspective kuhnienne est envisagée.

2. Une distinction sémantique : entre la disparité et l'anormalité

Une distinction explicitée par Isabelle Lefort (2015) en s'appuyant sur les travaux de Georges Canguilhem (1972) est ici reprise :

- Dans une première interprétation, l'anomalie désigne « un fait isolé, une disparité, une rugosité » (Lefort 2015). Cette signification s'appuie sur le sens originel du terme grec *omalos* dont la traduction est : inégal, irrégulier, fait d'aspérité. Isabelle Lefort fait d'ailleurs remarquer que c'est dans ce sens originel que Vidal de la Blache utilise ce terme d'anomalie. Il renvoie à l'impossibilité d'établir une loi générale dans de nombreux domaines comme la distribution de la population : « Une foule d'anomalies nous avertissent que la répartition actuelle de l'espèce humaine est un fait provisoire, issu de causes complexes, toujours en mouvements » (Vidal de la Blache 2015, 35). Les anomalies sont par conséquent, dans cette signification, consubstantielles des géographies vidaliennes et post-vidaliennes attentives aux spécificités, aux exceptions et aux aspérités du monde.
- La seconde signification d'anomalie est plus contemporaine. Contrairement au sens originel, elle renvoie à l'idée d'une « anormalité ». L'élément qui forme l'anomalie n'est plus un simple fait isolé que le système explicatif en place ne parvient pas à rendre intelligible. Il remet en cause la loi ou la théorie existante.

Thomas Kuhn joue habilement sur ces deux sens d'« anomalie », car contrairement à une conception poppérienne, ce n'est pas l'anomalie, dans son sens d'anormalité qui provoque le changement. La conversion au nouveau paradigme commence pour un ensemble important de chercheurs avant ce stade. Elle relève d'une conversion qui n'est pas totalement rationnelle. Toutefois, l'anomalie grandit suffisamment au point de mettre en lumière des contradictions fortes entre les résultats attendus et ceux obtenus, ce qui relève alors de la seconde signification.

Quand Olivier Orain développe une perspective initialement esquissée par André Meynier (qualifiant la période 1939-1969 du « temps des craquements »), puis qu'il analyse le positionnement de « géographes en malaise » dans les années 1950 et 1960, il s'appuie exclusivement sur le premier sens d'« anomalie » dans une acceptation plutôt faible. Le passage à la deuxième signification qu'exploite Thomas Kuhn n'est pas vraiment convaincant dans le cas de la géographie française. Il n'existe pas en effet de contradiction par rapport aux résultats attendus. Cette situation explique pourquoi la présentation d'Olivier Orain prend finalement la forme d'une longue évolution inachevée. Cette dynamique est contraire au schéma kuhnien où « justement parce que c'est une transition entre deux

incommensurables, la transition entre deux paradigmes concurrents ne peut se faire par petites étapes » (Kuhn 2008, 207).

Dans un article paru en 1961, Pierre George, promoteur de la géographie active, écrit ainsi :

> « Les géographes français, qui doivent une part importante de leur formation aux études historiques, se sont montrés très soucieux de ne pas dépasser le domaine des certitudes et se sont refusés à toute attitude perspective, donc constructive, toute construction comportant une part de spéculation. Dans ces conditions, le dilemme précédemment évoqué s'est présenté sous une apparence d'inéluctabilité. Mais, dans la mesure où l'on admet de plus en plus qu'une étude de situation — au sens de conjoncture — comporte la définition de rapports de forces, et par voie de conséquence, de virtualités à l'égard desquelles peuvent se déterminer des choix et des actions, il n'y a aucune raison de dénier à l'ensemble de la recherche une portée utilitaire. Par là même tombe toute justification d'une sécession d'une géographie appliquée. » (George 1961, 338).

Cette reconnaissance montre assez qu'il n'existe pas derrière la controverse entre la géographie active et la géographie appliquée une anomalie grandissante comme dans le descriptif kuhnien.

Une des raisons pouvant être invoquée pour expliquer cette absence d'« anomalie », pris dans le sens d'une « anormalité », dans la géographie française, renvoie à l'espace non-poppérien de Jean-Claude Passeron. L'élasticité du langage naturel a permis par exemple au concept de « région » d'acquérir un nouveau sens sans qu'un renversement paradigmatique soit nécessaire. Il n'a pas existé de remise en question expérimentale équivalente aux lunes de Jupiter pour le système géocentrique. Ces exemples permettent de comprendre cette référence[290] à Karl Popper même s'il a existé de vives oppositions entre ce dernier et Thomas Kuhn. Le système explicatif de *La structure des révolutions scientifiques* n'implique certes pas qu'il y ait une falsification poppérienne, mais il faut tout de même qu'il y ait « l'impression que la nature, d'une manière ou d'une autre, contredit les résultats attendus dans le cadre du paradigme qui gouverne la science normale » (Kuhn 2008, 83). Or, ce n'est pas à ce niveau que se situent les critiques effectuées par la « nouvelle géographie » à l'encontre de la géographie post-vidalienne. En effet, ce ne sont pas les résultats explicatifs concernant la nature, ni même les systèmes sociaux, ni même les interactions nature/sociétés, qui ont été directement remis en cause par la « nouvelle géographie ».

Par rapport à un schéma kuhnien incarnant le dépassement d'un paradigme face à des « anomalies » à l'intérieur d'une discipline circonscrite, l'explication par un mouvement

[290] Référence initialement réalisée par Jean-Claude Passeron et ici reprise.

critique multiforme (provenant des recherches quantitatives anglo-saxonnes[291], du marxisme, de structuralisme, puis de la « crise de la représentation »[292]…) est plus convaincante. À cet égard, la perspective internaliste d'Olivier Orain est particulièrement problématique puisque les remises en causes se retrouvent dans d'autres sciences sociales et bien au-delà de la France.

Cette faiblesse démonstrative me conduit à préférer et à continuer le développement de l'analyse précédente fondée sur les deux pôles (expérimental et herméneutique) en interaction.

3. Modifications des pôles expérimental et herméneutique

Les réflexions de Michel Philipponneau, promoteur de la géographie appliquée, montrent des changements du pôle expérimental autant au niveau des perspectives poursuivies que de la temporalité :

> « Était-il réaliste de penser que les travaux de l'universitaire libre du choix et du rythme de ses recherches auraient une valeur si évidente que les politiques en attendraient les conclusions pour les appliquer ? Les utilisateurs ont besoin de réponses rapides, précises, leur permettant de choisir entre plusieurs hypothèses. Si les géographes ne répondent pas à ces besoins, d'autres spécialistes le feront à leur place, d'autres disciplines bénéficieront de moyens importants et tendront même à se substituer à la géographie dans la formation des jeunes ? » (Philipponneau 1996, 272).

Les changements ici explicités sont importants, car ils touchent à une partie des finalités, des méthodes et des formations géographiques. Il ne s'agit pas pour autant d'une anomalie, car ce changement ne remet pas radicalement en cause le savoir et la communauté géographique alors en place. Il s'agit plutôt d'un ajout qui rentre en tension avec une position alors défendue par le chef de file de la géographie de l'époque, Pierre George.

Il faut souligner que ces changements impactent également le pôle herméneutique en donnant un sens différent aux recherches. Pour Pierre George, la géographie appliquée produit un affaiblissement du pôle herméneutique avec des travaux dont le caractère utilitaire nuit à la compréhension profonde des régions ou des phénomènes étudiés. *A contrario*, pour Michel Philipponneau, il existe une valorisation continue et cohérente de l'expertise du

[291] Ce point est bien mis en avant par Denise Pumain et Marie-Claire Robic pour la géographie théorique et quantitative française : « la découverte d'un nouveau continent – l'état des recherches anglo-saxonnes – qui a catalysé les énergies, provoquant une sorte de défi intellectuel et social (le sentiment d'un retard inadmissible par rapport aux États-Unis) à relever » (Pumain et Robic 2002, 128).

[292] Plusieurs références peuvent être consultées sur ce sujet : (Foucault 1990; Bougnoux 2019;…).

géographe en passant de la réalisation d'analyses à la prospective, puis aux propositions d'actions, et enfin à l'action.

Au-delà de la problématique de la qualité de l'expertise en fonction de la dimension utilitariste (ou non) de la recherche, une partie du débat renvoie à la question de l'implication du géographe en tant qu'acteur. Il faut ici rappeler que Michel Philipponneau s'est engagé dans de nombreux organismes d'aménagement du territoire. Cette situation pose évidemment le problème de l'objectivité. La question du positionnement des géographes, qui ne peuvent jamais totalement s'extraire de la société qu'ils étudient, se pose alors avec encore plus d'intensité. Toutefois, cette géographie appliquée ne remet nullement en cause le réalisme. Le diagnostic participe au contraire, le plus souvent, d'un renforcement réaliste de la situation et, ce faisant, le moment aménagementiste ne peut être considéré comme une anomalie.

Cette remise en cause conduit à aborder les travaux de Jean Labasse (1918-2002) et de Pierre George (1909-2006) avec des perspectives différentes de celle d'Olivier Orain qui considère ces deux auteurs principalement par le prisme d'une pré-supposée résistance du paradigme post-vidalien.

IV. Les évolutions apportées par Jean Labasse et Pierre George

La thèse de Jean Labasse (2012) parue en 1955 sur le commerce et les capitaux dans la région lyonnaise marque une évolution qui n'est pas explicitée par Olivier Orain alors qu'elle touche la problématique du réalisme. En effet, cette thèse met en lumière et développe de façon considérable l'importance d'un facteur immatériel, les capitaux, dans la production de l'espace. Les réflexions de Jean-Bernard Racine permettent de préciser les prolongements de cette voie :

> « C'est dans les années 60-70 qu'a été réellement explicitée l'idée qu'œuvre humaine, l'espace n'est pas produit à partir des seules réalités matérielles, la base infrastructurelle de la société, mais également à travers l'intervention des idées, des images, des codes de comportement, des systèmes de valeur, de tout ce qui, bien que n'étant pas matériel, a autant de réalité : les représentations mentales, ce que l'on pourrait appeler, dans une certaine mesure, l'idéologie » (Racine 2006, 234).

Certes, Jean Labasse n'a pas développé tous ces facteurs, mais en insistant sur l'importance d'un facteur immatériel d'explication, il a contribué à une géographie dépassant la prise en compte des seuls éléments visibles. Les éléments mis en avant par Jean-Bernard Racine témoignent du développement du pôle herméneutique, mais ce processus a mis plusieurs décennies à se réaliser.

Du côté du pôle expérimental, il existe également une évolution avec des changements qui tout en restant limités sont loin d'être négligeables. Joël Pailhé, dans son analyse de la géographie de la population de Pierre George, met en avant l'importance d'un « objet abstrait, que "l'on ne peut voir avec les yeux", ni "toucher avec les mains" » (Pailhé 1972, 62). Il rend compte d'une « rupture épistémologique » (*Ibid* 1981, 28) dans les années 1950 et écrit dans *L'Espace géographique* :

> « Si à l'époque, de nombreux géographes français découvrant à la suite de R. Brunet l'importance capitale de la catégorie de discontinuité, on pouvait considérer que ce terme cédait à une certaine mode, on doit reconnaître, huit ans plus tard, qu'il demeure adéquat aux travaux de P. George sur la géographie de la population » (Pailhé 1981, 28).

Ici, la nouveauté conceptuelle de la thèse complémentaire *Les phénomènes de discontinuité en géographie* de Roger Brunet (1965), souvent avancée comme un des éléments majeurs à l'origine de la « nouvelle géographie », est recontextualisée et remise en perspective. Joël Pailhé fait ainsi état d'une montée de l'abstraction, autant au niveau des calculs que de la cartographie, antérieure à la « nouvelle géographie ». Par exemple, ce n'est plus l'espace visible qui est directement étudié, mais des objets statistiques comme le déséquilibre emploi/résidence (Pailhé 1972, 62).

Outre la géographie de la population, il est ici utile de souligner que Pierre George a renouvelé un nombre important de domaines : géographie urbaine, économique, industrielle… Son engagement marxiste s'est traduit dans son travail par une importation de quelques concepts afférents à ce mouvement : classe sociale, analyse des changements de forme de production… mais il est vrai que cette importation conceptuelle est restée limitée puisque Pierre George n'a pas développé de géographie marxiste *stricto sensu*. La formalisation théorique est restée également assez réduite, en comparaison de celles proposées ultérieurement par d'autres géographes (Roger Brunet avec le structuralisme, le groupe *Dupont* avec l'orientation théorico-mathématique…). Finalement, même si la personnalité de Pierre George n'a pas laissé dans la discipline l'image d'un auteur en rupture avec celle de son temps, il demeure celui dont les apports intellectuels ont marqué les années 1950-1960 « donnant ainsi une dimension nouvelle à l'héritage vidalien » (Pailhé 1981, 23).

Plusieurs études thématiques permettent également de relativiser l'idée d'une rupture forte entre Pierre George et les « nouveaux géographes » :

- Sylvie Daviet (2005) dans son étude sur 30 ans de géographie industrielle (1970-1999) dans les *Annales de Géographie*, affirme que : « L'idée, *a posteriori*, que cette géographie [la géographie industrielle des années 1970] serait "démodée" parce qu'elle se serait enfermée dans des limites sectorielles étanches et

inopérantes est largement excessive » (Daviet 2005, 87). Ce qui domine l'article est l'absence de révolution conceptuelle sur cette période 1970-1999. Ce qui est plutôt mis en avant par Sylvie Daviet, ce sont quelques innovations comme celle de Jean-Paul Diry en 1979 avec la notion de « système », mais le tableau d'ensemble est loin de présenter une bascule conceptuelle et sémantique.

- Une seconde étude thématique, celle issue de la thèse de Serge Bourgeat (2007) sur la géographie urbaine pendant la période 1960-1984, contredit également l'idée d'une révolution kuhnienne et fait état plutôt d'une longue évolution.

Pour revenir à l'argumentaire d'Olivier Orain, ce dernier développe, après ce temps des « résistances » (incarné, selon lui par J. Labasse et P. George), l'idée d'une rupture historiquement appelée la « crise de la géographie ».

V. La « crise »

Rappelons d'abord que le vocable de la « crise » avait déjà été utilisé par Fernand Braudel en 1951 quand il écrit dans *Les Annales* : « Nul ne contestera qu'il n'y ait actuellement (en France spécialement, mais ailleurs aussi) une crise grave de la géographie dite "humaine" » (Braudel 1951, 485). Il est ensuite réutilisé par Pierre George, avec la publication d'un *Que sais-je ?* sur les méthodes de la géographie en 1970. Toutefois, il est vrai qu'il ne se diffuse de manière importante qu'au cours des décennies 1970-1980.

La multiplication des voies de recherches au début des années 1970, bien délimitées par plusieurs revues (*L'Espace Géographique, Hérodote, Espace-Temps*), peut sembler à première vue bien correspondre au descriptif kuhnien. Avant d'examiner cela en détail, il faut tout d'abord rappeler que le terme de « crise » est polysémique. Un travail plus en profondeur sur sa sémantique permet de bien différencier la « crise de la géographie » d'une crise kuhnienne.

1. Distinction sémantique : entre « décision » et « incertitude »

Edgar Morin précise l'étymologie du terme « crise » :

> « À l'origine, *Krisis* signifie décision : c'est le moment décisif, dans l'évolution d'un processus incertain, qui permet le diagnostic[293]. Aujourd'hui, crise signifie

[293] Par opposition à *Krasis*, qui signifiait la confusion.

> indécision : c'est le moment où, en même temps qu'une perturbation, surgissent les incertitudes » (Morin 1976, 135).

La « crise de la géographie », telle qu'elle est présentée par Olivier Orain dans sa thèse[294], est proche du sens originel de *Krisis*. En effet, elle est marquée par des diagnostics et des décisions de développer de nouvelles voies (R. Brunet, Y. Lacoste…). Le temps est alors à l'effervescence et aux affirmations plutôt qu'aux indécisions et aux incertitudes. *A contrario*, la crise kuhnienne me semble plus marquée par ce sens contemporain de crise lié aux incertitudes : tout d'abord, par rapport aux résultats inexpliqués par l'ancien paradigme ; ensuite, par rapport au choix de la théorie qui formera le nouveau paradigme parmi les multiples propositions existantes.

Dans la « crise de la géographie », il est certain qu'il a existé des débats vifs sur de nombreux sujets, mais les voies de recherche proposées n'ont pas visé à résoudre une anomalie précise et ont pu co-exister sur le long terme. Cette différence explique en grande partie pourquoi le descriptif kuhnien dans ses détails s'applique *in fine* mal à la géographie française. Par exemple, quand Thomas Kuhn affirme : « la recherche durant la crise ressemble beaucoup à celle de la période antérieure au paradigme, à ceci près que le foyer de divergence est à la fois plus petit et plus clairement défini » (Kuhn 2008, 123), il est évident que la « crise de la géographie » ne s'est pas réalisée sur ces modalités. De plus, les trois classes de problèmes de la science normale développées par Thomas Kuhn − « la détermination des faits significatifs ; la concordance des faits et de la théorie ; l'élaboration de la théorie » (Kuhn 2008, 59) − se retrouvent peu au niveau général où Olivier Orain situe le changement de paradigme. En effet, la « nouvelle géographie » et les autres voies de recherche (*Hérodote*, *Espaces-Temps*…) n'ont pas attaqué la géographie classique sur un défaut de concordance théorie/fait. Il n'y a pas eu de focalisation sur un problème, mais au contraire une multiplication des sujets et des approches, avec un foyer de divergence qui s'est agrandi et qui ne s'est jamais vraiment résolu.

Une difficulté importante par rapport à la « crise de la géographie » est de savoir ce qui relève de discours, d'affirmations identitaires par rapport à ce qui a été effectivement des évolutions ou des ruptures. Sylvain Cuyala souligne la dimension performative du terme « révolution » tout en précisant que « la plupart de ceux qui utilisent cette tournure n'explique pas réellement en quoi l'apparition de la géographie théorique et quantitative constitue une révolution » (Cuyala 2014, 88).

---

[294] La précision est ici importante, car Olivier Orain a retravaillé abondamment suite à sa thèse les sujets de « 1968 » et de la « crise » (2015; 2021). Cette partie est centrée sur la thèse du plain-pied et ne prend donc pas en charge ses travaux ultérieurs sur la question.

Cette remarque pose le problème difficile de l'évaluation de la crise. Au niveau du pôle expérimental, il existe un net enrichissement avec l'utilisation de nouvelles méthodes : AFC, modèle… L'emploi revendiqué des mathématiques au-delà de la géographie physique et de la statistique descriptive a créé une évolution notable par rapport à la géographie antérieure. Toutefois, par rapport au prisme de la révolution, il faut souligner que ces nouvelles méthodes n'ont pas remplacé les anciennes. Il a en effet pu être pensé que le développement de ce pôle expérimental allait reléguer les pratiques plus discursives, moins formellement construites, à des usages dépassés[295], mais cette croyance ne s'est jamais complètement réalisée.

Au niveau du pôle herméneutique, il y a eu des développements importants autour de la notion d'espace vécu, de la réflexion sur l'implication du chercheur sur le terrain, ou encore des avancées sur l'outillage des cartes mentales. Une partie de ces développements ont eu lieu en parallèle de la forte affirmation du pôle expérimental.

En approfondissant le contenu des discours de « crise », une majorité s'appuie sur la dichotomie approches inductives / déductives qu'il s'agit donc d'analyser plus en détail.

2. Réévaluation de l'approche duale : inductif / déductif

Au niveau d'une évaluation globale de la « crise », la construction de la géographie comme une science déductive fondée sur une axiomatique unifiée, est un échec. Toutefois, il existe une grande différence entre les géographes qui reconnaissent que ces deux grandes catégories, inductif et déductif, ne fonctionnent pas vraiment, mais continuent tout de même de les utiliser et ceux, moins nombreux, qui *a contrario* franchissent le pas et revendiquent un mode de scientificité différent.

Il est vrai que l'usage de problématique et l'explicitation des cadres théoriques sont certes devenus courants, mais la géographie est loin de se construire sur un fonctionnement de type déductif. Les bricolages théoriques sont nombreux pour s'adapter au mieux aux situations empiriques rencontrées et/ou aux présupposés scientifiques. Face à ces affirmations,

---

[295] La conclusion d'un texte d'André Fel intitulé *Deux géographies humaines ?* écrit dans le premier numéro de *L'Espace Géographique* marque bien cet espoir : « Deux géographies humaines : l'une sensible à la variété des faits, imprégnée d'histoire, se voulant compréhensive à l'égard des cultures régionales, l'autre qui cherche sa place dans les sciences humaines du présent, cherchant à expliquer les systèmes de transformation du monde et toute chargée de langage mathématique. C'est dans cette deuxième tendance que le philosophe E. Granger voit une des marques les plus vigoureuses de notre temps : le rationalisme appliqué. Mais il ne manque pas d'ajouter qu'il ne s'agit encore que des "débris d'une science future" » (Fel 1972, 112).

précisons tout d'abord qu'il n'est pas question ici de nier les efforts de conceptualisation réalisés par plusieurs chercheurs (Philippe Pinchemel, Roger Brunet, Franck Auriac…). Ensuite, il n'est pas non plus question de disqualifier les recherches basées sur des modèles, qu'ils soient très formalisés ou non. Si d'un point de vue local, des modèles peuvent faire preuve d'une certaine efficacité, ils n'ont pas contribué à des administrations de la preuve vis-à-vis des grandes théories du social. Aucune ne s'est imposée comme un nouveau paradigme et les deux grandes catégories, « inductif » et « déductif », dans leurs versions épurées, correspondent mal à la plus grande partie des travaux de recherche menés en géographie.

Les réflexions de Jean-Claude Passeron permettent de mettre en avant une modalité différente de fonctionnement des SHS :

> « Ces sciences ont en commun une autre manière de prouver : faire converger des preuves de forme logique différente dans un argumentaire d'ensemble, leurs arguments dans un langage de l'interprétation, leurs interprétations dans une théorie plausible » (Passeron 2001, 75).

Une théorie peut être bien entendu utilisée en amont sans remettre en cause cette forme explicitée d'un montage. D'ailleurs, la thèse d'Olivier Orain illustre bien ce point qui est reconnu à la fin du chapitre sur la « crise » :

> « Face à l'objection terminologique, l'ensemble du travail engagé dans ce chapitre se voudrait un dossier probatoire, constitué de textes, d'analyses socio-politiques, de témoignages. » (Orain 2003, 293).

Il s'agit en effet d'un « dossier », forme que Paul Vidal de la Blache avait lui-même commencé à mettre en avant avec divers éléments (cartes, textes…) qui s'articulent et se répondent. Même s'il est possible d'objecter qu'Olivier Orain a davantage explicité et questionné sa grille d'analyse, ce qui les relie, est au fond cette forme qui consiste à faire converger des éléments multiples dans un argumentaire d'ensemble. Ma thèse relève aussi bien entendu de cette forme. Il est ici intéressant de souligner que le travail quantitatif précédemment mené pourrait avoir l'apparence non pas d'une déductivité, mais d'une « nécessité » plus forte, tout en ne relevant pas d'un mode de scientificité représentatif des sciences expérimentales. S'il existe, bien entendu, des branches disciplinaires dans les SHS, comme la psychologie expérimentale, qui sont plus directement marquées par ce mode de scientificité, il me semble que ce dernier est très loin d'être représentatif de la géographie contemporaine.

Ce point ne sera pas ici plus développé, car ce n'est pas ce que revendique Olivier Orain dans sa thèse du plain-pied. Sa ligne de rupture repose sur une autre dualité, celle entre

réalisme et nominalisme. Cette perspective conduit à examiner la résolution, ou plutôt la semi-résolution, car le nouveau paradigme est reconnu par Olivier Orain comme incomplet.

VI. La semi-résolution

1. Une proposition anti-kuhnienne ?

La semi-résolution proposée dans la thèse du plain-pied sort complètement du cadre kuhnien tel qu'il est défini dans *La structure des révolutions scientifiques*. En effet, Thomas Kuhn y précise : « Rejeter un paradigme sans lui en substituer simultanément un autre, c'est rejeter la science elle-même » (Kuhn 2008, 117). Or, c'est bien ce que propose Olivier Orain.

Il est sûrement possible sur ce point de faire référence à des positions plus souples adoptées par Thomas Kuhn à la fin de sa vie dans lesquelles les révolutions n'aboutissent pas à un changement de paradigme, mais seulement à l'émergence d'une nouvelle spécialité. Dans ce cas, il faut souligner qu'une partie de la force du système explicatif de *La structure des révolutions scientifiques* est perdue, notamment en ce qui concerne l'incommensurabilité. Ceci pose la question de jusqu'où un chercheur est prêt à adapter un modèle pour le faire correspondre à une situation empirique… Cette question est d'autant plus légitime quand les adaptations réalisées ne résolvent pas les difficultés. Quand le modèle apporte plus de complexités que d'éclairages, la question de son remplacement se doit d'être envisagée. Cette affirmation est sans aucun doute mon point de divergence le plus important avec la thèse du plain-pied.

En considérant donc qu'une semi-résolution peut exister, Olivier Orain développe deux études de cas − celle de Franck Auriac (1935-2017) et de Claude Raffestin (né en 1936) − pour illustrer le pré-supposé nouveau paradigme (incomplet) qui prend la forme d'un nominalo-constructivisme en géographie

2. Franck Auriac et Claude Raffestin

Pour Franck Auriac, il s'agit d'une étude minutieuse de sa thèse montrant en fin de compte plutôt une ambiguïté qu'un nominalo-constructivisme. En effet, Olivier Orain reconnaît qu'il n'est pas possible de trancher entre deux positions : le système viticole languedocien, tel qu'il est décrit par Franck Auriac, existe-t-il en tant que tel ou est-il seulement une grille d'analyse ? Olivier Orain tend à donner du poids à la seconde hypothèse, mais un texte écrit par ailleurs par Franck Auriac peut légitimement réhabiliter la première position. En effet, Franck Auriac déclare en reprenant une citation de Roger

Brunet : « "Il n'y a pas de science sans quelque positivité" : dans positivisme c'est l' "isme" qui est de trop : je fais mienne cette assertion » (Auriac 2003, 9). Ce géographe exprime conjointement l'importance qu'il attache au matérialisme historique, c'est-à-dire à une conception de l'histoire dans laquelle les évènements ne sont pas déterminés par des idées, mais par des rapports sociaux et l'évolution des moyens de production. Il devient dès lors difficile de considérer que le vignoble languedocien, tel que décrit dans sa thèse, ne soit pour Franck Auriac qu'une grille de lecture sans connexion avec les rapports sociaux existants et l'évolution effective des moyens de production.

Cette ambiguïté est bien au cœur de la géographie contemporaine. En ayant l'objectif de fournir des éléments de compréhension du monde, les constructions, même si elles incluent des approches théoriques, peuvent très difficilement en être détachées. Cette étude de cas illustre de manière paradoxale pour la thèse d'Olivier Orain un inachèvement fondamental du constructivisme dans la géographie. Une situation développée par Markus Gabriel (2014) illustre concrètement le fond du problème : si deux personnes regardent par exemple le Vésuve de deux endroits différents, il n'existe pour un constructiviste radical que le Vésuve vu par la première personne et le Vésuve vu par la deuxième personne, mais il n'existe pas de volcan Vésuve. Une telle conception n'est pas sans poser de problème pour un grand nombre de géographes. Si le constructivisme a progressé sous des versions plus affaiblies, il n'y a pas eu de renversement gestaltique pour reprendre l'expression employée par Thomas Kuhn : il existe bien plutôt une diversité de positionnements entre réalisme et constructivisme.

L'étude de cas sur menée par Olivier Orain sur Claude Raffestin est également complexe, car les positions de cet auteur ont varié dans le temps. Un premier point fondamental qu'Olivier Orain développe peu, car il est contraire à sa thèse, se doit d'être souligné : Claude Raffestin revendique la recherche « de correspondances bi-univoques » (Racine et Raffestin 1978, 186) entre les concepts et leurs définitions. Rappelons ici l'affirmation de François Rastier à ce sujet : « l'idéal de correspondance terme à terme entre un mot, un concept et un objet a été durci par le positivisme logique » (Rastier 1995). Même s'il n'est guère loisible de considérer Claude Raffestin comme un néo-positiviste – beaucoup d'autres aspects de sa pensée bien développés par Olivier Orain empêchent une telle réduction – cette « bi-univocité » entre les mots et les choses définit pourtant un positionnement réaliste.

De plus, quand Claude Raffestin mobilise les réflexions d'Hilary Putnam, reprenant explicitement la distinction entre un « réalisme interne » et « un réalisme externe », il s'agit avant tout de deux formes de réalisme. S'il existe de la part de Claude Raffestin un rejet évident des externalistes, « c'est-à-dire des gens qui choisissent d'accoler un concept sur une

chose plutôt que de construire un système conceptuel et ensuite de retourner à la réalité » (Raffestin et Elissalde 1997, 90), il reconnait, en même temps, la persistance d'un « externalisme tragique » des géographes :

> « Toute notre géographie fonctionne comme cela parce qu'elle a été marquée par cette idée du "terrain". C'est pour cela que nous sommes des externalistes tragiques et c'est probablement un des verrous de la géographie » (*Ibid* 1997, 90).

Le nominalo-constructivisme est donc loin d'être reconnu comme une conception s'étant imposée partout dans la géographie française.

Enfin, Olivier Orain développe très peu[296] un texte de Claude Raffestin (1986), intitulé *Territorialité : concept ou Paradigme de la géographie sociale ?*, où il affirme pourtant la nécessité de prendre en compte un nouveau paradigme basé sur la territorialité. Par rapport à l'analyse d'Olivier Orain tendant plutôt à dévaloriser le concept flou de territoire par rapport à celui d'espace, il existe là une différence d'appréciation non négligeable. Claude Raffestin, présenté par Olivier Orain comme le « doxeur de la nouvelle épistémè » (2003, 315) est donc paradoxalement celui-là même qui souligne le maintien d'une perspective réaliste en géographie et affirme la fin du paradigme spatialiste.

Pour finir, deux points, un de désaccord et un d'interrogation, méritent à mon sens d'être explicités par rapport aux affirmations d'Olivier Orain sur cette phase d'avènement du nominalo-constructivisme. Ces deux points renvoient tous deux *in fine* à une reconsidération plus grande des jeux de pouvoir.

3. Des enjeux politiques cachés

Le premier point concerne un désaccord par rapport à la mise en avant par Olivier Orain de l'« abandon progressif du couperet du "ce n'est pas de la géographie" » (Orain 2003, 315). S'il est évident que la « nouvelle géographie » a contribué à un élargissement des objets d'étude, mes expériences personnelles me conduisent à contester cette idée d'un abandon progressif de ce couperet. Pour plusieurs projets, j'ai été en effet encouragé à développer en amont une perspective plus socio-spatiale afin d'éviter ce couperet. C'est pourquoi ce dernier n'a fait à mon avis que se déplacer. Il sert encore à discréditer des projets et s'inscrit dans des jeux politiques que les recherches de thèse d'Olivier Orain conduisent en grande partie à minorer (à l'exception de l'importance de mai 1968).

[296] Les limites temporelles du corpus de thèse d'Olivier Orain (1910 - milieu des années 1980) peuvent être invoquées pour tenter d'expliquer ce faible développement. Il faut toutefois souligner qu'Olivier Orain n'hésite pas par ailleurs à mobiliser dans sa thèse des textes à propos de Claude Raffestin postérieurs aux années 1990.

Par exemple, l'approche de Gilles Massardier (1996) concernant Roger Brunet rend beaucoup mieux compte de la position de force acquise par cet acteur, notamment suite à l'arrivée des socialistes au pouvoir à partir de 1981. Ce point est certes évoqué par Olivier Orain[297], mais de façon plus elliptique. L'importance qu'il donne à une lecture internaliste du schéma kuhnien explique cette situation. Il est alors possible de formuler une interrogation centrale à laquelle il est toutefois difficile de répondre : l'utilisation du schéma kuhnien dans une perspective internaliste n'est-elle pas une façon commode de transformer une domination institutionnelle en une domination cognitive rappelant les liens indissociables entre pouvoir et savoir (Foucault 1990) ? Dans le cas de la thèse d'Olivier Orain, une complexité peut être soulignée, car ce n'est pas la domination cognitive de la « nouvelle géographie » qui est finalement mise en avant. Elle n'est qu'une étape, selon Oliver Orain, vers l'avènement du nominalo-constructivisme. Toutefois, en étant à l'origine de la « révolution scientifique », son rôle historique s'en retrouve fortement valorisé.

Ce positionnement est d'autant plus discutable qu'il a été réalisé dans un contexte de montée en puissance des pensées post-modernistes[298] qui critiquent précisément l'attachement du modernisme au réalisme. La thèse d'Olivier Orain peut permettre de penser le spatialisme comme étant à l'origine d'une telle dynamique avec un prolongement jusqu'au tournant spatial[299] (Besse *et al.* 2017). Cette perspective pose tout de même un problème important, car les références mises en avant par le post-modernisme renvoient bien plus à Henri Lefebvre, Michel Foucault et aux post-structuralistes français (Deleuze, Derrida…) qu'à Thomas Kuhn. De plus, le travail concret, mené précédemment dans ma thèse pour rendre ma recherche aussi constructible que possible (*cf.* Part1 et 2), montre que le passage vers un constructivisme fort et un nominalisme n'a rien d'évident. Construire des problématiques plus précises, utiliser des outils mathématiques plus avancés, développer des réflexions critiques et réflexives par rapport à ses résultats, peut conduire, certes, à atténuer des formes de réalisme, mais ne pousse pas forcément à sortir du réalisme.

Mon travail de doctorat développe alors une autre voie que celle proposée par le post-modernisme en réactualisant les réflexions de Jean-Claude Passeron. Ce choix ne résulte pas

---

[297] « En revanche, il apparaissait intéressant de signaler la *convergence* qui s'est opérée à partir de 1980 entre un entrepreneur de projets géographiques, "capitaine d'industrie de la géographie française" (Lévy, 1996), doté d'une grande fécondité et un mouvement révolutionnaire, riche d'une expérience et d'une réflexion accumulées, mais longtemps cantonné à la littérature grise » (Orain 2003, 287). S'agit-il vraiment d'une convergence ou plutôt d'une évolution d'un mouvement qui s'est institutionnalisé ?

[298] La sortie de la thèse du plain-pied en 2003 est concomitante du débat ayant eu lieu sur le post-modernisme en géographie dans l'*Espace Géographique* (Collectif 2004)

[299] Le tournant spatial désigne une prise en compte renouvelée dans les SHS des phénomènes spatiaux qui pour Edward Soja (1989) trouve son origine dans la philosophie française des années 1960-1970 (Michel Foucault et Henri Lefebvre) et se développe dans les années 1980 (Anthony Giddens, Frederic Jameson…). L'expression s'est largement diffusée à partir des années 2000.

d'un état des lieux approfondi, ni d'une critique documentée du post-modernisme. Toutefois, il me semble être au cœur d'une difficulté du post-modernisme pouvant être abordée par l'intermédiaire d'une conclusion de la thèse de Yann Calbérac :

> « Plus que "l'image juste" (c'est-à-dire un traitement complet et définitif d'un corpus) je préfère mobiliser "juste l'image", c'est-à-dire la possibilité de sans-cesse déplacer la focale, de croiser les points de vue, de faire varier les angles, de tester des modèles analytiques parfois concurrents » (Calbérac 2010, 345).

De mon point de vue, il s'agit d'une pirouette rhétorique fort commode, car, au-delà de la définition donnée par Yann Calbérac de l'expression « juste l'image », qui s'accorde bien à son travail, il faut préciser que sa thèse est loin d'être juste une image. En effet, dans ce cas, toute une partie de sa construction serait difficilement explicable : pourquoi avoir interrogé des géographes de différentes générations, de différentes branches de la géographie, des deux sexes, si ce n'est pas pour obtenir une forme de représentativité (certes faible et non construite statistiquement, mais organisée et réfléchie) ? Pourquoi avoir construit toute une réflexion articulée à des exemples si ce n'est pas pour donner du poids à un discours scientifique ?

Même si Yann Calbérac rejette une lecture de l'histoire de la géographie en termes de paradigme, ce chiasme (« image juste » / « juste l'image ») est une rémanence d'une lecture épistémologique basée sur l'opposition. Il sert à mettre en avant, non pas le spatialisme comme chez Olivier Orain, mais une approche plus « culturelle » de la géographie. La prise en compte des réflexions de Jean-Claude Passeron conduit donc à une remise en cause bien plus radicale de la thèse d'Olivier Orain que celle effectuée par Yann Calbérac : au-delà du rejet du cadre kuhnien, il s'agit d'interroger plus profondément cette lecture épistémologique basée sur l'opposition autant dans ses appuis (couple inductif / déductif et réalisme / constructivisme) que dans ses finalités.

Le chapitre suivant poursuit cette analyse critique à partir d'une nouvelle entrée, celle des quatre éléments composant les paradigmes pour Thomas Kuhn.

Chapitre 10 :
Une lecture par les éléments de définition des paradigmes

Pour rappel, les quatre éléments composant un paradigme, développés successivement par Thomas Kuhn dans sa postface de *La structure des révolutions scientifiques* sont les suivants :

- Les généralisations symboliques
- La partie métaphysique
- Les valeurs
- Les exercices-types

L'ordre privilégié dans ce chapitre – les exercices-types, les valeurs, la métaphysique et les généralisations symboliques – est inverse. Il correspond à celui utilisé par Olivier Orain dans sa thèse que ma méthode (le commentaire de texte) m'a conduit à privilégier. Comme il est toujours préférable de lire le texte originel avant ses commentaires, il est conseillé avant la lecture de ce chapitre de se référer au passage de la page 120 à 129 de la thèse du plain-pied accessible à partir du lien suivant : https://theses.hal.science/tel-00082408/file/Orain_these.pdf

Il faut souligner enfin qu'aux quatre éléments, reconnus et spécifiés par Thomas Kuhn, Olivier Orain en a ajouté un cinquième – l'appareillage – qu'il a toutefois fort peu développé. Ce statut périphérique m'a conduit à ne pas l'analyser, car son traitement aurait été très déséquilibré par rapport à ceux des autres éléments.

I. Les exercices-types

1. Un écart au schéma kuhnien minoré

Tout d'abord, Olivier Orain multiplie les propositions correspondant selon lui à des exercices-types dans la géographie post-vidalienne : la leçon, le commentaire de carte, le bloc-diagramme… Il différencie ensuite les expériences de "terrain" des autres exercices académiques. La présentation initiale qu'il fait des exercices académiques comme visant « à reproduire dans un contexte *dénaturé* […] les expériences pleines et entières du géographe » (Orain 2003, 121) est fort discutable. En effet, une bonne partie des exercices académiques (leçon et commentaires de cartes) ne vise pas à reproduire un exercice de terrain, ni encore moins « les expériences pleines et entières des géographes ». Cette présentation tend à valoriser fortement les expériences de terrain dans les dispositifs de recherche post-vidaliens. S'il n'est pas question ici de remettre en cause cette importance, il faut rappeler que les post-vidaliens ne se contentaient pas du "terrain". La bibliographie et les archives étaient aussi utilisées et constituaient des sources majeures.

Ensuite, Olivier Orain introduit une distinction qui n'a jamais été développée par Thomas Kuhn entre « exercices-types » et « exemples ». Cette proposition constitue un changement fondamental, car les « exemples » dans la définition qu'en donne Olivier Orain, peuvent ne pas être soumis directement à la répétition (à la différence de ce qui fonde les « exercices-types » chez Thomas Kuhn). De cette manière, les grandes thèses régionales peuvent être qualifiées d'« exemple ». Ainsi, le fait que les exercices-types, présentés par Thomas Kuhn comme du *puzzle solving*, n'existent pas ou très peu dans la géographie post-vidalienne (à l'exception peut-être de la géomorphologie) est quasiment occulté. En effet, ce problème n'est mentionné par Olivier Orain que dans sa section sur la métaphysique en note de bas de page : « si tant est que *to solve a puzzle* soit une expression pertinente et congruente pour décrire les procédures de recherche que se donnent les géographes » (*Ibid* 2009, 121). La formule « si tant est » introduit ici un euphémisme volontairement utilisé par Olivier Orain pour montrer qu'il n'ignore pas que l'expression « to solve a puzzle », centrale dans les exercices-types kuhniens, n'est pas vraiment adaptée pour décrire les procédures de recherche dans les géographies post-vidaliennes et contemporaines. Toutefois, cette reconnaissance se fait à la marge (en note de bas de page) dans avec une formulation indirecte, qui évite de montrer l'écart existant avec le modèle kuhnien.

Les propos, précédemment cités de Roger Dion sur la difficulté éprouvée à écrire sa thèse (*cf.* section Chap9.II.2), illustrent bien la difficulté d'une discipline qui ne s'est jamais réduite au *puzzle-solving* de la science normale kuhnienne. Le fait de vouloir rendre au mieux la spécificité de chaque cas a empêché, sur le fond, l'établissement d'une normalisation importante. Dans le cursus de formation, il existe certes des exercices qui peuvent ressembler à une partie du descriptif kuhnien. Ainsi, l'étudiant apprend bien à reconnaître des formes définies à partir de cartes. Comme l'écrit Thomas Kuhn, il assimile « une manière de voir autorisée par le groupe et éprouvée par le temps » (2008, 258). Pour autant, il y a une différence fondamentale entre l'exercice de commentaire de carte et l'exercice-type kuhnien, au sens où le problème dépasse celui d'une "résolution de puzzle". Il y a bien des astuces à connaitre pour réaliser, par exemple, l'exercice de commentaire de cartes, mais il existe des situations spécifiques qui ne peuvent être résolues par la seule logique et qui font appel à ce qui est valorisé sous l'expression de « culture géographique ».

Cette dernière était déjà fortement encouragée par les post-vidaliens et elle continue de l'être dans les formations contemporaines. Cette remarque conduit à une autre problématique importante concernant les exercices-types, celle de la question de leur maintien dans le temps. En effet, dans une lecture kuhnienne, les exercices-types liés à un paradigme sont censés disparaitre en même temps que ce dernier. Or, ce qui est observé en géographie est

plutôt un maintien, avec certes des évolutions, mais sans transformation profonde de certains exercices emblématiques.

2. Non-disparition de certains exercices emblématiques

Deux exercices majeurs cités par Olivier Orain, la dissertation et le commentaire de carte topographique, sont des exercices scolaires encore très utilisés dans le cursus des géographes. La construction de modèles est loin de s'être généralisée et d'avoir remplacé ces exercices historiques. Cette situation, qui s'explique d'évidence mal en utilisant le schéma kuhnien, est d'ailleurs minimisée par Olivier Orain. Ce dernier met en avant une évolution au début des années 1990 avec l'introduction d'autres documents en précisant : « Les concours d'entrée aux ENS ont prolongé encore quelques années de plus cette prévalence du support cartographique » (Orain 2003, 121). Il faudrait ici préciser que le commentaire de carte a été prolongé pour plusieurs épreuves de concours de recrutement, bien davantage, de quelques décennies, sans n'avoir été jamais vraiment profondément remis en cause.

Un compte-rendu réalisé par Karine Bennafla et Claude Kergomard d'une épreuve de commentaire de documents géographiques réalisé en 2005 pour l'entrée à l'ENS permet de comprendre la raison principale de ce maintien. D'après eux, l'objectif de cette épreuve est de « détecter de futurs géographes capables d'associer l'analyse de faits concrets à la réflexion conceptuelle sur les espaces » (Bennafla et Kergomard 2005, 1). Ainsi, les « faits concrets » n'ont pas disparu du discours des géographes. L'objectif est plutôt celui d'une articulation de ces faits avec une problématique et quelques concepts. Aux explications valorisant la reproduction sociale et scolaire, il est nécessaire d'ajouter que les efforts de modélisations des « nouveaux géographes » et de leurs successeurs n'ont pas effacé la prise en compte du particulier et du concret.

Dans le même temps, il est vrai que quelques exercices cités par Olivier Orain connaissent une moindre utilisation actuelle que durant la période post-vidalienne : lecture de carte géologique, bloc-diagramme, excursion… Il est nécessaire de distinguer plusieurs causes, fort hétérogènes, à ces évolutions : plus faible poids de la géographie physique, mais aussi réduction des crédits, multiplication des supports accessibles pour créer des études de cas… Pour certains exercices, comme la sortie de terrain, il peut exister des changements de nature non négligeables. Par exemple, si les objectifs des sorties de terrain contemporaines peuvent être multiples, il ne s'agit plus comme dans les grandes excursions interuniversitaires post-vidaliennes de dresser des synthèses régionales (Calbérac 2010, 275). Toutefois, mes sorties de terrain, en tant qu'étudiant, me conduisent à privilégier l'hypothèse plus d'évolutions que

d'une révolution. En effet, le réalisme peut être variable suivant les enseignants et les situations[300], mais il est loin d'avoir disparu.

Au-delà de ce maintien d'un réalisme dans la géographie contemporaine, la réflexion sur les exercices-types m'a conduit à expliciter une orientation épistémologique dont il me semble important de montrer les implications et les limites.

3. Une conception contemporaine proche des exercices-types

L'analyse suivante présente une approche qui renvoie métaphoriquement à une démarche d'exercices-types. Ce choix provient du fait qu'il n'existe pas à ma connaissance de géographe revendiquant une conception disciplinaire directement appuyée sur des exercices-types kuhniens. Toutefois, quand Patrick Poncet (2017) utilise la métaphore du *Rubik's Cube* à la fin de son ouvrage intitulé *Intelligence spatiale*, il est très proche d'une conception basée sur la résolution de puzzle. Pour cet auteur, il existe six sciences du social (anthropologie, économie, médiologie, géographie, science politique et sociologie) qui peuvent être assimilées aux six faces d'un *Rubik's Cube* traditionnel. « On peut alors se représenter le travail de l'intelligence spatiale comme la résolution d'un *Rubik's Cube* "mélangé". Le désordre initial étant assimilable à la manière dont un "client" se représente son problème » (Poncet 2017). Le géographe, disposant d'une culture dans les sciences du social dépassant sa seule discipline, est alors chargé de résoudre le *Rubik's Cube*. De plus, Patrick Poncet insiste sur les temps records que les meilleurs réalisent, détail qui contribue à accentuer la vision mécaniste du processus de résolution. Ce dernier est réduit à un ensemble de permutations qu'il suffit d'effectuer pour trouver la bonne combinaison.

Sans remettre en question le fait que le géographe puisse apporter une expertise utile sur de nombreux problèmes, la métaphore du *Rubik's Cube* s'oppose à l'approche passeronienne qui met en avant l'imperfection des rapprochements de cas et le caractère non universalisable des intelligibilités produites. Ma formation et mes recherches me conduisent à être plus convaincu par la validité des réflexions passeroniennes que par la métaphore du *Rubik's Cube*. Plusieurs arguments peuvent être facilement mis en avant : contrairement à ce célèbre casse-tête, il existe, tout d'abord, dans le monde, nombre de situations sans solution parfaite. Ensuite, tout problème n'est pas réductible à un ensemble de schèmes limités, sur lequel le chercheur retomberait à chaque fois. Enfin, le chercheur est lui-même une partie du *Rubik's*

[300] Il est évident que suivant les objectifs d'un cours, les discours sont plus ou moins réalistes. Il est toutefois difficile d'imaginer une sortie de terrain « non-réaliste ». Il existe une diversité de situation où la fabrique pédagogique met en avant de façon variable des descriptions, des constructions ou des déconstructions de ce qui est observé.

Cube, ce qui change complètement la donne. Nul doute que certains objecteront que la métaphore ne doit pas être interprétée dans un sens aussi littéral. Il convient alors de définir la portée de l'analogie et nul doute que le problème dépasse alors celui d'une situation de *Rubik's Cube*.

Globalement, il est possible de mettre en avant des différences de valeurs entre des chercheurs croyants à des solutions de type *Rubik's Cube* et d'autres, plus sceptiques, voire en opposition avec cette conception du travail géographique. Ce passage par les valeurs est commode, car il permet de faire une transition avec la section suivante. Si ma formation du côté de l'ingénierie, ma culture des sciences du social et ma connaissance du mécanisme des *Rubik's Cube*, me conduisent à penser que le problème est bien en lien avec des valeurs, il convient, comme pour les exercices-types, de revenir aux précisions développées originellement par Thomas Kuhn et aux arguments d'Olivier Orain.

II. Les valeurs

1. La mise de côté des valeurs de prédiction

Alors que Thomas Kuhn précise que dans son système explicatif, ce sont les valeurs de prédiction qui sont les plus importantes, et qu'au sein de ces dernières, ce sont les prédictions quantitatives qui sont les plus cruciales, Olivier Orain ne mentionne pas ces précisions. Ceci se comprend aisément puisque la géographie post-vidalienne n'est, ni dans la prédiction, ni dans la quantification. Toutefois, cette situation est particulièrement problématique : dans l'examen d'un modèle, la réflexion sur ce qui est communément appelé les « écarts au modèle » est fondamentale. De plus, dans le cas présent, les différents éléments détaillés par Thomas Kuhn – valeurs de prédiction, exercices-types et généralisations symboliques – font système. Une formule comme celle de la chute des corps ($d = ½ gxt^2$, où d représente la distance parcourue, t la durée de la chute, et g correspond à l'accélération gravitationnelle exercée par la Terre sur tous les objets à sa surface et vaut 9,81 m/s^2 à 45° de latitude) permet de faire des prévisions et ouvre la porte à un ensemble d'exercices-types qui se résolvent par l'utilisation de la logique mathématique. En ce sens, la géographie post-vidalienne est de manière systémique anti-kuhnienne.

Pour développer sa perspective, Olivier Orain, ne pouvant s'appuyer sur les valeurs de prédiction, a privilégié une autre voie : celle des aphorismes de la géographie post-vidalienne.

2. Analyse à partir des aphorismes

Les aphorismes sont des petites phrases qui résument un conseil, une méthode ou un savoir. Par exemple, « on peut dire qu'on prouve le mouvement en marchant, et la réalité de la géographie en en faisant… » (Beaujeu-Garnier 1971). Il faut souligner que, dans le cas de la géographie post-vidalienne, ces aphorismes ont été largement repris par les critiques de cette même géographie pour mettre précisément en valeur ses faiblesses théoriques et épistémologiques. Le fait qu'une grande partie des aphorismes de la géographie post-vidalienne valorisent fortement la pratique de terrain est utilisé pour en déduire le rejet de toute approche théorique. Or, le poids de la géographie physique, avec notamment le rôle majeur de la théorie davisienne (Giusti 2004), montre qu'il n'y avait pas en soi de rejet de la théorie. C'est ainsi que des analyses fines, comme celles de Denis Wolff (2013) à propos des travaux d'Albert Demangeon, mettent en avant que la connaissance en amont de la théorie davisienne sur les cycles d'érosion constitue même un élément crucial pour comprendre l'élaboration de la géographie produite par ce géographe post-vidalien. Cela remet en cause totalement l'image classique d'observations réalisées « *de visu* sans théorie » (Wolff 2013, 7).

Il est vrai, cependant, que les approches expressément théoriques étaient peu développées dans la géographie humaine post-vidalienne et qu'elles n'ont été fortement promues qu'après la « crise de la géographie ». Il n'est pas question ici de dénier que derrière ces revendications se référant originellement, plutôt au terrain ou plutôt à la théorie, il y a des valeurs qui diffèrent et que cette différence a provoqué de fortes tensions. Toutefois, en rentrant dans le détail, il faut souligner qu'il existe une complexité mal rendue par une lecture en termes de changement de paradigme. Par exemple, quand Roger Brunet revient sur sa formation, il met en avant plusieurs valeurs : « la découverte de la géomorphologie et de l'explication de carte topographique fut une sorte d'émerveillement, par les vertus des méthodes d'investigation, du raisonnement logique et du dessin réunis, du moins tels que François Taillefer les faisait vivre » (Brunet 2003, 14). Ces caractéristiques dans les manières de faire de la géographie classique se retrouvent au cœur de la « nouvelle géographie ». Cette situation montre que ce n'est pas toute la géographie post-vidalienne qui a été critiquée et remise en cause, mais seulement une partie. Olivier Orain développe peu cet aspect discordant par rapport au schéma kuhnien.

3. Une continuité des valeurs ?

Il me semble important de ne pas tomber dans l'excès inverse en affirmant par opposition une continuité des valeurs. Un article de Jean-Bernard Racine (2006), notoire acteur de la « nouvelle géographie », témoigne d'une continuité par rapport à la géographie classique sur les valeurs « scientifiques » et « humanistes », et ce, tout en mettant en avant des discontinuités, notamment un « radicalisme réflexif » et une « pluralité d'approches » (Racine 2006). À ce niveau, il faut reconnaitre que l'élément des « valeurs », une fois détourné de son sens le plus important pour Thomas Kuhn – les valeurs de prédiction –, revêt une pertinence pour lire et comprendre la « crise de la géographie ». Il a existé des confrontations de valeurs entre, d'un côté, des géographies post-vidaliennes qui ont évolué de façons diverses, mais qui n'ont globalement pas remis en cause l'héritage vidalien et, de l'autre, des géographies des années 1970, qui se sont affirmées, en utilisant des mathématiques un peu plus avancées (AFC, Bayes, modèle…), en s'inspirant du structuralisme ou/et en se politisant[301] (avec l'importance de la référence marxiste).

Toutefois, ces changements d'horizons, plus ou moins profonds, peuvent tout à fait être lus et compris sans référence au schéma kuhnien. En effet, une lecture en termes de conflit de valeurs entre différents acteurs scientifiques est intrinsèquement réalisable. Un tel changement permet de s'affranchir de la problématique du modèle sous-jacent des sciences dures et d'une incommensurabilité difficilement justifiable.

III. La métaphysique

Par rapport à la section précédente sur les « valeurs », il faut souligner que l'exclusion des prédictions quantitatives – l'élément le moins métaphysique – aboutit à une différenciation difficile avec la « métaphysique », alors que ces deux caractéristiques sont bien distinctes dans la définition kuhnienne du paradigme.

1. La différenciation « métaphysique » / « valeurs »

Conscient de cette difficulté, Olivier Orain propose la différenciation suivante : les valeurs sont explicites[302] alors que la métaphysique reste « implicite voire "invisible" » (Orain 2003, 126). Ce critère est peu convaincant par rapport au schéma kuhnien originel.

[301] Ces divers éléments ne sont pas vraiment en correspondance avec la lecture internaliste du changement qu'essaye de mettre en avant Olivier Orain.

[302] Et même, parfois, « ressassées » (Orain 2003, 126).

En effet, dans la section sur la métaphysique, les croyances mises en avant par Thomas Kuhn (comme « la chaleur est l'énergie cinétique des parties constituantes des corps »[303]) ne sont pas forcément invisibles et implicites. Elles constituent plutôt des postulats. Inversement, certaines valeurs (comme la cohésion interne ou externe) ne sont pas forcément explicitées.

Cette proposition d'Olivier Orain peut se comprendre comme un choix astucieux pour renforcer l'importance du réalisme dans la géographie post-vidalienne. En effet, la thèse du plain-pied repose sur l'idée d'un réalisme explicité en partie par certains post-vidaliens (Lucien Gallois, Emmanuel de Martonne…), mais aussi avec une part restée implicite, à l'état latent. Différencier « valeurs » et « métaphysique » selon la modalité « explicite » / « implicite » permet de mettre en valeur cette idée que le réalisme exprimé n'est que la face émergée de l'iceberg. La thèse d'Olivier Orain peut être alors vue comme le révélateur de la face immergée, directement incarnée par la « métaphysique » du paradigme, ce qui contribue à renforcer en retour la pertinence d'une lecture kuhnienne.

Par rapport à cette construction, il est intéressant de faire un parallèle avec une autre analyse réalisée par Olivier Orain (2007a) en dehors de sa thèse. Cette analyse s'attaque à la notion de « possibilisme », notion souvent utilisée pour qualifier la conception des vidaliens et des post-vidaliens par rapport à leur approche des relations hommes/milieux. Cette conception a été parfois résumée par la formule : « la nature propose, l'homme dispose ». Selon Olivier Orain, le possibilisme est un « descripteur *a posteriori* de l'attitude des vidaliens, au pouvoir explicatif faible, et ce d'autant plus qu'il n'y a pas eu de posture stable, y compris chez nombre de représentants fameux de cette "école" » (Orain 2007a). À partir de cette affirmation, Olivier Orain enchaîne : « Ce faisant, le rabâchage de *la vulgate du possibilisme* a pour effet d'impatroniser une représentation simpliste et pour partie erronée de ce que pensaient "les classiques" » (*Ibid* 2007a). Ce que reproche *in fine* Olivier Orain au possibilisme est une réduction de la complexité des régimes de causalité.

Si je rends compte et développe ici cette critique, c'est parce qu'elle est très proche de celle adressée par mon travail au « réalisme ». Il y a là un terme subsumant, qui a été développé principalement *a posteriori* et qui fait courir le risque d'une « représentation simpliste et pour partie erronée » (pour reprendre les termes d'Olivier Orain à propos du possibilisme). Par rapport à ce parallèle, une objection possible consiste dès lors à contester que le réalisme soit un descripteur *a posteriori*. Toutefois, cette remarque, si elle est prise au sérieux, interroge par là même les analyses d'Olivier Orain sur le possibilisme. En effet, ce concept a été explicité préalablement par Lucien Febvre (1922) et n'est donc pas complètement *a posteriori*. De plus, le réalisme mis en avant par Olivier Orain pose un

[303] *Cf.* Bilibliographie : (Kuhn 2008, 250).

problème bien plus étendu que celui du possibilisme car le cadre kuhnien tend à cliver les positions pour être au plus proche d'une description en termes de paradigmes incommensurables[304]. De ce fait, ce n'est pas seulement la représentation de la géographie post-vidalienne qui est impactée, mais aussi les analyses de la « nouvelle » géographie.

Par rapport à ce parallèle, je dois préciser que les étiquettes de « représentation simpliste » et de « positions clivées » (pour reprendre la terminologie précédente) s'appliquent mal au travail complexe d'Olivier Orain. Ce qui est en jeu est bien plutôt, à mon sens, le maintien (ou non) d'une lecture paradigmatique qui prend pour caution scientifique cette thèse du plain-pied sans que toute la complexité de l'application du modèle kuhnien soit vraiment explicitée. Pour une partie, cette complexité est développée par Olivier Orain, mais il existe également de nombreux éléments, en particulier au niveau des limites de l'application du modèle, qui sont peu développés, voire occultés. Le travail approfondi et détaillé, ici engagé, vise à mettre en lumière ces éléments.

Dans cette optique et pour poursuivre sur la métaphysique, il faut rappeler que dans le schéma kuhnien, cet élément renvoie à l'adhésion des scientifiques à des modèles ontologiques ou heuristiques. Thomas Kuhn fait référence aux métaphores et aux analogies qui contribuent « à déterminer ce qui sera accepté comme une explication et comme une solution d'énigme » (Kuhn 2008, 251). Or, plusieurs épistémologues ayant montré l'importance des métaphores organicistes dans la géographie classique (Bachimon 1979; Berdoulay 1982), Olivier Orain est conduit à aborder ce sujet problématique dans sa thèse.

2. Le problème de l'organicisme dans la thèse du plain-pied

Le principal problème de l'organicisme par rapport à la thèse d'Olivier Orain est qu'il existe une continuité de l'emploi de ces métaphores entre Paul Vidal de la Blache et les post-vidaliens. Ce point est évidemment contradictoire avec l'approche kuhnienne qui postule une discontinuité entre le fondateur de la géographie classique française et ses successeurs. Par rapport à cette remarque, il est possible d'arguer qu'un même thème métaphorique, comme celui de l'organicisme, peut être utilisé à des fins différentes. Or, le principal objectif de ce thème métaphorique, celui de valoriser la nécessité d'une approche synthétique dépassant les études analytiques (géologie, climatologie, botanique,

[304] Le caractère a posteriori de l'étiquette « onctructiviste » mériterait aussi d'être étudié en détail car le terme est peu utilisé par les « nouveaux géographes ». De plus, il existe sur ce sujet une spécificité du travail d'Olivier Orain : ce terme et toutes ses formes dérivées (constructivismes, constructiviste…) sont par exemple absents de la thèse de Sylvain Cuyala (2014) pourtant consacrée à la diffusion de la géographie théorique et quantitative européenne francophone.

économie…), n'a pas été globalement remis en cause par les post-vidaliens (même s'il existe une difficulté majeure concernant sa réalisation). La métaphore organiciste est par conséquent un élément qui soutient plutôt la perspective d'une continuité intellectuelle entre vidaliens et post-vidaliens.

Dans cette optique, Olivier Orain a eu tout intérêt à minorer l'importance des métaphores organicistes dans la géographie post-vidalienne. Son positionnement consiste alors à affirmer que ces métaphores n'ont pas le statut de modèle heuristique par rapport au sens que Thomas Kuhn donne à cette expression[305]. Les raisons avancées sont multiples : les métaphores organicistes n'ont pas « débouché sur des opérations directes de justification » (Orain 2003, 126) ; elles n'ont pas fait « office de problème à résoudre » (*Ibid* 2003, 126) ; enfin, « les géographes classiques n'en ont pas tiré de gain explicatif (c'était plutôt une clause *a priori*) » (*Ibid* 2009, 122). Mais, ces arguments n'empêchent en rien l'organicisme de faire théoriquement partie de la métaphysique. Il s'agit par conséquent d'arguments spécifiques à la façon dont Olivier Orain, lui-même, conçoit la métaphysique et les modèles heuristiques. Ces changements de sens (« valeur », « métaphysique », « modèle heuristique », « exercice-type »…) posent la question de la pertinence de conserver une référence au modèle originel alors même que les complexités engendrées s'avèrent nombreuses.

La section suivante aborde une de ces complexités en détail. Il s'agit du glissement[306] opéré par Olivier Orain vers la notion d'« opérateur épistémologique » qui n'existe pas originellement chez Thomas Kuhn.

3. Une analyse des « opérateurs épistémologiques » d'Olivier Orain

Cette notion d' « opérateur épistémologique » créée par Olivier Orain n'est pas vraiment définie. Dans un premier temps, une série de couples (« induction/idiographie », « global/local », « particulier/général ») est mentionnée. Selon lui, ces couples sont conçus par le paradigme classique « comme complémentaires » alors qu'ils ont « la caractéristique d'être *a priori* dialectiques (voir antinomiques) » (Orain 2003, 127). Pour appuyer cette

[305] Les modèles heuristiques ne sont pas strictement définis chez Thomas Kuhn. Deux exemples sont cités : « le circuit électrique peut être considéré comme un circuit hydrodynamique en état d'équilibre ; les molécules de gaz se comportent comme de petites boules de billard élastiques » (Kuhn 1986, 250). Ces deux exemples conduisent à penser que la dimension heuristique repose avant tout sur leur caractère imagé.

[306] Ce glissement est effectué de la manière suivante dans la thèse du plain-pied : « À un niveau second et par contraste, je serais tenté de traduire les "principes" (explicités entre autres par E. de Martonne dans son *Traité de géographie physique*) comme des "*modèles*" ou plus exactement des *opérateurs*, non pas "ontologiques", mais plutôt épistémologique compte tenu de leur fonction explicative récurrente pour la géographie classique » (Orain 2003, 126).

affirmation, Olivier Orain réitère l'argument du monisme d'Emmanuel de Martonne. Si j'ai déjà critiqué cet argument (*cf.* section Chap9.II.1), son usage pose ici d'autant plus question qu'il sert à appuyer une affirmation générale sur l'ensemble de la géographie classique.

Par rapport à ces couples (« induction/idiographie », « global/local », « particulier/général »), l'argument ensuite avancé par Olivier Orain mérite d'être détaillé :

> « De surcroît, ces couples complémentaires ont un avantage paradigmatique évident : ils permettent une interprétation assez souple des finalités de la recherche géographique : décrire puis expliquer puis classer (procédure inductive), ou décrire une "combinaison" en utilisant un répertoire de schèmes géomorphologiques ou morpho-fonctionnels préalables (procédure idiographique) » (Orain 2003, 127).

Si le couple ici mis en avant est « inductif / idiographique », il est possible de s'interroger sur ce choix, notamment par rapport à une autre alternative, celle du couple « inductif / déductif », et cela pour deux raisons :

- La première est que les autres couples sont construits sur une opposition manifeste (« global / local » et « particulier / général »). Il y a donc sur le couple « inductif / idiographique » une exception qui n'est pas justifiée.
- La seconde raison provient du fait que l'expression « procédure idiographique » peut être remplacée dans la citation précédente par « procédure déductive ». En effet, « décrire une "combinaison" en utilisant un répertoire de schèmes géomorphologiques ou morpho-fonctionnels préalables » peut aussi être considérés comme une démarche déductive dans un sens faible[307].

Dans ce choix opéré d'utiliser « idiographique » plutôt que « déductif », le modèle du plain-pied a sûrement joué en amont un grand rôle pour éviter de nommer un type de démarches scientifiques rentrant mal dans le cadre de ce modèle. De plus, par rapport à ces couples, l'idée qu'ils permettent « une interprétation assez souple des finalités du chercheur » est finalement un argument plutôt défavorable à la thèse d'Olivier Orain. En effet, cette souplesse permet une pluralité d'approches qui ne sont pas toutes marquées par une inductivité radicale.

Pour finir sa section sur la métaphysique, Olivier Orain met aussi en avant d'autres opérateurs épistémologiques : les principes de causalité, de complexité et d'exhaustivité

[307] La notion de « déduction » n'est pas utilisée ici pas dans son acception logique pouvant être définie comme un type de raisonnement qui conduit d'une ou de plusieurs proposition(s) à une ou plusieurs conclusion(s) nécessaire(s) si l'ensemble des règles est accepté. Le fait ici de reconnaitre qu'il s'agit ici d'un usage de « déduction » au sens faible ne discrédite en rien cette proposition, car, parallèlement, la notion d'induction est elle-même utilisée par Olivier Orain dans un sens faible : le terme « classer » dans la citation précédente sous-entend en effet que l'aboutissement des démarches dites inductives est plus l'établissement de typologies qu'une loi générale (à la différence d'une induction au sens fort du terme).

avant d'arriver à la conclusion d'un « noyau dur » de la discipline. L'analyse suivante étudie et interroge cette nouvelle proposition complexe d'Olivier Orain. En effet, la notion de « noyau dur », inexistante originellement chez Thomas Kuhn, renvoie aux propositions épistémologiques d'un autre penseur : Irme Lakatos.

4. La reprise de l'épistémologie d'Irme Lakatos

Olivier Orain développe cette référence à Irme Lakatos dans une note de bas de page dans laquelle il précise que « le modèle des "noyaux durs" a été énoncé contre le relativisme supposé de *La structure des révolutions scientifiques* » (Orain 2003, 127), sans donner plus de détails. Cette explicitation minimale est problématique : Olivier Orain se sert finalement d'une proposition épistémologique concurrente à celle de Thomas Kuhn pour valider l'utilisation du modèle kuhnien. Cette situation paradoxale mérite quelques éclaircissements.

L'expression de « noyau dur » renvoie à la proposition épistémologique d'Irme Lakatos (1994) basée sur la notion de « programme de recherche ». À la différence des paradigmes, les programmes de recherches ne sont pas incommensurables. Ils peuvent co-exister. Le noyau dur d'un programme de recherche est sa base théorique admise comme provisoirement irréfutable par une communauté. Il faut reconnaitre qu'il existe une pertinence à rapprocher ce concept lakatossien de « noyau dur » de la notion kuhnienne de « métaphysique ». En effet, ces deux catégories renvoient à l'idée de postulat. De surcroît, par rapport à la lecture internaliste d'Olivier Orain, cette approche d'Irme Lakatos en termes de « noyau dur » est plutôt cohérente avec une focale davantage centrée sur les facteurs scientifiques que socio-politiques. Enfin, il faut souligner que la proposition d'Irme Lakatos a le mérite d'éviter les problèmes de l'incommensurabilité et d'une discontinuité absolue difficilement soutenable pour la géographie.

Toutefois, les concepts de programme de recherche et de noyau dur sont également difficilement applicables aux géographies post-vidalienne. Olivier Orain le reconnaît dans sa thèse, là encore sous la forme d'un euphémisme : « peut-être que l'idée même de "programme de recherche" est-elle trop précise pour rendre compte de ce qui donne une unité au style épistémologique de l'"école française de géographie" » (Orain 2003, 35). En prenant appui sur le travail effectué par Jean-Michel Berthelot sur de possibles programmes de recherche dans les sciences sociales (*cf.* section Chap11.III), il est certain que ces concepts de programme de recherche et de noyau dur s'appliquent mal aux géographies post-vidaliennes.

Si Olivier Orain utilise tout de même ces concepts dans sa section sur la métaphysique, c'est qu'ils sont favorables à sa thèse qui a besoin d'une solidification[308] du positionnement épistémologique post-vidalien. Il est bien plus facile de s'imaginer un paradigme en partant des idées de programme de recherche et de noyau dur, plutôt qu'en mettant en avant les libertés permises par l'œuvre vidalienne et la diversité des géographies post-vidaliennes. La thèse du plain-pied est prise dans cette tension. Pour ne pas fragiliser la démonstration, la reconnaissance que les concepts de programme de recherche et de noyau dur correspondent à une proposition épistémologique concurrente à celle de Thomas Kuhn et qu'ils sont peu adaptés pour lire la géographie classique, n'est effectuée qu'à la marge.

Concernant ces points − les libertés permises par l'œuvre vidalienne et la diversité des géographies post-vidaliennes −, le développement du quatrième élément, avancé par Thomas Kuhn pour définir un paradigme, permet de mieux comprendre une de leurs raisons.

IV. Les généralisations symboliques

1. Un élément absent de la géographie post-vidalienne

Les généralisations symboliques sont sans aucun doute l'élément le plus problématique dans le transfert du schéma kuhnien à la géographie post-vidalienne. En effet, les exemples donnés par Thomas Kuhn, formel (du type « f= ma ») ou verbal (du type « l'action est égale à la réaction »), n'existent pas ou sont très marginaux dans la géographie de cette époque.

Face à cette situation, Olivier Orain propose tout d'abord que les grandes notions de la discipline soient considérées comme des généralisations symboliques kuhniennes. Il cite ainsi les termes de région, de paysage, de milieu, de nature… Mais, ce changement introduit un écart majeur par rapport à la définition initiale des généralisations symboliques puisqu'il n'y a plus d'égalité de type « formel » ou « verbal » (pour reprendre la terminologie kuhnienne). De plus, quand Olivier Orain affirme ensuite que ces « "généralisation symboliques" y ont pris une forme particulièrement "indiscutable" durant les années 1910-1970 » (Orain 2003, 128), le travail que j'ai mené montre que ce n'est pas le cas au moins pour les notions de « milieu » et de « région ».

Dans un second temps, toujours pour tenter de trouver un équivalent à ces généralisations symboliques kuhniennes dans la géographie post-vidalienne, Olivier Orain mentionne un autre argument : celui des « répertoires de forme » (*Ibid* 2003, 128).

[308] Cette métaphore fait référence à une formalisation incomplète et à une fluidité du positionnement épistémologique vidalien permettant des interprétations et des travaux très diversifiés comme le montrent les exemples de Jean Bruhnes, Camille Vallaux, Maximilien Sorre, Eric Dardel…

2. Les répertoires de formes

Deux domaines sont cités par Olivier Orain : la géomorphologie et la géographie agricole avec, pour ce dernier, une remarque concernant la dimension « sans doute moins canonique » (*Ibid* 2003, 128) de ce répertoire par rapport à celui de la géomorphologie. Les deux sous-sections suivantes présentent successivement des analyses sur ces deux répertoires de formes en s'appuyant sur les travaux d'Henri Chamussy (1996) sur la formation des concepts en géographie.

a. Réflexions à partir du répertoire de formes géomorphologiques.

Le fait que la généralisation symbolique la plus convaincante, mise en avant par Olivier Orain, se réfère au domaine des sciences de la nature, est loin d'être anodin. Il est possible d'y voir une confirmation des réflexions de Thomas Kuhn sur l'importance de la détermination de la référence dans les sciences physiques pour expliquer leurs dynamiques différentes de celles des sciences sociales (*cf.* Chap2.II.1)

Pour approfondir cette remarque, les analyses d'Henri Chamussy sur la « cuesta » sont particulièrement intéressantes, car elles insistent sur la nécessité de sortir du langage naturel ou de le subvertir en redéfinissant strictement les termes. L'objectif est d'éviter la polysémie et de créer des « semi-abstractions » (Chamussy 1996) qui dépassent la simple description des formes. Le terme de « semi-abstraction » est ici parfaitement adapté au sens où la « cuesta » provient des observations des reliefs de l'est du bassin parisien. Ces réflexions renvoient exactement au processus mis en avant par Jean-Claude Passeron quand il montre que le rapprochement de cas spécifiques est à la base du travail conceptuel en SHS. Cet exemple montre donc qu'une partie de la géomorphologie s'est construit sur une dynamique intellectuelle similaire, avec toutefois, selon Henri Chamussy, un effort plus marqué que dans le reste de la géographie pour s'extraire[309] du langage naturel.

Malgré ce rapprochement, soulignons qu'Henri Chamussy et Jean-Claude Passeron sont sur des positionnements très différents, puisque le premier a été, précisément, le promoteur d'une lecture de l'histoire de la géographie française en termes de changement de paradigme. Dans cette perspective, il soutient l'idée que la formation des concepts a été initialement « parfaitement inconsciente » (*Ibid* 1996, 42). Il en vient ainsi à affirmer qu'« avant le renversement épistémologique des années 60 (en France), l'abstrait se fabriquait tout seul,

[309] Cette « extraction » passe par la création de nouveaux termes spécialisés et par une recherche accrue de précision pour définir les termes ainsi proposés.

le géographe ne savait pas que son discours en fabriquait, ou s'il s'en rendait compte, c'était à son corps défendant ! » (*Ibid* 1996, 54). Cette réflexion, bien dans la facture ironique d'Henri Chamussy, est explicitement contradictoire avec ce qui a été, par exemple, précédemment présenté à propos d'Henri Baulig (*cf.* section Chap9.II.3.b) : la conscience de l'abstraction chez certains géographes est sans conteste bien antérieure[310] aux années 1960.

Mais plus largement, par rapport à cette mise en avant d'un répertoire de formes, il est nécessaire de bien comprendre qu'il est, dans tous les cas, particulièrement difficile de valoriser en même temps sa dimension de « généralisation symbolique » et sa participation à une approche de « plain-pied ». En effet, les généralisations symboliques telles qu'elles sont ici entendues, c'est-à-dire comme des idéaux-type, tendent à définir une grille de lecture formalisée, contradictoire avec l'approche de plain-pied qui valorise avant tout les connaissances issues directement du terrain et occulte les médiations dans la construction scientifique.

Enfin, il faut ici rappeler qu'il existe une différence considérable entre ces généralisations symboliques idéal-typiques et celles plus formelles mise en avant par Thomas Kuhn. Il suffit pour s'en convaincre de comparer par exemple une formule comme celle régissant la chute des corps avec des idéaux-types qui décrirait différentes trajectoires[311]. L'établissement d'une formule mathématique, marque par rapport à une typologie générale de trajectoires, une montée en abstraction considérable. Dans un contexte précis (qui exige en amont une décontextualisation avec par exemple dans un premier temps la non-prise en compte des frottements de l'air pour établir la formule la plus simple), il y a un gain en précision majeur. L'établissement d'une formule mathématique ouvre la porte à un ensemble d'exercices-types qui se résolvent par l'utilisation du raisonnement mathématique.

Il y a donc dans la mise en avant du répertoire de formes géomorphologiques, une insuffisance cruciale par rapport aux généralisations symboliques kuhniennes[312]. Des analyses sur le second répertoire de formes, celui de la géographie agricole, permettent d'approfondir ce constat.

[310] Henri Baulig prend par exemple sa retraite universitaire en 1947.

[311] Par exemple, celle de l'objet lourd, celle de l'objet léger (de type feuille ou plume) sans vent et celle de l'objet léger avec du vent.

[312] Il existe certes aussi des approches plus mathématisées en géomorphologie, mais elles sont très spécifiques à ce domaine et peu représentatives de la géographie post-vidalienne dans son ensemble.

b. Réflexions à partir du répertoire de formes de la géographie agricole

La remarque d'une moindre canonicité de ce répertoire de formes par rapport à celui de la géomorphologie que met en avant Olivier Orain, est également développée par Henri Chamussy d'une manière un peu différente :

> «Ce qui était manipulé en géographie était, en fait, souvent à la limite de l'observation généralisée et de la conceptualisation : de l'openfield au village-rue, de l'étagement de la végétation à l'agriculture de plantation, on navigue toujours entre les classes, - qu'on appelle en général des types – dans lesquelles on range des objets distincts, uniques, mais qui "se ressemblent", et une sorte de conceptualisation floue, quelque part entre le zoom et le croquis » (Chamussy 1996, 52).

Le processus mis en avant par Jean-Claude Passeron d'une montée en généralité à partir de cas spécifiques, déjà invoquée pour le répertoire de formes géomorphologiques, se retrouve avec toutefois une conceptualisation moindre.

Il est remarquable que, dans l'ensemble de cet article sur la formation des concepts en géographie (Chamussy 1996), le processus décrit par Jean-Claude Passeron, reste valable pour tous les exemples donnés. Bien qu'une révolution scientifique soit affirmée par Henri Chamussy, elle n'apparait aucunement au niveau de la formation des concepts[313]. Il faut reconnaitre toutefois qu'il existe une évolution épistémologique importante par rapport à la géographie post-vidalienne, quand Henri Chamussy mentionne qu'il n'existe pas de critère de rationalité dans le choix de ce qu'il appelle l'« axiomatique ». Si ce terme donne une illusion de déduction logique à partir d'axiomes de départ, les détails donnés par Henri Chamussy sur la formation des concepts en géographie n'illustrent en rien une telle dynamique. Ce qui est décrit relève *a contrario* parfaitement du raisonnement sociologique tel qu'il est explicité par Jean-Claude Passeron. Par exemple, toute la tension entre des concepts trop précis qui ne permettent pas de monter en généralité et des concepts trop larges qui perdent en précision, mais favorisent des rapprochements (détaillée par Jean-Claude

[313] Pour justifier un changement de paradigme, Henri Chamussy utilise l'opposition d'Emmanuel Kant à l'empirisme de Hume. Toutefois, le projet kantien de refondation de la métaphysique à partir de jugements *a priori* est quelque peu détourné par Henri Chamussy pour l'appliquer à toute la connaissance. En effet, la distinction que Kant établit entre les mathématiques qui forment bien des jugements a priori et les sciences expérimentales qui font appel à des jugements a posteriori, n'est ni reprise, ni développée par Henri Chamussy. Seule l'idée d'une connaissance *a priori* est exposée, soutenant l'idée d'une révolution par rapport au mode de connaissance empirique.

Passeron sous la dialectique entre termes « sténographiques »[314] et « polymorphes »[315]) est très présente dans les réflexions d'Henri Chamussy.

L'ensemble de ces réflexions conduisent dès lors à mettre en évidence un paradoxe dans l'existence d'une proximité si grande entre Jean-Claude Passeron et Henri Chamussy quant à leurs réflexions sur la formation des concepts, alors que leurs positionnements théoriques par rapport à la pertinence d'une lecture kuhnienne sont totalement opposés. Pour comprendre ce paradoxe, il faut sans doute se rappeler que les affirmations de rupture ont été si importantes dans les années 1970 (notamment de la part du *Groupe Dupont* dont Henri Chamussy était une cheville ouvrière) qu'il a existé une longue rémanence durant laquelle une lecture kuhnienne s'est perpétuée.

Pour finir avec l'examen des généralisations symboliques, il reste encore possible d'argumenter l'idée d'une discontinuité majeure en mettant en avant le fait que ces généralisations n'existaient pas (ou presque pas) dans la géographie post-vidalienne alors qu'elles sont plus présentes dans la « nouvelle géographie » : ce changement marque-t-il une rupture épistémologique pouvant être lue sous le prisme de la révolution scientifique ?

3. Un marqueur de la révolution ?

Par rapport à une approche discontinuiste s'appuyant sur l'argument du développement des formules mathématiques dans une partie de la géographie française après les années 1970, il est intéressant de mentionner quelques réflexions développées à partir d'un des premiers articles de *L'Espace Géographique* : *La mobilité rurale en Aquitaine* de Pierre Duboscq (1972). Une analyse en guise d'introduction de cet article a été rédigée par Alain Reynaud. Ce dernier met en avant un élément crucial : l'article de Pierre Duboscq est bien marqué par l'utilisation d'une approche mathématique, mais cette utilisation se différencie nettement d'une approche statistique ou d'une démarche modélisante (comme a pu en développer la géographie quantitative anglo-saxonne). Les mathématiques sont utilisées par Pierre Duboscq comme un outil d'exploration visant à examiner méthodiquement un ensemble de combinaisons possibles. Alain Reynaud souligne la nouveauté du procédé et conclut par une question dont le futur a montré toute la pertinence : « À côté de la *quantitative geography* au sens strict, n'y a-t-il pas une place pour une géoanalyse ? » (Reynaud dans Duboscq 1972, 24). Cette distinction entre une géoanalyse et la géographie

[314] Les termes sténographiques sont « trop peu théoriques (c'est-à-dire trop particuliers pour disposer d'un pouvoir utilisable de généralisation ou d'analogie une fois abstrait du matériel limité dont ils se bornent à sténographier les relations) » (Passeron 2006, 97).

[315] À l'inverse, pour Jean-Claude Passeron, les concepts polymorphes sont trop théoriques, vagues et pas assez univoques.

quantitative anglo-saxonne s'appuie en amont sur cet usage différencié de l'outil mathématique.

Une partie de la « nouvelle géographie » française, notamment sous l'impulsion de Roger Brunet, s'est en effet davantage développée vers des travaux de géoanalyse que des modélisations totalement mathématisées. Il existe donc une diversité d'usages des formules mathématiques : outil d'exploration, modélisation, transfert sous forme de réflexion analogique, statistique… Cette diversité relativise l'idée d'une rupture forte basée sur l'apparition de formules mathématiques. Il existe une évolution indéniable marquée par une multiplication de leurs usages, mais, dans le détail, les modalités de leurs utilisations sont variées et renvoient à des positionnements épistémiques différents. Contrairement à l'idée qui a pu exister, d'une phase préliminaire, d'une pré-science amenée à disparaitre avec l'adoption d'une axiologie définie, cette diversité d'usages des formules mathématiques ne s'est jamais réduite. Dans cette optique, les conceptions de Jean-Claude Passeron montrent bien que les formules mathématiques constituent un outillage utile pour les SHS, mais qui reste momentané, obligeant à des processus de traduction des résultats ensuite en langage naturel et devant le plus souvent être complété par des examens complémentaires.

Il revient enfin au chapitre suivant de prendre en charge les usages démonstratifs qu'Olivier Orain fait de trois auteurs hors du champ géographique : Ian Hacking, Hilary Putnam et Jean-Michel Berthelot.

Chapitre 11 :
Analyses de trois références externes à la géographie

Les trois auteurs en question (Ian Hacking, Hilary Putnam et Jean-Michel Berthelot) sont mobilisés par Olivier Orain dès l'introduction de sa thèse, et ce, de manières différentes : Ian Hacking et Hilary Putnam y revêtent des rôles plus centraux que celui accordé à Jean-Michel Berthelot mais aussi parce que ces références n'interviennent pas aux mêmes lieux de la démonstration. Les réflexions de Ian Hacking, sont ainsi mobilisées dès l'introduction, dans la présentation de la dichotomie réalisme / nominalisme, constitutive de la thèse du plain-pied. S'appuyer ainsi sur un épistémologue et un sociologue des sciences permet à Olivier Orain de donner une profondeur et une légitimité plus grande à son travail. Il convient donc d'examiner la validité de cet appui. De même pour le philosophe Hilary Putnam qui est mobilisé du fait de son utilisation par Claude Raffestin, chercheur auquel Olivier Orain donne une place majeure dans sa thèse (comme principal théoricien du nominalo-constructivisme dans le champ de la géographie française : *cf.* Chap9.VI.2).

Pour comprendre plus globalement cette situation, il faut souligner qu'après les années 1970, la géographie française a été marquée par une diversité d'affiliations (structuralisme, marxisme, clivage idéologique et politique gauche/droite…), mais avec peu d'ancrage philosophique affirmé. Claude Raffestin fait partie des exceptions à cette situation générale. Cette particularité est d'ailleurs renforcée par le travail d'Olivier Orain lui-même : sa focale d'analyse, centrée sur le couple réalisme / constructivisme, est évidemment ancrée dans des schèmes et catégories du champ philosophique.

Quant à Jean-Michel Berthelot, qui fut membre du jury de thèse d'Olivier Orain, son positionnement épistémologique soulève des questions intéressantes : il est, en effet et sur plusieurs points en désaccord avec la thèse du plain-pied, mais également avec les propositions de Jean-Claude Passeron (2006). Il m'a paru pour cette raison intéressant de reprendre et de détailler sa position, même si, au demeurant, Olivier Orain n'a pas été trop disert à son endroit.

La plus grande importance des réflexions de Ian Hacking permet de comprendre pourquoi les analyses suivantes commencent par cet auteur.

I. Ian Hacking

Olivier Orain mobilise, principalement dans l'introduction de sa thèse, un élément de réflexion développé par Ian Hacking (2008) dans son ouvrage : *Entre science et réalité : la construction sociale de quoi ?*. Cet élément utilisé par Olivier Orain renvoie à ce que Ian Hacking appelle des « points de blocage » dans les débats autour de la problématique de la construction sociale. Par rapport aux trois points de blocage développés par cet auteur (« la contingence », « le nominalisme », « les explications de la stabilité »), Olivier Orain en

développe surtout un seul : le nominalisme. Ce choix s'explique facilement par rapport à la thèse du plain-pied. En effet, le nominalisme s'oppose philosophiquement et historiquement au réalisme, controverse qui s'est incarnée notamment dans la querelle des universaux[316] (Libera 1996). Il faut toutefois souligner qu'il existe au moins deux autres raisons expliquant l'absence de développement des autres points de blocage.

Tout d'abord, Ian Hacking critique fortement dans la partie dédiée au deuxième point de blocage (les « explications de la stabilité ») la pertinence du système explicatif kuhnien. Il souligne que les propositions de *La structure des révolutions scientifiques* sont liées à un contexte particulier provenant de l'instabilité des sciences du début du XX^{ème} siècle. Comme cette instabilité a disparu, la pertinence du système kuhnien se retrouve, selon lui, remise en cause. Cette critique est abordée par Olivier Orain en note de bas de page au début de sa deuxième partie[317], mais non développée. Ce dispositif lui permet de ne pas expliciter ce positionnement de Ian Hacking, éminemment défavorable à des lectures kuhniennes appliquées aux sciences contemporaines, au moment où ces mêmes réflexions sont mobilisées par Olivier Orain comme des éléments de justification d'une approche kuhnienne.

En outre, la présentation que fait Ian Hacking de Thomas Kuhn en développant ces points de blocage insiste fortement sur l'externalisme de ce dernier alors qu'Olivier Orain développe plus fondamentalement dans sa thèse une vision internaliste. Par conséquent, le choix d'Olivier Orain de ne pas trop développer l'ensemble des points de blocage explicités par Ian Hacking se comprend aisément au regard de ces différences significatives.

L'élément utilisé et développé dans la thèse du plain-pied est donc stratégique. En effet, Ian Haking en utilisant la dénomination « point de blocage » peut donner du poids à l'idée d'un changement opposant des paradigmes incommensurables. L'expression « point de blocage » illustre l'impossibilité d'une simple adaptation. Le passage suivant reproduit la longue citation de Ian Hacking directement reprise par Olivier Orain dans son introduction :

---

[316] Les universaux doivent se comprendre par rapport aux « particuliers ». Par exemple, la circularité est un universel qui subsume toutes les formes de cercles possibles. La querelle des universaux a porté sur la question de savoir si ces universaux ont une existence en soi et comment ils s'articulent avec l'existence des particuliers. Le débat remonte à Platon. Il a été aussi particulièrement vif au Moyen Âge avec Pierre Abelard, Thomas d'Aquin, Guillaume d'Ockham… Enfin, il a été réactivé par la philosophie analytique du XX^{ème} siècle, notamment avec Bertrand Russel et David M. Amstrong.

[317] L'argument d'Olivier Orain est que le système kuhnien permet aussi de comprendre la stabilité. Cette argumentation est tout à fait valable, mais elle ne répond que partiellement à la critique de Ian Hacking. En effet, s'il n'existe plus vraiment de révolution scientifique, alors le prisme kuhnien perd de sa légitimité même s'il peut expliquer en partie la stabilité. L'intérêt de l'œuvre de Kuhn reste fondamentalement lié aux successions paradigme − révolution.

> « Plutôt que de répéter quelque épisode de l'histoire de la philosophie, je vais tenter de proposer une version contemporaine des vieux enjeux du nominalisme […]
>
> L'un des camps espère que le monde peut, par sa propre nature, être structuré de la manière dont nous le décrivons. Même si nous n'avons pas tout à fait raison sur la nature des choses, il est au moins possible de considérer que le monde est structuré de cette manière. L'enjeu principal de la recherche est de découvrir le monde. Les faits sont là, ordonnés à leur manière, indépendamment de celle dont nous les décrivons. Penser autrement, c'est manquer de respect à l'univers et souffrir d'*hubris*, pour exalter ce couinement de foutriquet : l'esprit humain.
>
> L'autre camp dit qu'il a un respect encore plus profond pour le monde. Le monde est si autonome, si à lui-même, qu'il n'a même pas ce qu'on appelle une structure en lui-même. Nous fabriquons nos représentations dérisoires de ce monde, mais toutes les structures que nous pouvons concevoir se situent dans nos représentations. Celles-ci sont sujettes à des contraintes sévères, naturellement. Nous avons des attentes en ce qui concerne nos interactions avec le monde matériel, et, quand elles ne sont pas satisfaites, nous ne nous mentons pas à nous-mêmes à leur propos, ni à personne d'autre. Dans le domaine assez public de la science, les ruses de l'appareillage et le génie de la théorie sont utiles au maintien de cette honnêteté » (Hacking 2008, 118).

À l'issue de cette citation, Olivier Orain associe fortement les notions de « nominalisme » et de « constructivisme ». Cette association, utile à sa thèse, est contestable. En effet, la « crise de la géographie » a débouché sur des approches qui peuvent être qualifiées de plus « constructivistes », mais pas forcément de plus « nominalistes ». Des géographes ont revendiqué des postures constructivistes alors qu'ils se rattachaient fortement au premier camp décrit par Ian Hacking, espérant « que le monde a une structure qui peut être partiellement atteinte par la dynamique de connaissance » (*Ibid* 2008, 118).

De plus, il est ici utile de rappeler, comme cela a été précédemment développé (*cf.* section Chap8.I) que pour désigner le premier camp, Ian Hacking ne retient pas le terme de « réalisme » trop flou sémantiquement, mais propose l'appellation de « structurisme inhérent » (*Ibid* 2008, 119). Olivier Orain occulte cette réflexion de Ian Hacking, car cette dernière appellation rendrait explicite le fait que la « crise de la géographie » ne s'est pas jouée sur ce point de blocage. En effet, Roger Brunet, acteur majeur de la « crise », en mettant en avant une géographie structuraliste reste fortement attaché au premier camp. Sur ce plan, il se situe, donc, dans une continuité avec la géographie classique sur ce plan.

Il reste possible de penser que la tension, voire la crise entre structurisme inhérent et nominalisme se joue après la période qui précisément a été nommée « crise de la géographie ». Or, ce que montre Olivier Orain sur cette période, ce sont des évolutions qui ont permis de formaliser un positionnement plus constructiviste (notamment avec le *Géopoint* de 1978) et des dynamiques intellectuelles multiples (avec parfois des positionnements un peu plus orientés vers le nominaliste comme ceux de Claude Raffestin). Aucune crise kuhnienne n'est mise en avant à cette période alors que le cœur du point de blocage tel que défini par Ian Hacking aurait dû avoir lieu à ce moment-là.

Enfin, dans la conception de ces points de blocage, Ian Hacking fait valoir l'importance de la perspective historique. Par exemple, après avoir détaillé un débat récent opposant des 'relativistes culturels et historiques' et un physicien (Weinberg), Ian Hacking conclut :

> « Il [Weinberg] écrit comme s'il touchait du doigt un quelconque débat éphémère qui s'est enrichi au cours de ces trente dernières années environ. Je suggère que son doigt indique une paire d'attitudes qui se sont opposées, l'une à l'autre au cours d'au moins 2300 années. Mon analyse 'en points de blocage' entend mettre l'accent sur le fait que ce n'est pas la première fois que d'honnêtes gens, profondément engagés, se retrouvent, disons, bloqués » (Hacking 2008, 127).

Olivier Orain ne prend pas en compte une telle perspective. Il est vrai qu'un problème durant depuis 2300 ans et continuant de revenir périodiquement constitue difficilement un changement de paradigme. Ces explicitations montrent bien qu'Olivier Orain s'est servi à dessein d'une proposition de Ian Hacking en passant sous silence tous les développements problématiques vis-à-vis de la thèse du plain-pied. La section suivante consacrée à Hilary Putnam reste marquée par la tension entre réalisme et nominalisme. En effet, celle-ci est au cœur de l'œuvre de ce philosophe mobilisée par Olivier Orain par l'intermédiaire de Claude Raffestin.

II. Hilary Putnam

Hilary Putnam (1926-2016) est un philosophe américain qui a revendiqué son appartenance au courant pragmatiste. L'importance du réalisme dans ce courant a été montrée et expliquée par Claudine Tiercelin (2017). Cette importance se retrouve dans les positionnements complexes d'Hilary Putnam qui ont d'ailleurs évolué tout au long de sa vie :

- Avant les années 1980, d'Hilary Putnam a défendu un réalisme dit « métaphysique » qui affirme une correspondance entre les mots et les choses (par exemple, entre l'eau et les molécules H_2O).
- Au cours des années 1980, il a, au contraire, fermement combattu le réalisme métaphysique et a milité pour un réalisme dit « interne » qui conçoit la vérité plus comme une acceptabilité rationnelle : « une sorte de cohérence idéale de nos croyances entre elles et avec nos expériences telles qu'elles sont représentées dans notre système de croyances » (Tiercelin 2013).
- Enfin, au cours des années 1990, son positionnement a finalement évolué vers un réalisme dit « naturel » où il remet en cause sa conception de la vérité réduite à l'acceptabilité rationnelle dans le réalisme interne.

Sans détailler au-delà ces revirements de positionnement au cours du temps, il est possible de remarquer que leur point commun est la dénomination « réalisme ». En effet, Hilary Putnam a fermement combattu une approche strictement nominaliste. Il s'est opposé plus spécifiquement à l'approche kuhnienne qui réduit la rationalité à « ce que dit notre culture locale » (Putnam 1981, 128). Il existe donc une certaine ironie à rapprocher ainsi Hilary Putnam du nominalisme pour soutenir une lecture kuhnienne, deux perspectives en contrepoints notoires des réflexions de ce philosophe.

Pour tenter de comprendre, tout de même, la mobilisation de cet auteur dans une perspective nominaliste, précisons que le réalisme interne marque, avec sa conception de la vérité comme « acceptabilité rationnelle », un positionnement qui trouve des échos avec le nominalisme sans toutefois s'y confondre. Si dans le réalisme interne, la vérité repose sur une acceptabilité rationnelle et qu'elle dépend donc, aussi, en partie de facteurs sociaux, le discours d'une communauté ne suffit pas. Comme l'affirme Hilary Putnam : « La "vérité" d'une secte khomeyniste n'est pas digne de ce nom, parce qu'elle ne réagit à rien si ce n'est à la volonté du leader »[318] (Putnam 1992, 420). L'enjeu est ici d'éviter les problèmes liés au relativisme.

Par rapport à cette problématique, une réponse des pragmatistes est de faire reposer la rationalité, non pas sur une grande théorie globale, mais plutôt sur l'objet de quêtes locales et multiples. En ce sens, la valorisation d'enquêtes n'acquérant jamais le statut de lois universelles chez Jean-Claude Passeron correspond assez bien à ce processus, même si ce dernier ne s'est jamais revendiqué comme étant « pragmatique »[319]. Ce rapprochement théorique reste partiel et délicat du fait de l'absence de définition univoque du pragmatisme.

Enfin, il faut souligner que l'utilisation de la référence d'Hilary Putnam n'est jamais approfondie par Olivier Orain jusqu'à traiter du problème du rapport entre les mots et les choses. Quand il évoque, par exemple, le problème du réductionnisme linguistique, il met de côté la question en écrivant que cette dernière dépasse l'ambition de son propos (2003, 133). Il s'agit pourtant d'un des cœurs du problème. Sur cette problématique, il faut reconnaître que la prise en compte des développements de Jean-Claude Passeron ne s'avère pas d'un grand secours. En géographie, le chercheur ayant le plus poussé la réflexion par rapport à cette problématique, est sûrement Augustin Berque (2014). Ce dernier a résumé son positionnement par la formule générale suivante :

---

[318] Il faut souligner que ce sujet prend avec l'actualité des « fake news » une importance fondamentale et il n'est pas étonnant que ces questionnements soient aussi réactualisés scientifiquement.

[319] Une des raisons de cette non-affiliation provient aussi du fait que la sociologie pragmatique renvoie à un courant s'ancrant dans les travaux de Luc Boltanski et Laurent Thévenot. Jean-Claude Passeron ne fait pas partie de ce courant et utiliser ce terme de « pragmatique » peut apporter en ce sens une certaine confusion. Le rapprochement ici effectué doit se comprendre à un niveau plus philosophique avec cette attention à la pratique.

r= S/P

Cette formule se traduit par le fait que la réalité (r) est toujours un sujet (S) saisi par un ou des prédicat(s) (P). Par rapport à ce rapport S/P, oublier P, ça serait tomber dans le scientisme et oublier S, ça serait tomber dans le fanatisme. Cette réflexion recoupe la préoccupation explicitée par Hilary Putnam précédemment et interrogeant la valeur des discours d'une secte khomeyniste. Indirectement, ce rapport S/P, peut être rapproché de l'entre-deux passeronien. Dans ce dernier, le pôle expérimental vise à ne pas tomber dans une interprétation libre (P) et le pôle herméneutique évite une naturalisation des variables et des assertions (S).

Une des questions de fond est dès lors celle de l'articulation des deux pôles. La section suivante examine l'entre-deux présenté par Jean-Michel Berthelot dans son ouvrage *Épistémologie des sciences sociales* (2018), car il présente une articulation différente de celle de Jean-Claude Passeron.

III. L'entre-deux de Jean-Michel Berthelot

L'entre-deux développé par Jean-Michel Berthelot (2018) pour penser les sciences sociales s'appuie sur le concept de programme de recherche d'Irme Lakatos. Contrairement à Olivier Orain (*cf.* section Chap10.III.4), Jean-Michel Berthelot utilise ce concept dans une perspective épistémologique semblable à la proposition originelle d'Irme Lakatos. Son application aux sciences sociales le conduit à élaborer une véritable architecture interdisciplinaire, définie par plusieurs programmes de recherche. Sans revenir sur les détails de cette élaboration, les programmes correspondent à six schèmes avec deux grands modèles épistémologiques (moniste ou dualiste) et trois pôles (naturaliste, intentionnaliste, symboliciste) pouvant être graphiquement synthétisés de la manière suivante.

Pôles	Modèles épistémologiques	Schèmes	Programmes
Naturaliste	Moniste (modèle des sciences de la nature)	Causal	De type factoriel : y = f(x1,x2,x3…) De type systémique (S1 → S2)
		Fonctionnel	Simple : fonction de X pour S Complexe : articulation fonctionnelle de systèmes complexes
		Dialectique / évolutionniste	Simple : a et non a → X Complexe : transformations successives d'un système
Intentionnaliste	Moniste (application du modèle déductif-nomologique)	Actanciel	Utilitarisme Actionnisme Théorie de l'action organisée
	Dualisme (irréductibilité de la compréhension)		Sociologie phénoménologique Interactionnisme symbolique Ethnométhodologie
Symboliciste	Moniste (idem)	Structural	Structure générative (par exemple combinatoire) Homologies structurales
	Dualisme (irréductibilité de sens)	Herméneutique	Recherche de signifié divers : sens du monde, de l'expérience vécue, archétypes, pulsions collectives…

Tableau n°27 : Méta-architecture des SHS selon (Berthelot 2018, 504).

L'objectif n'est pas de détailler l'ensemble du contenu de ce tableau. Le lecteur intéressé pourra trouver plus de détails dans l'ouvrage cité. Soulignons que par rapport à la construction de Jean-Claude Passeron (2006), cette architecture se caractérise aussi par un positionnement entre la grammaire du pôle expérimental (souvent proche du schème causal situé en haut du tableau) et la rhétorique du pôle herméneutique (en bas du tableau).

L'articulation entre les deux pôles passe chez Jean-Michel Berthelot par la définition de grandes voies théoriques alors qu'elle se construit davantage dans des bricolages spécifiques sans que de grands courants soient nommés et formalisés chez Jean-Claude Passeron. Ce dernier ne refuse pas de reconnaitre qu'un ensemble de chercheurs peut utiliser des procédés similaires (et ainsi définir une école ou un courant), mais il n'existe pas pour lui une méta-architecture théorique (contenant un ou des paradigmes) qui pourrait subsumer l'ensemble des recherches. Les paragraphes suivants tentent d'appliquer l'architecture de Jean-Michel Berthelot à l'histoire de la géographie pour analyser si la lecture kuhnienne d'Olivier Orain peut trouver une légitimité dans ce déploiement différent de l'entre-deux.

Une première constatation d'ordre général est qu'un transfert de cette architecture à l'histoire de la géographie est loin d'être évident. Les dénominations d'une grande partie des programmes de recherche (utilitarisme, interactionnisme symbolique…) sont très marquées par des écoles sociologiques[320]. Or, la géographie a été historiquement beaucoup moins marquée que la sociologie par ces grandes différenciations méthodologiques. Il est tout de même possible d'essayer de lire le passage de la géographique classique à la « nouvelle géographie » en s'appuyant sur les schèmes les moins marqués par des écoles sociologiques. L'objectif sous-jacent est de déterminer si une révolution scientifique, correspondant à un changement de schème, peut être mise à jour de cette manière.

Au niveau des schèmes, la géographique classique a surtout mobilisé celui qui est intitulé « causal » par Jean-Michel Berthelot. Toutefois, toute la problématique autour du déterminisme et du possibilisme tempère ce schème causal dans un positionnement épistémologique qui ne renvoie pas strictement au modèle des sciences de la nature, contrairement à ce qu'affirme la classification proposée (*cf.* la colonne « modèle épistémologique » du Tableau n°27). Ce pas de côté effectué par la géographie classique a permis de mettre en avant d'autres schèmes, mais d'une manière beaucoup moins affirmée, c'est-à-dire sans l'existence de véritables programmes de recherche associés. Par exemple, l'importance de la perspective historique (Paul Vidal de la Blache, Roger Dion, Xavier de Planhol…) a eu comme conséquence de recourir dans certains cas au schème « dialectique / évolutionniste ». Sur un autre plan, l'étude des structures agraires (Albert Demangeon, André Cholley, René Lebeau…) renvoie en partie au schème « structural » avec une application dans la géographie humaine post-vidalienne qui, pour autant, y est restée limitée.

La « nouvelle géographie » a évidemment beaucoup plus activé ce schème structural, notamment à travers la perspective chorématique, mais sans abandonner le schème causal qui demeure au cœur de beaucoup d'analyses systémiques. Dans cette perspective, il existe

[320] Cette importance d'un lexique sociologique s'explique facilement du fait du positionnement de ce chercheur avant tout dans ce champ.

une continuité indéniable entre la combinaison vidalienne et l'approche systémique qui relèvent toutes deux de ce même schème. Si les recherches sur l'espace produit et vécu ont pu mettre en avant des approches relevant davatange des schèmes actanciel et herméneutique, il n'y a pas eu de basculement net vers un nouveau programme de recherche. Le choix réalisé par Olivier Orain de mettre en avant Claude Raffestin correspond à un choix stratégique de valoriser des réflexions développant le pôle symboliste dans une distance évidente avec un pôle naturaliste plutôt marqué historiquement par des approches réalistes. Cette position de Claude Raffestin explique aussi sa rupture avec le schème structural qui a marqué initialement la « nouvelle géographie », rupture qui s'est incarnée par le passage conceptuel qu'il revendique de l' « espace » au « territoire » (*cf.* section Chap9.VI.2).

Ces développements sont loin de rendre compte d'une révolution scientifique, car il n'y a pas de net basculement d'un schème vers un autre, mais plutôt des mobilisations de plusieurs schèmes. Les évolutions ne s'effectuent pas de manière unilatérale et irréversible. Par rapport aux réflexions passeroniennes, cette approche de Jean-Michel Berthelot privilégiant de grandes voies de recherches et des schèmes déjà définis, cadre mal *in fine* avec les évolutions de la géographie. L'idée de recherches qui négocient localement leur positionnement entre plusieurs schèmes, avec aux deux extrémités, le modèle (schème causal) et le récit (schème herméneutique), me semble plus opérant. Cette idée d'un bricolage spécifique destiné à faire tenir au mieux chaque recherche me semble également plus proche de mon travail doctoral et de l'ensemble des travaux que j'ai pu réaliser ou suivre de près.

IV. Réflexions conclusives

Cette troisième grande partie a permis d'explorer plus en détail des difficultés qui adviennent quand on met en oeuvre une lecture paradigmatique de la géographie française telle que l'a effectuée Olivier Orain. Sans revenir sur l'ensemble de ces difficultés, nombreuses et importantes, je souhaite simplement rappeler quelques points sur les trois plans étudiés, en guise de synthèse :

- Au niveau, tout d'abord, de l'ensemble des étapes du schéma kuhnien:
    - Du stade pré-paradigmatique au premier paradigme : l'évolution entre vidaliens et post-vidaliens a été accentuée par Olivier Orain pour créer une rupture qui reste très discutable. Le passage d'une proto-science à une science qui marque la première étape du schéma construit par Thomas Kuhn est

également particulièrement problématique dans son application à la géographie française.

- L'« anomalie » : la sphère aménagementiste n'a pas développé d'élément incommensurable avec la géographie classique, même si des tensions ont existé.
- La « crise » : les rêves du passage de l'inductif au déductif avec une géographie construite à partir d'axiomes, a tourné court. De plus, si l'expression « crise de la géographie » est commune, cela ne signifie pas et n'équivaut pas à une « crise kuhnienne ». De nouvelles voies scientifiques ont été explorées sans remplacer les anciennes. Il n'a pas existé deux mondes incommensurables.
- Le nouveau paradigme : Olivier Orain reconnaît son absence, ce qui est totalement contraire au schéma kuhnien.

- Au niveau, ensuite, des éléments originels, définissant un paradigme et décrits dans la postface de la *Structure des révolutions scientifiques* (2008, *1970*) :
  - Il existe évidemment très peu de « généralisations symboliques », telles que les entend Thomas Kuhn, dans les géographies post-vidaliennes.
  - Pour les « exercices-types » et les « valeurs », il existe des écarts importants entre les descriptions kuhniennes et les pratiques effectives des géographies post-vidaliennes : pas de valeurs de prédictions quantitatives, pratique réduite et insuffisante du *puzzle solving*…
  - La question de la « métaphysique » est complexe dans le transfert effectué par Olivier Orain. Les nouveautés introduites (opérateur épistémologique et noyau dur) sont peu convaincantes et posent la question de savoir s'il s'agit encore du modèle kuhnien, tant les adaptations et les différences sont importantes.

- Enfin, sur les trois auteurs hors de la sphère géographique convoqués par la thèse du plain-pied et étudiés ici plus en détail (Ian Hacking, Hilary Putnam et Jean-Michel Berthelot), de nombreuses inadéquations entre leurs réflexions et leurs utilisations ont pu être soulignées. Ces décalages sont loin d'être mineurs et sont souvent non développés par Olivier Orain alors qu'ils touchent le cœur même de sa problématique.

Si les réflexions passeroniennes ont permis de développer en contrepoints plusieurs propositions, il faut reconnaitre que le thème central de cette partie – l'histoire de la géographie française – reste un sujet de niche. Or, l'étude des relations entre pôle

expérimental et herméneutique, au cœur des développements passeroniens, a un cadre d'application beaucoup plus grand et reste, à mon sens, d'une grande actualité. Cette réflexion m'a conduit à analyser comment ces relations jouent et sont à l'œuvre dans les sciences contemporaines, notamment dans les humanités numériques (auxquelles se rattache ce travail) et dans les données massives qui constituent pour certains un nouveau « paradigme » (Hey 2012).

Beaucoup de chercheurs en sciences sociales ont utilisé à dessein d'autres termes et expressions que « révolution » et « changement de paradigme ». Par exemple, Valérie Carayol et Franck Morandi (2019) ont utilisé plus modestement le syntagme de « tournant numérique ». Gilles Bastin et Paola Tubaro (2018), celui de « moment big data »… Ces dénominations plus nuancées n'empêchent pas de mobiliser une partie des réflexions ici menées pour penser ces dynamiques contemporaines marquées par un retour du quantitatif.

Il s'agit alors de se servir des développements réalisés sur une problématique historique pour tenter de mieux comprendre une partie de la situation scientifique actuelle. Il convient toutefois de ne pas conférer à cette partie une trop grande ambition, car la « situation scientifique actuelle » se caractérise évidemment par sa multiplicité et ne peut être envisagée qu'avec un recul limité.

Quatrième partie

Humanités numériques et données massives : mises en perspective contemporaines

L'objectif de cette partie est de comprendre en quoi les réflexions précédemment menées ont un intérêt au-delà de leur cadre initial qui est caractérisé par une problématique centrée sur l'histoire et l'épistémologie de la géographie française (*cf.* Chap2). L'utilisation des propositions et conceptions de Jean-Claude Passeron (2006) a déjà permis d'insérer la réflexion menée dans le cadre plus large des SHS (*cf.* section Chap8.II). Il s'agit maintenant de développer plus spécifiquement la portée et l'intérêt de ces réflexions dans le cadre de l'actualité autour des humanités numériques et des données massives. Il s'agit de relever le défi d'une inter-compréhension entre les dynamiques de production scientifique de deux époques, les années 1960-70 et les années 2010-20, décennies durant lesquelles une partie des SHS a été marquée par une montée en puissance des approches quantitatives. Ce rapprochement est complexe, car les contextes et les évolutions sont loin d'être semblables.

Un phénomène, qui a grandement facilité la réalisation de cette partie, est la mobilisation par bon nombre d'auteurs contemporains de la référence kuhnienne, et cela, aussi bien dans les humanités numériques que dans les données massives. Jean-Edouard Bigot *et al.* (2016) écrivent à ce propos :

> « c'est donc un principe de rupture qui semble caractériser l'ensemble de la pensée des "humanités numériques". La mutation technologique globale que constituerait le développement du numérique entraînerait une transformation sociale de grande ampleur, voire une "rupture anthropologique" (Stiegler, 2014), à laquelle devrait correspondre un changement paradigmatique profond des SHS (Kuhn, 2008). C'est, en quelque sorte, l'avènement d'une nouvelle société qui sert de béquille à l'avènement d'une nouvelle façon de faire de la recherche sur la société. À une nouvelle société devrait correspondre une nouvelle science sociale. Ce discours d'accompagnement nous semble devoir être interrogé pour mieux comprendre la réalité des pratiques en SHS face à la montée en puissance du numérique » (Bigot, Julliard, et Mabi 2016, 4).

Ma thèse partage la nécessité d'interroger les discours d'accompagnement des humanités numériques sans aborder, ni prendre en charge la seconde partie complexe du programme de Jean-Edouard Bigot *et al.*, à savoir une compréhension de la « réalité » des pratiques numériques en SHS.

Sans rentrer ici dans la question complexe de définition des humanités numériques (*cf.* Chap12), on peut au préalable facilement observer et affirmer que leurs discours d'accompagnement sont loin d'être univoques. L'état des lieux réalisé m'a conduit à identifier un argument souvent avancé par les tenants, plus ou moins affirmés, de la révolution scientifique (Zighed 2014; Morandi, Baltz et Delamotte 2021…) : dans une perspective historique, le numérique peut ainsi être vu comme constituant une troisième révolution (après la première marquée par l'invention de l'écriture et la deuxième découlant de l'invention de l'imprimerie). Ce changement technique profond ne peut, selon ces auteurs, qu'aboutir à une transformation des manières de produire de la connaissance. Le contenu

des savoirs n'est en effet nullement indépendant des outils pour obtenir des données, les stocker, les analyser et aussi communiquer : avec sur tous ces plans, des transformations non négligeables dues aux développements informatiques des dernières décennies. Autrement dit, ces mutations instrumentales ne peuvent qu'avoir, à plus ou moins long terme, des conséquences épistémologiques sur les manières de produire les SHS.

Une des subtilités du positionnement issu de mon travail est qu'il ne remet globalement pas en cause ce dernier argument. Ce que j'interroge, et après analyses, ce que je critique fortement est la pertinence d'utiliser le prisme kuhnien pour penser les conséquences de ces évolutions techniques sur les SHS. Comme pour la géographie française, il n'existe pas deux mondes incommensurables. Il n'est pas question de nier ici l'existence de divergences dans les manières de penser et de faire, ainsi que la présence de profondes controverses entre les prometteurs de solutions technologiques et ceux qui sont plus réfractaires. Toutefois, il peut exister de vives controverses sans changement de paradigme. Sur ce sujet, l'état des lieux réalisé n'a pas identifié de référence examinant finement les éléments constitutifs d'un paradigme selon Thomas Kuhn (généralisation symbolique, métaphysique, valeur, exemple type) pour les humanités numériques. Cette absence ne peut qu'étonner comparativement à l'importance des emplois des concepts de « paradigme » et de « révolution scientifique » dans ce champ.

Malgré ce constat introductif, la partie qui suit n'est pas orientée vers la reproduction du travail réalisé pour la géographie française : chaque élément de la matrice disciplinaire kuhnienne n'a pas été examiné pour les humanités numériques. Les développements de Jean-Claude Passeron n'ont pas fait l'objet non plus de reprises multiples. Il m'a semblé moins redondant et plus fécond d'explorer plusieurs pistes de travail à partir de textes produits en humanités numériques, sachant que les auteurs dont je discute les positionnements ont été sélectionnés en fonction de l'intérêt qu'ils présentaient au regard de mon travail.

Ces analyses ont été complétées ensuite par un focus sur les données massives (*big data*). Ce choix peut paraître à première vue étonnant – les données mobilisées pour cette thèse n'ont rien de « massives » – mais il s'explique par le fait qu'une partie des enjeux contemporains relatifs à ma problématique gagne à mon avis en compréhension et en intensité si les réflexions se nouant autour des données massives y sont développées.

Il est vrai que les expressions « *big data* » et « données massives » ont été surtout utilisées entre 2005 et 2020 et sont actuellement en recul devant le recours de plus en plus important aux expressions d'« apprentissage profond » (*deep learning*) et d'« intelligence artificielle ». Toutefois, les données massives restent fortement présentes, car une grande partie des avancées en intelligence artificielle s'appuie sur des données massives. Pour cette raison, ma

thèse prend en compte la période la plus actuelle malgré une utilisation moindre de cette expression de « données massives ».

Par rapport aux années 1960-1970 où les mouvements quantitativistes se sont développés à partir des disciplines elles-mêmes (géographie, mais aussi histoire, linguistique…), la situation avec les données massives est souvent différente. Comme le souligne Pierre Mounier (2018), dans certains cas, des ingénieurs s'emparent eux-mêmes des problématiques de SHS et les traitent directement avec leur outillage en laissant une très faible place aux spécialistes de ces disciplines. Dans ce contexte, l'utilisation du concept de paradigme, conçu par Thomas Kuhn pour expliquer en partie la stabilité des sciences « dures » par rapport aux SHS, pose d'autant plus question. En outre, quelques auteurs, comme Chris Anderson (2008), revendiquent pour les données massives un changement de paradigme en affirmant qu'il est possible et pertinent de se passer d'hypothèse et de modèle. L'observation des corrélations et des motifs sur de grandes masses de données serait alors suffisante. D'où l'hypothèse que j'ai réalisée, d'un parallèle avec les affirmations de « réalisme » et de plain-pied mobilisées par Olivier Orain (2003) pour les post-vidaliens : cette proposition conduit à un grand écart entre deux époques, mais constitue épistémologiquement une piste que j'ai estimée assez pertinente pour être explorée.

Pour finir, les limites des humanités numériques et des données massives sont, à mon sens, trop floues pour tenter d'objectiver précisément leurs intersections et leurs interactions. C'est, en partie pour cette raison, que j'ai abordé ces deux domaines successivement et non simultanément. Néanmoins, il se peut qu'en lisant la section sur les données massives, certains lecteurs pensent que quelques exemples ou réflexions auraient pu trouver leur place dans la section précédente sur les humanités numériques (et inversement). Ce qui a déterminé la place des développements est surtout l'angle que les auteurs mobilisés ont choisi[321]. Plus globalement, le travail suivant ne vise aucunement à l'exhaustivité, mais seulement à expliciter les mises en perspective que j'ai jugées les plus intéressantes.

[321] Par rapport, bien entendu, à la partie de leur réflexion analysée dans ma thèse.

Chapitre 12 :
Les humanités numériques

Définir les humanités numériques relève de la gageure. Une solution conciliante, mais assez minimaliste d'un point de vue épistémologique est de faire référence à l'existence de communautés de pratiques et d'échanges travaillant à la fois sur les humanités et le numérique. Cette définition minimale pose toutefois quelques problèmes dès qu'il s'agit de la développer plus amplement. Devant les difficultés rencontrées pour définir les humanités numériques, ma stratégie a été de procéder de façon plus indirecte et moins ambitieuse en explicitant tout d'abord quelques tensions au cœur de ce mouvement.

I. Tensions au cœur des humanités numériques

L'expression française « humanités numériques » est marquée tout d'abord par une question de traduction, car elle provient du syntagme anglophone « digital humanities »[322].

1. Tensions liées à la traduction

Claire Clivaz (2017) a détaillé dans un article intitulé *Lost in translation ? The odyssey of 'digital humanities' in French* ces tensions liées à la traduction. La controverse la plus importante a concerné le choix à effectuer, dans la traduction française de « digital humanities », entre l'adjectif « digital » et « numérique ». Il y a eu dans un premier temps un refus de trancher et c'est pourquoi le manifeste écrit à Paris en 2010 et cosigné par plus de 250 chercheurs a été appelé : *Manifeste des Digital Humanites.* Ce n'est que quelques années plus tard que l'usage en français a consacré l'expression « humanités numériques », mais cette évolution ne correspond pas à un choix scientifique unanimement reconnu et assumé. Plusieurs auteurs comme Olivier Le Deuff et Frédéric Clavert (2014) ont défendu le choix inverse.

Sans rentrer dans les détails des débats, il faut préciser que les termes « digital » et « numérique » ne se recouvrent pas. L'adjectif « digital » renvoie globalement aux doigts alors que « numérique » recouvre deux grandes significations : la première fait référence aux nombres[323] et la seconde, plus récente, renvoie à une démocratisation des pratiques

[322] Ce sont les auteurs Susan Schreibman, Ray Siemens et John Unsworth qui ont créé ce syntagme avec l'ouvrage *A Companion to Digital Humanities* (2004). John Unsworth est revenu sur le choix de ce titre de livre qui répond avant tout à des considérations marketing : « Ray [Siemens] voulait "A Companion to Humanities Computing", car c'était le terme couramment utilisé à l'époque ; les responsables éditoriaux et marketing de Blackwell voulaient "Companion to Digitized Humanities". J'ai suggéré "Companion to Digital Humanities" pour ne pas mettre l'accent sur la simple numérisation » (traduction personnelle de Kirschenbaum 2012, 5). Cette anecdote montre que les *digital humanities* s'inscrivent pour ces auteurs dans une continuité avec les *humanities computing*.

[323] L'expression « valeur numérique » renvoie à cette dimension numérale.

informatiques avec une prise en compte de l'évolution culturelle liée à ces usages. Suivant le choix de l'adjectif et de sa définition, les grands ancêtres mis en avant pour les humanités numériques varient.

En effet, une entrée par le « numérique » dans son acceptation « informatique » conduit plutôt à privilégier une figure comme Roberto Busa (prêtre jésuite qui a œuvré dès la fin des années 1940 pour utiliser les capacités informatiques dans l'étude de l'œuvre de Thomas d'Aquin) alors qu'une entrée par le « digital » peut par exemple revendiquer les bouliers comme premiers mécanismes de calculs utilisés par des sociétés. Sans remonter nécessairement aussi loin, les travaux de Paul Otlet[324] et de Vannevar Bush[325] sont cités comme précurseurs par certains francophones revendiquant une préférence pour l'adjectif « digital » (Deuff et Clavert 2014).

Pour autant, dans les *digitals humanities*, le mouvement anglophone qui utilise originellement cet adjectif, la reconnaissance de Roberto Busa comme ancêtre originel est développée, notamment du fait de sa mention dans l'ouvrage fondateur *A Companion to Digital Humanities* (Schreibman, Siemens, et Unsworth 2004). La situation est par conséquent complexe : la revendication de l'adjectif « digital » dans la sphère francophone valorise une histoire longue au-delà des pratiques computationnelles alors qu'originellement, dans la sphère anglophone, cet adjectif a été utilisé dans une grande continuité par rapport à ces pratiques.

Ce développement montre donc qu'en deçà du problème de la traduction, il existe quelques tensions autour de la place à accorder à la computation dans les humanités numériques. Plus globalement, cet enjeu peut s'interpréter comme des tensions entre un pôle lié aux calculs et aux modélisations et, un autre, lié à des approches plus interprétatives et phénoménologiques.

2. Tensions épistémologiques et articulations entre deux pôles

Cette tension a donné lieu à un premier manifeste publié en 2008, moins connu que celui de Paris, coordonné par Jeffrey Schnapp, Todd Presner, Peter Lunenfeld et Johanna Drucker.

[324] Paul Otlet a créé en 1905 le système de classification toujours utilisé dans les bibliothèques. Son *Traité de documentation* écrit en 1943 est considéré comme un ouvrage fondateur avec des préfigurations de ce que sera internet. Son œuvre a donné lieu au projet de recherche *HyperOtlet* : https://hyperotlet.hypotheses.org (consulté le 18/09/2023).

[325] Dans un article intitulé *As we may think* paru en 1945 dans le magazine *Atlantic Monthly*, Vannevar Bush décrit un système, appelé *Memex*, qui est une sorte d'extension de la mémoire de l'homme et qui préfigure l'ordinateur et le système du Web.

Ces auteurs déclarent que la première vague des *digitals humanities*[326] a été quantitative : « mobilisant les capacités de recherche et de récupération au sein des bases de données, [en] automatisant les corpus linguistiques [et en] empilant des hypercartes au sein de séries critiques » (traduction de Julien Saavedra et Citon de Schnapp *et al*. 2015). Par opposition, une seconde vague « qualitative, interprétative, expérientielle, affective et générative » (*Ibid* 2015) est affirmée.

Pour plusieurs raisons, mon travail doctoral est en désaccord avec une telle affirmation. Tout d'abord, s'il n'est pas question de nier l'existence dans les humanités numériques de travaux plus directement qualitatifs et interprétatifs, aucun travail de recherche et d'enquête n'a clairement identifié l'existence et la succession de ces « vagues ». De plus, mon travail doctoral me conduit à penser que la production d'« hypercartes au sein de séries critiques » (censée caractériser la première « vague ») peut s'effectuer dans des perspectives « qualitatives » et « interprétatives ». Plus globalement, Jean Guy Meunier dénonce les formulations[327] excessives qui exacerbent un paradoxe opposant **:**

> « le qualitatif au quantitatif, l'interprétatif à l'explicatif, l'intuitif au formel, le déductif à l'inductif, l'interprétatif à l'empirique, et où finalement se révèlent une position de type herméneutique et une autre de type positivisme logique *déguisée* en informatique […] ? » (Meunier 2019, 20).

En rentrant dans le détail des réflexions de plusieurs chercheurs, il existe, comme précédemment pour la géographie française (*cf.* Chap9), des positionnements complexes qui sont mal rendus par des approches trop duales[328]. Dans cette optique, les deux sous-sections suivantes détaillent les positionnements de Jean-Guy Meunier et Johanna Drucker qui développent des articulations bien différentes des deux pôles.

a. Jean-Guy Meunier : un développement poussé des modèles

Jean-Guy Meunier est professeur au département de philosophie de l'Université du Québec à Montréal, associé à des programmes de recherche en informatique cognitive et en sémiologie. Ses écrits sur les humanités numériques sont nombreux et cités de manière

---

[326] J'ai ici préféré ne pas traduire l'expression pour marquer le fait qu'en 2008, le mouvement s'était surtout développé dans la sphère anglophone.

[327] Ces formulations excessives sont dues pour Jean-Guy Meunier autant aux promoteurs des humanités numériques qui déclarent que ce mouvement est « une "révolution théorique", la "voie de l'avenir", un "changement de paradigme", un "nouvel espace théorique et épistémologique", etc » (*Ibid* 2019, 20), qu'à ces détracteurs qui *a contrario* dénoncent et insistent sur le manque de fondement épistémologique des humanités numériques.

[328] Il est ici possible de remarquer de manière critique que les réflexions de Jean-Claude Passeron reposent sur ces pôles semblant opposés, mais ses développements, comme les miens, portent surtout sur l'articulation de ces deux pôles.

répétée notamment par Julien Longhi dans son article *Contours, perspectives et tensions des "humanités numériques"* (2019).

Pour Jean-Guy Meunier, un objectif prioritaire est de développer la modélisation à plusieurs niveaux :

- Les modèles conceptuels exprimés en langage naturel rendent compte du ou des positionnement(s) épistémique(s) adopté(s). Les biais, les ambiguïtés et les limites d'un tel exercice ne sont pas occultés. Reprenant une formule des chercheuses Margaret Morrisson et Mary Morgan, Jean-Guy Meunier reconnaît que ces modèles peuvent servir à « rabouter ensemble *(fit together)*… des morceaux qui viennent de sources disparates » (Morgan et Morrison 1999, 15).
- Ensuite, les modèles formels « ont pour rôle de traduire certains éléments du modèle conceptuel en langage formel (mathématique, géométrique, logique, grammatical, etc.) » (Meunier 2019, 26). Les symboles de ce langage doivent être sans ambiguïté.
- Enfin, les modèles computationnels traduisent les énoncés calculables du modèle formel dans des algorithmes.

Ces modèles traduisent en partie les choix interprétatifs réalisés par les chercheurs. Une dernière partie de l'interprétation correspond à ce qui n'a pas pu être modélisé. Par conséquent, l'interprétation est réduite à des constructions de modèles et en dernière instance à la portion congrue qui n'a pas pu être modélisée. Toutefois, Jean-Guy Meunier reconnaît également que ce « sera toujours le lecteur humain lui-même, qui par sa culture, son expertise, finalisera les interprétations » (Meunier 2014, 23). Donc, même si les modélisations (avec les différents niveaux précédemment explicités) restent l'objectif central et prioritaire des humanités numériques, il existe chez Jean-Guy Meunier une reconnaissance de l'importance *in fine* du pôle interprétatif, qui est, sur ce point, assez proche du raisonnement sociologique passeronien (2006, *1991*).

Une caractéristique des réflexions de Jean-Guy Meunier reste cette priorité, ce rôle central accordé au développement des modèles, qui est plus secondaire chez de nombreux chercheurs dont fait partie Johanna Drucker.

b. Johanna Drucker : force et revendication de la dimension interprétative

Johanna Drucker est professeure au département des *Information Studies* de l'université de Californie, à la fois historienne de l'art, théoricienne des représentations visuelles et

artiste. Elle est connue dans le domaine des humanités numériques pour ses travaux autour d'une épistémologie visuelle critique qui dénonce la croyance en une neutralité des images et affirme *a contrario* le caractère situé et engagé de toute construction iconographique. Sa réflexion ne se limite pas pour autant aux images et s'inscrit dans des visées plus larges, notamment celle de la conception des interfaces, et plus généralement encore, celle des constructions scientifiques dans leur ensemble. Elle a notamment proposé de transformer la terminologie actuelle de « *data* » par celle de « *capta* », pour mieux souligner les dimensions capturées, contextuelles, historiques de toute donnée.

De plus, elle a mis en avant une opposition conceptuelle entre *mathesis* et *aesthesis*. La *mathesis* provient d'un concept repris d'auteurs antiques (Rabouin 2015), développé ensuite par René Descartes pour désigner la méthode permettant de s'orienter vers la vérité. *A contrario,* l'*aesthesis*, qui désigne en grec la faculté de percevoir par les sens, renvoie pour Johanna Drucker à une connaissance toujours située et partiale. Cette auteure insiste sur le fait que le numérique fait resurgir dans les SHS des approches marquées par la *mathesis* sans pour autant qu'il y ait de renversement : l'*aesthesis* reste le pôle central. Un des intérêts des humanités numériques est alors d'explorer les complémentarités et les tensions via ce que Johanna Drucker nomme habilement l'« interprétation modélisante » (Drucker 2020). Tout en soulignant la prééminence du pôle interprétatif, l'expression ouvre la porte à de multiples expérimentations avec des modèles plus ou moins formalisés.

Ces deux focus montrent donc des positionnements épistémologiques ne se réduisant pas à un pôle, mais présentant des articulations très différentes des deux pôles. Les tensions ne sont pas forcément vives entre ces différents positionnements, car l'existence d'une diversité épistémologique des humanités numériques a été originellement affirmée[329] et perdure. Toutefois, la diversité des positionnements épistémologiques peut donner lieu à quelques discussions animées sur des listes de diffusion, comme celle en juillet 2019 sur la liste *DH* faisant suite à la demande de conseils d'un étudiant pour s'orienter dans une carrière en humanités numériques. Alors que la demande était factuelle, les réponses[330] ont abordé de nombreuses sous-problématiques (place et rôle de l'ingénieur, niveau d'hybridation des compétences, différenciation avec les sciences physiques…) révélatrices de plusieurs tensions et désaccords.

[329] Par exemple, dans le *Manifeste des Digital Humanities* (Collectif 2010), aucun positionnement épistémologique n'est privilégié et revendiqué.

[330] L'ensemble du fil de la discussion est accessible à partir du lien suivant : https://groupes.renater.fr/sympa/arc/dh/2019-07/msg00012.html (consulté le 18/09/2023).

Il est évident que le prisme des articulations entre les deux pôles, précédemment développés, n'a pas vocation et n'est pas adapté pour expliquer l'ensemble des tensions et des désaccords dans leurs détails. Il offre toutefois une focale large dans l'approche des humanités numériques avec la possibilité aussi d'une perspective diachronique. Dans cette perspective, un point qui a retenu mon attention du fait de ma problématique de réflexion entre deux époques est l'importance de l'héritage chez plusieurs acteurs des humanités numériques francophones (Ruiz 2019; Clavert et Schafer 2019) de l'erreur d'appréciation des années 1960-1970, incarnée par la tribune d'Emmanuel Le Roy Ladurie dans laquelle il avait prédit en 1968 : « L'historien de demain sera programmeur ou ne sera plus ».

c. Éviter le « Syndrome de Stress Post-Quantitativiste »

Plus de 50 ans après, il existe certes quelques historiens qui programment, mais leur nombre est très loin d'être une majorité. L'acculturation au numérique et les changements méthodologiques sont très inégaux suivant les laboratoires de recherche. Emilien Ruiz appelle à la prudence dans la volonté des humanités numériques d'impulser et de porter une augmentation des usages numériques dans les autres disciplines pour éviter le « Syndrome de Stress Post-Quantitativiste » (Ruiz 2019). Cette expression renvoie à une formule que beaucoup d'historiens ont entendue dans les années 1970 : « vous n'êtes pas de vrai·e·s historien·ne·s si vous ne faites pas d'histoire quantitative ! » (*Ibid* 2019). Cette injonction a conduit ensuite à une dynamique dans le sens inverse avec un retour en force d'une production de connaissances historiques effectuée selon des modalités non quantitatives.

Pour éviter ce syndrome, la revendication d'une diversité épistémologique est un garde-fou pratique. Néanmoins, l'absence de positionnement épistémologique fait courir le risque d'une dérive vers une loi de la jungle. C'est pourquoi, afin de rester dans une perspective humaniste, Johanna Druckers insiste sur la nécessité de théories et Rafaël Alvarado en appelle également à renouer « avec la production de théories, un domaine dans lequel les SHS interprétatives ont développé une expertise » (Alvarado 2019, 79). Ici, le positionnement épistémologique de Jean-Claude Passeron (2006, *1991*) permet d'utiliser cette perspective historique et de préciser à la fois l'importance de l'expérimentation (qui peut être aidée par un outillage technologique) et de la relativiser (le pôle herméneutique reste central avec l'impossibilité d'une réduction à des approches totalement mathématiques et physiques).

Au-delà de l'articulation entre ces deux pôles, les indéterminations théoriques et méthodologiques des humanités numériques ont aussi conduit à plusieurs autres tensions.

3. Tensions autour des indéterminations théoriques et méthodologiques

Une formule utilisée en 2008 dans une rencontre à l'université de George Mason a fait couler beaucoup d'encre : « more hack, less yack »[331]. Bethany Nowviskie (2016) est revenue sur la genèse de cette expression afin d'éviter des oppositions stériles et de revendiquer une position plus consensuelle, se réclamant à la fois du *hack* et du *yack*. Toutefois, cette position ne résout en rien le problème définitionnel de ce mouvement qui ne peut se définir ni par un certain niveau d'utilisation du numérique, ni par une théorie spécifique, ni par des articulations définies entre « humanité » et « numérique ».

De plus, comme le soulignent Marin Dacos et Pierre Mounier (2015), ces deux termes ont plusieurs acceptations :

- le numérique peut être utilisé, tout aussi bien, et parfois dans le même temps, comme instrument de recherche, outil de communication ou/et objet de recherche.
- Les « humanités » peuvent également désigner l'objet de recherche, une orientation méthodologique ou/et un objectif plus ou moins affirmé de recherche. Sur ce dernier point, la référence latine à l'humanisme de la Renaissance avec la quête du beau, du bien et de l'homme cultivé, n'est pas forcément partagée par les « humanities » anglophones qui désignent sous ce terme plus pragmatiquement l'ensemble des disciplines liées à la culture.

Cette situation permet de comprendre pourquoi, comme l'affirme Aurélien Berra en 2012, il y a au cœur du débat sur les humanités numériques,

> « une question récurrente, et même permanente : celle de la définition. Les *digital humanities* désignent-elles en propre certaines pratiques, des méthodes, une discipline ? De l'aveu de certains praticiens, le terme constitue une sorte de "signifiant flottant" » (Berra 2012).

Les deux cartes intellectuelles proposées par Willard McCarty (cf. Figures n°25 et 26) à 8 ans d'écart montrent bien que le développement des humanités numériques n'a pas permis de répondre à cette question identitaire. La première carte a été publiée en 2003 et décrit le champ précurseur des « humanities computing ».

[331] Littéralement : plus de bricolages, d'actions, de piratages, moins de discours, de théorie, de « bla-bla ».

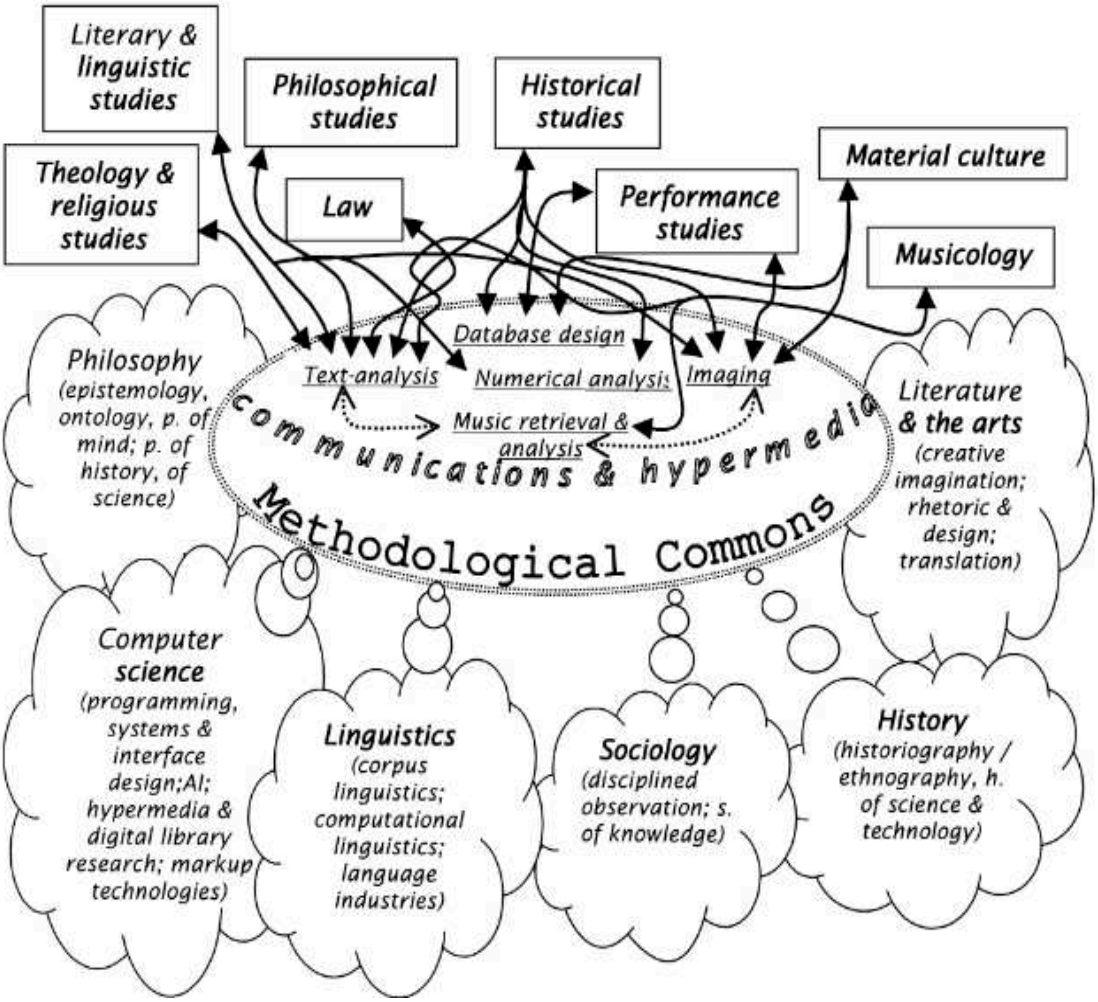

Figure n°25 : Carte intellectuelle approximative des « humanities computing » (McCarty 2003, 1225).

L'idée d'un « commun méthodologique » est au centre de cette carte. Les différents types de données (image, texte, musique, numérique…) sont reliés par l'intermédiaire de flèches à des groupes disciplinaires (études linguistiques, philosophiques, historiques…). En bas, McCarty a fait apparaitre des groupes qu'il qualifie d'interdisciplinaires (philosophie, science informatique, linguistique) reliés de manière plus floue (sous forme de bulles) au « commun méthodologique ». L'objectif n'est pas ici de détailler cette carte (et notamment cette distinction groupes disciplinaires / interdisciplinaires qui pourrait être amplement discutée), mais de la comparer avec celle présentée également par McCarty en 2011 à un colloque à Lausanne.

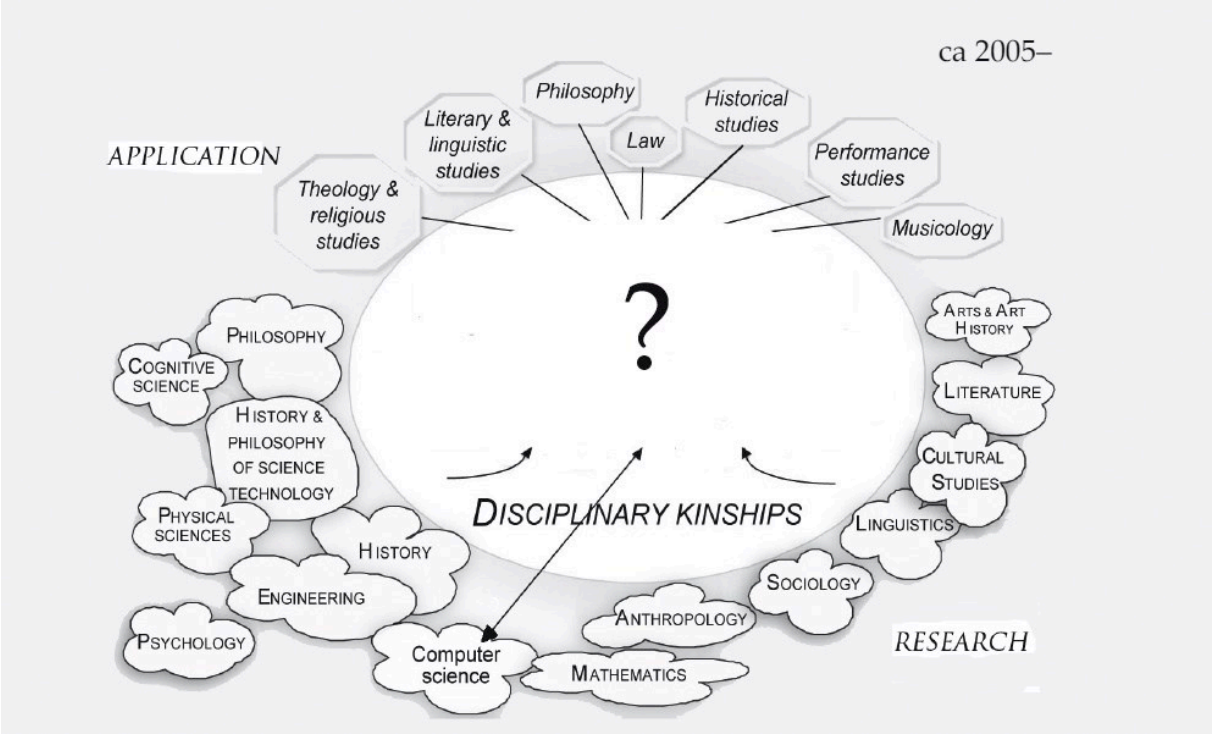

Figure n°26 : Carte intellectuelle approximative des *Digital Humanities* proposée par McCarty en 2011.

Source : https://www.slideshare.net/infoclio/prof-willard-mccarty-mimesis-to-poesis-in-the-digital-humanities

Le point d'interrogation central de cette nouvelle carte remet en cause l'affirmation précédente d'un « commun méthodologique ». Si des outils transversaux à ces données ont ensuite pu être massivement développés comme les réseaux de neurones, le champ des humanités numériques est loin de pouvoir revendiquer un socle méthodologique commun qui pourrait entièrement les définir.

Une autre notion qui a servi aussi historiquement d'essai de base définitionnelle, notamment suite au *manifeste des Digital Humanities* (Collectif 2010), est celle de « transdisicipline ». Or, Stephen Robertson, dans l'ouvrage collectif *Debates in the Digital Humanities* (Gold 2012) constate *a contrario* un non-effacement des différences disciplinaires. Il est évident que certaines recherches font appel à de multiples champs, mais cela relève dans la majorité des cas du pluridisciplinaire (et dans le meilleur des cas de l'interdisciplinaire) et non du transdisciplinaire (*cf.* section Chap1.III.2).

Ces difficultés à trouver une base théorique de définition expliquent pourquoi une partie non négligeable des chercheurs se revendiquant des humanités numériques les définissent

par le bas, par leurs pratiques effectives. Cet exercice n'est pas pour autant aisé et il faut souligner qu'une partie des tensions est aussi liée à cette structuration progressive.

4. Tensions en pratique : structurations et concurrences

De multiples activités relèvent des humanités numériques : collecte, structuration, conservation, traitement et visualisation des données, mais aussi publication, organisation et communication autour des projets, développement de réflexions plus épistémologiques… Certains acteurs sont engagés dans plusieurs de ces domaines alors que d'autres sont plus spécialisés. Des structurations plus ou moins fortes selon les pays peuvent être observées. Elles se réalisent autour de laboratoires[332], de programmes de recherche[333], de revues[334], de publications[335], d'associations[336], de rencontres[337], de manuels[338], d'infrastructures[339], de forums[340] et de formations[341]… Sans effectuer ici une recension exhaustive, il est important de souligner que cette structuration progressive, couplée à la difficulté définitionnelle des humanités numériques précédemment explicitée, a participé à des interrogations vives, provenant notamment d'acteurs eux-mêmes concurrencés par ces dynamiques. Les questions polémiques concernant le statut et la légitimité des humanités numériques ont été nombreuses : sont-elles un moyen « de penser une recomposition de disciplines universitaires dominées, voire assiégées (lettres classiques ou modernes, philosophie, etc.), qui cherchent à retrouver un second souffle » (Denizot 2015, 7) ? Ne sont-elles qu'une expression à la mode destinée à aller chercher des financements en jouant sur le créneau en vogue de l'interdisciplinarité ? Sont-elles un label amené à disparaitre étant donné que tous les chercheurs utilisent plus ou moins l'informatique et qu'il n'existe aucun critère

---

[332] *Center for History and New Media, Digital Humanities Laboratory, Medialab SciencePo* …

[333] *Valley of the Shadow*, *Venice Time Machine* puis *Time Machine Europe*, *Savoirs*…

[334] *Digital Scholarship in the Humanities, Digital Humanities Quarterly, Digital Studies/Le Champ numérique, humanités numériques, Journal of Digital History…*

[335] *Debates in the digital humanities* (Gold 2012), *Read/Write Book 2 : une introduction aux humanités numériques* (Mounier 2012)…

[336] *ADHO, EADH, Humanistica, DHd…*

[337] Conférence internationale organisée par l'*ADHO*, conférence nationale organisée par *Humanistica*, Rencontres *Huma-Num*, multiples séminaires locaux…

[338] *Doing digital humanities : pratice, training, research* (Crompton 2019), *Introduction aux humanités numériques : méthodes et pratiques* (Wilde *et al*. 2016), The Programming Historian (https://programminghistorian.org/)...

[339] *Digital research infrastructure for the Arts and Humanities* (DARIAH), *Common language ressources and technology infrastructure* (CLARIN), *Huma-Num*.

[340] Digital Humanists (https://hcommons.org/groups/digital-humanists/forum/), Liste DH (https://groupes.renater.fr/sympa/info/dh)…

[341] *Executive Master Digital Humanities* à Science Po, *Master humanités numériques* à l'Université Paris Sciences et Lettres, à l'Université Lyon 2…

permettant de définir à partir de quand un travail peut être considéré comme relevant des humanités numériques ?

Mon objectif n'est pas ici d'éclairer ces questions et controverses, mais plutôt de faire dialoguer réflexivement mes développements précédents avec quelques éléments des humanités numériques, sélectionnés en conséquence et présentés dans la partie suivante de façon chronologique.

II. Prolongements réflexifs : de Thomas Kuhn à mon propre positionnement épistémologique

1. La révolution : figure des discours des pionniers et des revues fondatrices

Le travail de Julianne Nyhan et Andrew Flin, publié sous le titre *Computation and the humanities, Towards an oral history of digital humanities* (2016), aborde la question de la révolution pour les « pionniers » et la « première génération » ayant vu les *digital humanities* s'affirmer. Il est intéressant de noter que ce thème n'était pas l'objectif de la recherche menée par ces chercheurs. Après avoir réalisé un corpus d'entretiens auprès d'une quinzaine de chercheurs identifiés comme majeurs dans les premiers temps des *digital humanities*, deux figures, celle de l'« outsider » et celle de la « révolution », ont été reconnues comme étant transversales à toutes les interviews et sont devenues ainsi les cadres d'analyse choisis par Julianne Nyhan et Andrew Flin pour construire leur conclusion. Analysant plus finement le terme de « révolution » en prenant appui sur les articles de trois revues majeures du domaine (*Literary and Linguistic Computing*, *Digital Humanities Quarterly* et *Computers and the Humanities*), les deux chercheurs montrent que les occurrences de ce terme sont très variées : « la "révolution informatique", la "révolution de l'information", la "révolution de la communication", la "révolution quantique", la "révolution technologique", la "révolution du livre électronique", la "révolution de l'interaction homme-machine", la "révolution communautaire", la "révolution des métadonnées", la "révolution de l'imprimerie", la "révolution numérique", la "révolution mobile" et même la "révolution de la rentabilité" » (traduction personnelle de Nyhan et Flinn 2016, 264).

Une mise en parallèle peut être ici réalisée avec l'évolution de la géographie française étudiée en première partie, avec pour rappel les propos de Jean-Paul Ferrier, Jean-Bernard Racine et Claude Raffestin, dénonçant le fait qu'

> « en moins d'une génération on pourrait croire, à lire ceux qui l'écrivent, que la géographie (quantitative) a subi plus d'une demi-douzaine de révolutions : l'initiale d'abord, la "quantitative", puis la révolution "méthodologique", "conceptuelle",

> "statistique", "révolution des modèles", révolution "behaviorale", "radicale", et tout dernièrement "axiomatique"… » (Ferrier, Racine, et Raffestin 1978, 292).

Les épithètes ne sont pas les mêmes, mais leur accumulation traduit dans les deux cas des imprécisions fortes quant aux contenus des révolutions en question. Il s'en suit des problématiques communes, notamment dans la détermination de ce qui relève des facteurs, externes ou internes, explicatifs des changements. Julianne Nyhan et Andrew Flin distinguent ainsi des positionnements très différents de la part des interviewés :

> « Parfois, la révolution est considérée comme extérieure au domaine […] Parfois, des chercheurs individuels sont dépeints comme des révolutionnaires […] Parfois, le domaine dans son ensemble est caractérisé comme ayant une intention révolutionnaire » (traduction personnelle de Nyhan et Flinn 2016, 265).

Ces multiples discours co-existent comme dans la géographie française des années 1970-1980 avec une rémanence du problème concernant le choix à privilégier *a posteriori* entre des lectures internalistes et externalistes des dynamiques passées.

Julianne Nyhan et Andrew Flin proposent ensuite des pistes explicatives pour rendre compte de cette figure de « la révolution scientifique » : le contexte de la technologie et de l'informatique, si prodigue en discours sur les changements révolutionnaires, est mentionné, ainsi que l'importance de l'ouvrage de Thomas Kuhn, *La structure des révolutions scientifiques* (2008, *1962*).

Il est particulièrement intéressant de souligner ici la proposition de Julianne Nyhan et Andrew Flin d'interpréter les deux motifs de la révolution et de l'outsider « d'une manière moins littérale en les considérant non pas en termes de leur véracité, mais plutôt en termes de leur fonction potentielle et de leur symbolisme pour le groupe qui les manie » (traduction personnelle de Nyhan et Flinn 2016, 268). Le motif de la « révolution » joue alors un rôle d'objectif commun mobilisateur et d'étiquette symbolique protégeant au moins dans les premiers temps le groupe qui le manie. Par apport à cette idée, Julianne Nyhan et Andrew Flin mettent bien en avant le fait que le maintien de cette étiquette peut être contre-productif et se retourner ultérieurement contre le groupe. En effet, la révolution scientifique n'ayant pas vraiment lieu ou ses contours restant flous, les opposants peuvent facilement attaquer cette qualification.

Là aussi, on peut établir un parallélisme avec les positions d'une partie des acteurs de la géographie théorique et quantitative française qui ne s'est pas attachée à cette perspective « révolutionnaire », ni au maintien d'une lecture kuhnienne. À l'aval de ces rapprochements, une question se pose alors : celle de la « fonction potentielle » et du « symbolisme » pouvant expliquer l'attachement d'une école épistémologique (portée par une partie de l'équipe *Épistémologie et Histoire de la GéOgraphie* de l'UMR Géographie-cités) à cette lecture.

Dans les hypothèses effectuées sans avoir été toutefois approfondies, l'importance d'un héritage collectif, repris et progressivement approfondi sans être remis en cause (avec ici aussi des notions d'objectif commun mobilisateur et d'étiquette symbolique) est centrale.

Pour revenir à la sphère des *digital humanities*[342], le premier auteur à ma connaissance développant une lecture kuhnienne un peu plus approfondie est Alan Liu dans l'article *Digital Humanities and academic change* (2009).

2. Alan Liu : une lecture kuhnienne plus approfondie, bien que partielle

Tout d'abord, Alan Liu montre qu'à l'origine, les *digital humanities* se sont surtout incarnées dans des « petits projets », ce qui le conduit à reconnaitre une tension entre les perspectives de l'« évolution » et de la « révolution ». Ce qui permet à Alan Liu de tout même soutenir une perspective révolutionnaire de type kuhnien est l'utilisation de deux métaphores : la première est celle de la « saturation d'un milieu finissant par précipiter » (Liu 2009, 25) qui permet au lecteur de se figurer un changement irréversible. La seconde est celle de l' « alien » (*Ibid* 2009, 25) qui signifie étymologiquement « extraterrestre », « étranger » et qui symbolise l'anomalie par rapport aux méthodologies traditionnelles. Le travail de transfert du modèle kuhnien est bien moins développé par Alan Liu que celui réalisé par Olivier Orain dans sa thèse (2003). Ils partagent toutefois l'utilisation d'un registre métaphorique avec le maintien d'une ambiguïté fondamentale par rapport à ce qui relève ou non de la métaphore.

Alan Liu nomme le nouveau paradigme dont il proclame la venue l'« humanisme global ». Derrière ce terme fourre-tout, la principale revendication est celle de la diversité sans exclure aucune tradition disciplinaire, ni épistémologie et méthode. Tout un ensemble de théories, structuralistes et post-structuralistes, sont citées comme pouvant potentiellement servir les *digital humanities* sans qu'aucune de ces théories ne soit vraiment analysée et que leurs différences soient explicitées. Plus globalement, l'article d'Alan Liu est très différent du travail plus fin, conséquent et approfondi d'Olivier Orain (2003), car c'est plus l'enthousiasme des premiers temps qui porte ici son travail. Comme pour la géographie française, cette perspective kuhnienne s'est ensuite prolongée tout en rencontrant quelques critiques.

[342] L'expression anglophone a été ici gardée, car le mouvement francophone est postérieur à l'article d'Alan Liu (2009) analysé dans la section suivante.

3. Adhésions et critiques de la perspectives kuhnienne

David Berry (2011) reprend l'idée de « vague » du premier manifeste dans un billet de blog intitulé *Digital Humanities : First, Second and Third Wave.* Il déclare l'existence d'une troisième vague se caractérisant par la capacité des chercheurs à lire et à écrire du code ainsi que par l'émergence d'une *« nouvelle subjectivité computationnelle critique et réflexive » (*traduction Citton et Masure de Berry *2015, 199).* Il revendique un changement de paradigme :

> *«* Pour les disciplines universitaires d'enseignement et de recherche, la mutation numérique pourrait représenter un moment de "science révolutionnaire", dans le sens défini par Thomas Kuhn d'une transformation dans l'ontologie constitutive de la discipline et de l'émergence d'une nouvelle "science normale" (Kuhn 1962). Cela reviendrait à dire que différentes disciplines universitaires pourraient trouver dans le computationnel un "noyau dur" ontologique très similaire, selon la définition qu'en propose Imre Lakatos (1980). Ce point a des implications considérables pour ce qui concerne l'unification des savoirs et l'idée de l'université (Readings 1996). Les sciences computationnelles pourraient jouer le rôle de fondation envers les autres sciences, soutenant et dirigeant leur développement au point d'énoncer des directives lucides pour leurs modes d'investigation » (*Ibid* 2015, 199).

L'argumentation de David Berry utilise, comme celle d'Olivier Orain (2003), la référence à Irme Lakotos, mais dans un sens différent puisque pour lui, c'est le computationnel qui « pourrait » constituer le noyau dur. L'emploi du conditionnel est en revanche un point de convergence avec Olivier Orain : la révolution scientifique est dans les deux cas remise à un futur indéterminé.

A contrario, Steven E Jones (2013) critique précisément l'application du modèle kuhnien aux *digital humanities :*

> « L'apparition du nouveau modèle des *digital humanities* vers 2004-2008 n'était pas un changement de paradigme - au sens où Thomas Kuhn entend ce terme surutilisé - dans lequel des modèles concurrents au sein d'une discipline conduisent à un nouveau point de vue dominant. Il s'agissait plutôt d'une "bifurcation" de l'informatique des sciences humaines (qui était déjà un domaine interdisciplinaire) qui a créé une nouvelle "branche" »[343] (Jones 2013, 5).

Cette notion de « branche », qui est une métaphore du système *Git* utilisé en développement informatique, me semble en effet mieux correspondre que le modèle kuhnien à la création des humanités numériques ainsi qu'à la « crise de la géographie » où les créations de branches ont été multiples.

[343] Traduction personnelle de : « The appearance on the scene of the new-model DH around 2004–2008 wasn't a paradigm shift — in Thomas Kuhn's sense of that overused term — in which competing models within a discipline lead to a new dominant view. It was more like a "fork" of humanities computing (already an interdisciplinary field of practice) that established a new "branch" » (Jones 2013, 5).

Si cette sélection de quelques auteurs peut faire croire à une importance de l'épistémologie kuhnienne (et de sa critique) dans les humanités numériques, plusieurs analyses permettent plutôt d'en relativiser fortement la présence.

4. Une présence réduite des épistémologies kuhniennes et de leurs critiques

L'article écrit par Rabea Kleymann, Andreas Niekler et Manuel Burghardt (2022), intitulé *Conceptual Forays : À Corpus-based Study of "Theory"*[344], tente d'identifier les théories mobilisées dans les publications de quatre grandes revues : *Computers and the Humanities* (1966-2004), *Digital Humanities Quarterly* (2007-2019), *Literary and Linguistic Computing* (1986-2014) et *Digital Scholarship in the Humanities* (2014-2020). Les références qui apparaissent le plus sont les suivantes : « hermeneutics », « linguistic theory », « cultural studies », « Roland Barthes », « new critics », « critical theory », « Michel Foucault », « Jacques Derrida », « phenomenology »… Ces références situent les *digital humanities* loin de l'épistémologie kuhnienne. Pour autant, le terme « paradigm » est utilisé comme un synonyme de « model », « framework », « theory », « approach » ; ce qui témoigne d'un usage relâché du concept kuhnien sans que la référence historique et épistémologique soit creusée.

Concernant la sphère francophone, dans les sept premiers numéros de la revue *humanités numériques* (soit un peu moins de 80 articles), un constat similaire peut être établi. Si le terme « paradigme » est cité dans 15 articles, il revient plus de 5 fois seulement dans 4 articles. La référence kuhnienne n'est explicitée que dans 2 de ces 4 articles. De plus, les paradigmes mis en avant sont très divers : paradigme « participatif », « de l'ouverture des données », « réticulaire », « éditorial numérique », « du Web des données »… Nous retrouvons ici le constat réalisé précédemment par Julianne Nyhan et Andrew Flin (2016).

Quant aux critiques kuhniennes, l'ensemble de mes lectures me conduit à affirmer qu'elles sont peu nombreuses. L'usage plutôt faible et relâché de l'épistémologie kuhnienne, permet d'expliquer en miroir cette présence critique réduite. Plus spécifiquement, par rapport aux développements de Jean-Claude Passeron qui ont été très utilisés dans cette thèse, je n'ai

[344] Traduction personnelle de : « Incursions conceptuelles : Une étude *corpus-based* de la "théorie" dans les revues de *Digital Humanities* ». J'ai préféré dans cette traduction garder les expressions de *corpus-based* et de *digital humanities*. La première fait référence à la distinction effectuée par Tognini-Bonelli en 2001. Les études basées sur le corpus (*corpus-based*) utilisent les données d'un corpus afin d'explorer une théorie ou une hypothèse. Les études guidées par le corpus (*corpus-driven*) s'appuient sur les données pour construire la problématisation et la théorie. Cette distinction a souvent été traduite abusivement par la dualité inductif / déductif alors qu'il s'agit plutôt à mon sens plutôt d'un côté de problématiques et d'explorations larges et de l'autre de problématique et d'explorations précises. Concernant la deuxième expression, *digital humanities,* l'étude de Rabea Kleymann, Andreas Niekler et Manuel Burghardt est exclusivement tournée vers la sphère anglophone.

identifié qu'une seule citation les explicitant dans le champ des humanités numériques. Cette dernière a été réalisée par Hélène Bourdeloie (2013) dans un article intitulé *Ce que le numérique fait aux sciences humaines et sociales. Épistémologie, méthodes et outils en question*. Cette auteure affirme :

> « J.-C. Passeron concède à ce titre que " (…) la sociologie, dont l'observation porte sur des configurations jamais réitérées intégralement dans le cours de l'histoire ou dans l'espace des civilisations, rencontre nécessairement comme limite de ses aspirations expérimentalistes la singularité des contextes historiques, dont la richesse déborde toujours les possibilités d'une analyse expérimentale, qui ne peut maîtriser *stricto sensu* que ce qu'elle est capable d'énumérer ou de définir analytiquement (Passeron, 1991, p. 367) " » (Bourdeloie 2013).

Même si cette citation fait une part belle aux termes originellement employés par Jean-Claude Passeron, la réflexion de ce dernier me semble ici plutôt assez mal rendue, car ce n'est pas une concession,[345] mais bien plutôt une affirmation et une revendication forte de cet auteur.

L'éclipse de Jean-Claude Passeron dans l'épistémologie des humanités numériques s'explique par de nombreuses raisons :

- L'absence d'une école passeronienne et, donc, une méconnaissance de cet auteur ;
- Une épistémologie des humanités numériques plus tournée historiquement vers le post-structuralisme et le post-modernisme ;
- Enfin, il est vrai que des propositions de Jean-Claude Passeron (2006, *1991*) peuvent sembler mettre trop en avant une dichotomie SHS / sciences modélisantes et expérimentales, rebutante dans une première approche pour des chercheurs en humanités numériques…

Toutefois, en plus de la distinction *mathesis* / *aesthesis* précédemment explicitée de Johanna Druckers (*cf.* section Chap12.I.2.b), il faut souligner l'existence de nombreuses autres réflexions qui sont proches du positionnement de Jean Claude Passeron. Par exemple, Sébastien Broca (2016) met en avant l'insuffisance d'un programme de recherche réduit à la programmation informatique et insiste sur la nécessité de l'argumentation, ainsi que de la narration, dans les SHS. Il précise que « l'argumentation est ce qui empêche de réduire le travail des sciences humaines à un gigantesque *problem solving* » (Broca 2016, 15). Face à ces réflexions, les réflexions de Jean-Claude Passeron permettent d'appréhender ces enjeux avec un peu plus de profondeur historique et de recul épistémologique. Cette formalisation

[345] *Cf.* le premier verbe de la citation précédente : « J-C Passeron concède… » (Bourdeloie, 2013).

en termes de « recul » me conduit à aborder la notion de « surplomb épistémologique » que l'épistémologie des humanités numériques m'a aidé à approfondir via les réflexions d'Aurélien Berra.

5. Aurélien Berra : analyses du « surplomb épistémologique »

Aurélien Berra est historien, spécialiste de l'Antiquité et des humanités numériques. Lors de la conférence inaugurale du colloque *DHnord* en 2014 intitulée *Connaitre aujourd'hui. L'épistémologie problématique des humanités numériques*, il a explicité une conception initiale que l'on peut avoir de l'épistémologie et qu'il compare au célèbre tableau du peintre romantique allemand Caspar David Friedrich : *Le voyageur contemplant une mer de nuages.*

Figure n°29 : *Le voyageur contemplant une mer de nuages* de Caspar David Friedrich

Aurélien Berra illustre ainsi une conception de l'épistémologue comme une position de domination, ayant une vue globale des théories, un homme du « meta » (Berra 2014) s'étant élevé au niveau de la science de la science. Comme il l'affirme, beaucoup de chercheurs en humanités numériques ne poursuivent pas ce rêve de surplomb de l'épistémologie soutenu par l'idéal d'une maitrise des diverses manières de faire science. La métaphore du bateau de Neurath, déjà employée (*cf.* section Chap7.I.2), qui insiste *a contrario* sur le recul limité du fait de l'impossibilité de s'extraire du monde, offre une image et une perspective opposées.

Même si les modalités pratiques et le positionnement épistémologique de mon travail doctoral sont plus proches du bateau de Neurath que du voyageur contemplant une mer de nuages, une critique possible est de considérer que mon rendu propose *in fine* à travers les réflexions de Jean-Claude Passeron (2006, *1991*) une vue surplombante. Une question sous-jacente à cette critique est celle de la place centrale donnée à cette référence. Ce choix provient de ma problématique et du constat d'une connaissance réduite − en géographie comme dans les humanités numériques − des propositions de Jean-Claude Passeron, mais ces dernières ne sont évidemment pas considérées comme l'alpha et l'oméga de toute réflexion épistémologique.

Toutefois, il est nécessaire de reconnaitre que la conclusion de son ouvrage *Le raisonnement sociologique* (2006, *1991*), structurée sous forme de propositions et de scolies formant un système théorique, peut être vue comme une grille de lecture globale, et par là même surplombante, du travail en SHS. Par rapport à cette remarque, il est important de préciser que, chez Jean-Claude Passeron, la pratique de l'enquête étant spécifique à chaque cas d'étude, il n'existe pas de mode d'emploi. Si de grandes lignes théoriques peuvent être définies, c'est, *in fine* toujours localement, par le bas, que se construisent les recherches en SHS. La double non-clôture des expérimentations et des interprétations rend toute vue surplombante partielle et inachevée.

De manière plus concrète, suite à la présentation de mon travail au séminaire EGHO à la fin de l'année 2019, cette critique de position de surplomb, de juge, a été formulée, associée à d'autres arguments. Mon travail, en donnant trop d'importance à la lecture kuhnienne de l'histoire de la géographie proposée par Olivier Orain, l'aurait essentialisé à tort. Marie-Claire Robic a ainsi énoncé l'idée qu'il serait préférable de considérer le modèle kuhnien comme un schème d'intelligibilité parmi d'autres. Je ne peux qu'agréer sa proposition et dans cette optique, il me semble qu'il faille considérer et formuler les limites de ce schème d'intelligibilité avec autant d'attention que ses apports. Or, mes recherches m'ont conduit à constater qu'il y avait un gros travail de la part d'Olivier Orain pour cumuler tous les indices favorables à une lecture kuhnienne associée à ce que je considère comme un très faible développement des points critiques. Par rapport à cette situation, mon travail vise à un rééquilibrage que je considère comme intellectuellement légitime. En outre, mon positionnement non relativiste revendique que tous les schèmes ne se valent pas. Des passages de l'un à l'autre, au gré des besoins et des envies, sans analyse de leurs différences et de leurs oppositions, n'est pas non plus une voie à laquelle j'adhère.

Enfin, une autre entrée par rapport à cette notion de « surplomb » concerne les figures produites dans cette thèse. En effet, ces dernières peuvent donner l'impression d'être des cartes sémantiques offrant une vue surplombante pour le lecteur. Ce qui est ici en jeu relève en partie de la conscience, plus ou moins grande, des processus de construction de ces figures. Si cette thèse a fait une grande place à l'explicitation de ces constructions, il existe toujours une tension entre les affirmations de résultats, les nuances critiques et des déconstructions plus radicales. Ainsi, les tensions entre réalisme / constructivisme, précédemment développées (*cf.* section Chap8.I), se doivent également d'être examinées pour les humanités numériques.

6. Réalisme et constructivisme dans les humanités numériques

William McCarty, tout en affirmant le contexte constructiviste de naissance des humanités numériques, n'omet pas totalement de montrer une tension avec une perspective réaliste :

> « Nous pouvons trouver une profonde parenté dans l'idée complexe et constructiviste selon laquelle, pour dire les choses crûment, la connaissance scientifique est à la fois trouvée et fabriquée. Le modèle de science expérimentale de Hacking, dans lequel l'enquêteur rend réelles des entités hypothétiques en apprenant à les manipuler, nous est particulièrement utile. »[346] (McCarty 2003, 1233).

Le réalisme exposé reste ici très théorique n'étant qu'une conséquence des « manipulations ». L'appui sur les travaux de Ian Hacking est utilisé pour reprendre le modèle et le vocabulaire des sciences expérimentales. La tension réalisme / constructivisme développée par ailleurs par Ian Hacking n'est pas reprise et le grand cadre du constructivisme est adopté sous l'argument d'une « profonde parenté ». Par rapport à cette affirmation rapide, il est possible de trouver des témoignages relatant des situations en pratique beaucoup plus complexes. Tim Hitchcock explicite par exemple dans un billet de blog[347] les dérives positivistes qu'il a remarquées en travaillant avec d'autres chercheurs, archéologues, sur un projet d'humanités numériques :

> « le fait d'être confronté à des gens heureux de définir un point sur la surface de la Terre par trois simples chiffres, et de prétendre qu'il en a toujours été ainsi, a été un choc. Cela ne veut pas dire que les archéologues étaient naïfs, loin de là, mais du fait de ma formation d'historien du texte - essentiellement un critique textuel - j'ai été

[346] Traduction personnelle de : « we may find deep kinship in the complex, constructivist idea that, to put the matter crudely, scientific knowledge is both found and made. Particularly useful to us is Hacking's model of experimental science, in which the investigator makes hypothetical entities real by learning how to manipulate them » (McCarty 2003, 1233).

[347] http://historyonics.blogspot.com/2013/12/big-data-for-dead-people-digital.html (consulté le 18/09/2023).

> confronté, lors de ces réunions, à l'existence d'un autre type de connaissance »[348] (Hitchcock 2013).

Tim Hitchcock développe ensuite sa position constructiviste en opposition à celle des archéologues. Si je conçois tout à fait la difficulté d'une rencontre entre ces positions épistémologiques (et c'est aussi ce qui s'est passé dans la géographie des années 1970-1980 et ce qui continue parfois encore d'avoir lieu), mes analyses de thèse me conduisent à penser que cette tension réalisme / constructivisme s'incarne aussi à des échelles plus fines. Beaucoup de travaux d'humanités numériques sont pris entre la volonté d'affirmer leurs résultats et celle de reconnaitre tous les ressorts de leurs constructions. Affirmer la fin du réalisme ou penser que celui-ci reste attaché seulement qu'à quelques chercheurs archaïques, est, à mon sens, une erreur d'appréciation, similaire à celle de la thèse d'Olivier Orain. L'objectif de faire correspondre leurs discours au(x) monde(s) reste un objectif partagé par bon nombre de chercheurs avec, il faut évidemment le reconnaitre, des conceptions plus ou moins critiques par rapport à la réalisation effective de cet objectif.

Dans le contexte actuel, cette tension réalisme / constructivisme est de plus renforcée par les données massives avec l'utilisation de méthodologies accentuant l'importance de la *mathesis* (*cf.* section Chap12.2.b). Ceci permet de comprendre pourquoi le chapitre suivant examine une partie des réflexions épistémologiques qui anime la sphère des données massives même si cette thèse n'en fait pas directement partie.

[348] Traduction personnelle de : « Confronted by people happy to define a point on the earth's surface as three simple numbers, and to claim that it was always so, was a shock. This is not to say that the archaeologists were being naïve, far from it, but that having been trained up as a text historian - essentially a textual critic - in those meetings I came face to face with the existence of a different kind of knowing » (Hitchcock 2013).

Chapitre 13 :
Les données massives

Aborder le champ des données massives conduit dans un premier temps à se heurter au problème de leur définition. Après une première approche générale, j'ai choisi d'ajouter trois études de cas afin d'étudier des exemples concrets dans différents domaines. Enfin, des analyses de trois chercheurs (Boris Beaude, Rob Kitchin et Bruno Bachimont) sont plus particulièrement développées, car les rapprochements de leurs réflexions avec mon travail doctoral m'ont semblé assez intéressants pour être présentés.

I. Un problème de définition

Ce problème complexe de la définition des données massives est ici abordé en prenant comme appui principal la thèse d'Eglantine Schmitt (2018). Son travail s'appuie sur l'analyse d'un corpus de 107 définitions existantes et aboutit à l'identification de quatre grandes familles définitionnelles :

- La première se fonde sur les caractéristiques des données elles-mêmes. L'appellation de « données massives » met au premier plan l'importance de la quantité, même s'il n'existe pas de seuil permettant de dire à partir de quelle quantité de données il est légitime d'utiliser cette appellation. De plus, la définition la plus couramment employée excède le critère du nombre avec la formule des « 3V » : « Volume, Variété, Vélocité » (Laney 2001) :

- La variété renvoie autant à la provenance des données (issues de la numérisation / nativement numériques, issues de contenu web, de réseaux sociaux, de capteurs, de bases de données…) qu'à leurs formats (texte/audio/image/vidéo…).
- La vélocité désigne quant à elle la fréquence à laquelle les données sont produites, mais aussi partagées et traitées.

En fait, cette définition par les « 3V » ne résout en rien le flou définitionnel des données massives puisqu'elle ajoute seulement deux critères sans aucun seuil discriminant. À la suite de Doug Laney, plusieurs autres chercheurs ont ajouté d'autres critères : exhaustivité (Mayer-Schönberger 2013), finesse de la granularité (Dodge et Kitchin 2005), caractère relationnel (Boyd et Crawford 2012), extensibilité et capacité à passer à l'échelle (Marz et Warren 2015), véracité (Marr 2014)… Rob Kitchin et Gavin MacArdle (2016) ont étudié ces caractéristiques sur 26 jeux de données considérés comme massifs. Leur conclusion est que les corpus sont hétérogènes vis-à-vis de cet ensemble de critères. Pour ces auteurs, la « vélocité » et l' « exhaustivité » sont les caractéristiques les plus importantes. Néanmoins,

la position d'Eglantine Schmitt (2018) sur la question, substituant le terme d'« indice »[349] à celui de « caractéristique » ou de « critère », rend mieux compte du flou définitionnel et des multiples propositions sur ce sujet.

Par rapport à la recherche quantitative précédemment menée (*cf.* Part2), l'argument pour exclure la recherche effectuée du champ des données massives concerne le nombre d'articles constituant le corpus, mais aussi le processus de construction du sujet. En effet, la prise en compte des difficultés liées à l'hétérogénéité et la représentativité des données a précédemment conduit à une problématisation plus précise et à la réduction importante du nombre de revues travaillées (*cf.* section Chap2.II.5.a). Cette construction s'oppose aux dynamiques de recherche caractérisant les « données massives », marquées par un mouvement inverse avec la multiplication grandissante du nombre de données traitées. Cet argument relève ici de la méthodologie, ce qui conduit à présenter la deuxième famille de définitions identifiée par Eglantine Schmitt.

- Celle-ci, centrée sur les outils et les méthodes, est composée par des techniques de collecte (numérisation, aspiration de données…), de stockage (entrepôt, lac de données[350]…) et de traitement (apprentissage automatique[351] avec plusieurs variantes possibles : supervisé / non supervisé / semi-supervisé, profond[352] ou non…). Une grande partie de ces outils est utilisée pour travailler sur des données massives, mais aussi sur des plus petits corpus, renouvelant alors le problème d'un critère de délimitation. Certes, il faut souligner que le fait de travailler avec beaucoup de données conduit parfois à des changements techniques et méthodologiques non négligeables. La question du « passage à l'échelle » est une problématique classique de développement d'applications ou de logiciels pouvant s'avérer complexe. Il est vrai aussi que le fait de devoir stocker et traiter des données, non plus sur un seul ordinateur mais en utilisant des ressources distribuées, conduit également dans certains cas à des modifications dans les pratiques de recherche.

---

[349] Le terme peut évidemment faire penser au « paradigme indiciaire » de Carlo Ginzburg. Ce dernier emploie le terme de paradigme « sans tenir compte des précisions et distinctions que l'auteur T Kuhn a introduites dans sa "postface" » (Ginzburg 1980, 3). L'approche développée par Carlo Ginzburg est finalement bien plus proche des développements de Jean-Claude Passeron que ceux de Thomas Kuhn. En effet, les disciplines indiciaires sont définies comme ayant « pour objets des cas, des situations et des documents individuels, et c'est précisément pour ce motif qu'elles atteignent des résultats qui conservent une marge aléatoire irréductible » (Ginzburg 2010, 250). Cette mise en avant de la spécificité des cas ne permettant pas l'établissement de lois universelles, rentre tout à fait en résonance avec les développements de Jean-Claude Passeron dans *Le Raisonnement sociologique* (2006, *1991*).

[350] *Data lake* en anglais.

[351] *Machine learning* en anglais.

[352] *Deep learning* en anglais.

Un aspect important du développement méthodologique de la dernière décennie est lié à l'apprentissage profond mobilisant des réseaux de neurones. En effet, la croissance de la quantité de données à l'intérieur des corpus[353] est allée de pair avec l'augmentation du nombre de neurones utilisés dans les architectures d'apprentissage[354]. Ces deux dynamiques se renforcent mutuellement : l'augmentation du nombre de données en entrée améliore les résultats obtenus par les réseaux de neurones et, inversement, plus les réseaux de neurones sont utilisés, plus il existe une demande importante pour construire et disposer de jeux de données massifs.

Dans ce cadre, une étude de Dominique Cardon *et al.* (2018) met en lumière, à partir d'une analyse des publications du *Web of Science*[355] dans le domaine de l'intelligence artificielle, un changement majeur depuis le milieu des années 1990. Avant cette période, l'intelligence artificielle dite « symbolique » dominait, et ce, depuis la fin des années 1970. Elle se caractérise, pour un problème donné, par la construction en amont de règles avec l'aide d'experts du sujet. Ces règles sont ensuite implémentées informatiquement. Le système expert ainsi créé permet de produire facilement des réponses sur le problème en question. *A contrario*, l'intelligence artificielle, aujourd'hui prééminente, est dite « connexionniste ». Elle repose sur l'utilisation de réseaux de neurones qui prennent en entrée avec les données, non pas les règles à appliquer comme dans un système expert, mais les résultats attendus. Ce qui est produit par le réseau de neurones est une optimisation appelée « apprentissage ». Ce dernier permet ensuite pour une nouvelle donnée de prédire le résultat le plus probable en fonction de l'apprentissage réalisé. La démarche est considérée comme beaucoup plus « inductive », car les règles ne sont pas fixées à l'avance. De plus, ce qui est appris ne ressemble pas à des règles, mais à une optimisation mathématique difficilement interprétable, ce qui constitue un changement notable en termes de production de connaissances.

Un épisode marque une rupture relatée par un chercheur dont les propos ont été anonymisés dans l'article de Cardon *et al.* (2018). Il revient sur une compétition

---

[353] Par exemple, pour les corpus d'images : *MNIST* (*Modified National Institute of Standards and Technology database*), une référence datant d'avant les années 2000, contient 60 000 images alors qu'*ImageNet* atteint en 2016 plus de 10 millions d'images annotées. Pour les corpus textuels, une forte croissance existe aussi comme en témoigne le dernier corpus *C4* (*Colossal Clean Crawled Corpus*) avec 1,1 billion de documents dans sa version la plus massive, mais aussi la moins nettoyée.

[354] *GPT-3* possède par exemple 175 milliards de paramètres, 100 fois que *GPT-2* sorti un peu plus d'un an auparavant.

[355] Le *Web of Science* est une plateforme majeure de données bibliographiques multidisciplinaires. Elle est produite par la société *Clarivate Analtics*. Elle est très utilisée notamment du fait de sa production d'indicateurs bibliométriques. Elle est accessible à l'adresse suivante après inscription : https://clarivate.com/webofsciencegroup/solutions/web-of-science/ (consulté le 15/03/2021).

internationale annuelle de reconnaissance automatisée d'objets sur des images (*European Conference on Computer Vision*) :

> « Alors à la compétition de 2012, qui débarque ? C'est Hinton [le "père" du renouveau des réseaux de neurones] et c'est le séisme. Il ne connaît rien au domaine de la vision par ordinateur et il prend deux petits gars pour tout faire sauter ! Un [Alex Krizhevsky] qu'il a enfermé dans une boîte et il lui a dit : "Tu ne sors pas tant que ça ne marche pas !" Il a fait tourner des machines énormes, qui avaient des GPU à l'époque qui n'étaient pas ultra, mais qu'il faisait communiquer entre eux pour les booster. C'était un truc de machinerie complètement dingue. Sinon, ça n'aurait jamais marché, un savoir-faire de geek, de programmation qui est hallucinant. À l'époque, les mecs de computer vision s'excitaient sur ImageNet depuis deux, trois ans [*une base de données de 1,2 million d'images étiquetées avec 1 000 catégories servant de benchmark pour comparer les résultats en classification des différents compétiteurs*]. Le number one, il était à 27,03 % d'erreur, le number 2 à 27,18 %, le number 3 à 27,68 %. Et Hinton, il envoie son mec sorti de nulle part : "on a fait tourner un gros deep, on est à 17 !". Il met 10 points à tout le monde ! Comme ça, le jeune geek, il arrive, il annonce le résultat, la salle bondée à craquer. Enfin, il comprend rien à rien, genre il a 17 ans ! Il ne sait pas pourquoi les trucs sont là. Lui, il était enfermé dans sa boîte, il ne connaissait rien au domaine. Et là, il est face à Fei Fei ! Et tu as LeCun qui est assis au fond de la salle qui se lève pour répondre aux questions [*Li Fei Fei, professeur d'informatique qui dirige SAIL le laboratoire historique d'intelligence artificielle de Stanford ; Yann LeCun, aujourd'hui directeur de FAIR, le laboratoire d'intelligence artificielle de Facebook et un des acteurs centraux du renouveau des réseaux de neurones*]. Et tu as tous les grands manitous du computer vision qui essayent de réagir : "Mais en fait c'est pas possible. Ça va pas marcher pour la reconnaissance d'objet quand il faut…" Enfin, les mecs étaient tous par terre parce que grosso modo cela foutait en l'air 10 ans d'intelligence, de *tuning*, de sophistication. C'est pas forcément des gens qui font de la logique formelle, mais c'est des gens qui sont quand même dans cette idée qu'il faut comprendre, qu'il faut savoir expliquer pourquoi on met les branches comme ça et qu'on raisonne comme ça, et que l'on avance comme ça, et qu'il faut toute cette intelligence des *features* qui va avec et qui aide à dire que l'on comprend parfaitement ce que l'on fait et que l'on sait pourquoi c'est là. Et le mec il arrive avec une grosse boîte noire de deep, il a 100 millions de paramètres dedans, il a entraîné ça et il explose tout le domaine ». (cité par Cardon, Cointet, et Mazières 2018, 175).

Ce récit doit être évidemment considéré comme un témoignage personnel tout empreint de la subjectivité de son auteur, mais il rend compte d'un changement, qui, tout en s'ancrant dans un historique des pratiques (les méthodes à base de réseaux de neurones sont bien antérieures à 2012), marque une évolution importante. Plus globalement, comme le soulignent Dominique Cardon, Jean-Philippe Cointet et Antoine Mazières : « depuis 2010, domaine après domaine, les réseaux de neurones profonds provoquent la même perturbation au sein des communautés informatiques traitant du signal, de la voix, de la parole ou du texte » (*Ibid* 2018, 176). À partir d'une telle observation, la question d'un changement de paradigme peut être posée, avec notamment ce développement important de méthodologies considérées comme plus inductives. Dans son travail de doctorat, Eglantine Schmitt déconstruit toutefois fortement la thèse de l'inductivisme qui entretient « un rapport plus

complexe avec les pratiques des *big data* que ne le laissent entendre les discours mythologiques associés. Comme la corrélation, l'exhaustivité, l'empiricité, il s'agit davantage de traits projetés et espérés » (Schmitt 2018, 65).

Au-delà de ces deux premières approches mettant l'accent sur des dimensions techniques (en se basant sur les données pour la première et sur les outils pour la seconde), il est possible de privilégier une focale orientée sur les dimensions sociales. Ce choix conduit à la troisième famille de définitions présentée par Eglantine Schmitt.

- Cette troisième famille est centrée sur les usages et les acteurs. Ces derniers dépassent largement le cadre de la recherche scientifique, avec une importance majeure de la sphère industrielle (production et communication), et notamment des GAFAM[356]. D'ailleurs, il faut souligner que les discours sur un changement de paradigme lié aux données massives proviennent à l'origine d'auteurs qui ne font pas partie de la sphère universitaire. La référence la plus connue, *The end of theory : the data deluge makes the scientific method obsolete* (2008), est un article, non pas de recherche, mais issu d'un magazine, écrit par Chris Anderson, alors rédacteur en chef du journal en question (*Wired*). Une autre référence célèbre, *The Fourth Paradigm* (2012) a été écrit par Tony Hey, Steward Tansley et Kriskin Tolle, alors à des postes de direction dans le département recherche de l'entreprise Microsoft. Il faut ici souligner que si la sphère universitaire a, en grande partie, critiqué le discours naïf d'une fin de la théorie, plusieurs de ses travaux ont aussi développé la perspective d'une révolution scientifique liée à l'utilisation des données massives (Mayer-Schönberger 2014; Domingos 2015).

Distinguer les mondes universitaires et non universitaires pose évidemment le sujet complexe de leurs interactions : il existe indéniablement de la compétition entre les laboratoires des universités et des entreprises, mais aussi des coopérations (García Robles *et al*. 2021). Il serait toutefois dommageable d'effacer cette distinction, car la question de la propriété, privée ou publique, des découvertes, reste majeure. Elle peut avoir des impacts non négligeables sur les dynamiques de production de connaissances. Enfin, ces dynamiques peuvent être également appréhendées à travers la quatrième famille de définitions des données massives, mise en avant par Eglantine Schmitt et privilégiant comme critère les changements culturels.

- Cette dernière famille participe conséquemment à une production de travaux et de discours en termes de « changement de paradigme ». Elle dépasse largement le cadre de la

[356] GAFAM est l'acronyme de plusieurs grandes compagnies du Web : *Google* (*Alphabet*), *Apple*, *Facebook* (*Meta*), *Amazon* et *Microsoft.*

recherche scientifique. Les données et les algorithmes prennent de plus en plus de place dans notre vie quotidienne (sélections des informations dans les médias, diagnostic médical…) et font l'objet de beaucoup d'espoirs et d'inquiétudes (surveillance des individus, perte de contrôle des régulations…). Ce faisant, il devient particulièrement difficile de démêler ce qui relève des pratiques de leurs fantasmes. Comme l'affirme Eglantine Schmitt, en abordant cette quatrième famille, « les contours déjà flous de la définition se noient davantage » (2018, 28).

Cette approche des données massives permet de déplier plusieurs dimensions de cette dynamique multiforme (technique, sociale et culturelle). Les difficultés rencontrées pour approfondir cette recherche définitionnelle m'ont conduit à compléter cette approche par des analyses d'études de cas. La section suivante en présente trois : une première ancrée dans les sciences naturelles, une deuxième axée sur un retour d'expériences allant des sciences expérimentales aux SHS, et enfin, une troisième plus orientée sur les sciences du texte.

II. Mises en perspectives variées : trois études de cas

1. *AlphaFold2 : un changement de paradigme ?*

Ce premier cas revient sur la mise au point par la société *DeepMind*, filiale de *Google*, d'un programme nommé *AlphaFold2*, qui s'attache à une problématique spatiale : la détermination des structures tridimensionnelles de protéines à partir de leur composition. Ce programme, mis au point en 2020, s'attaque ainsi à un sujet de biologie moléculaire existant depuis plusieurs décennies. À partir de deux bases de données (*Protein Data Bank* qui contient environ 180000 structures et *UniProt* qui contient 200 millions de protéines), un système d'apprentissage automatique a été construit pour déterminer la structure tridimensionnelle la plus probable d'une protéine à partir de la connaissance des éléments la composant. Les résultats obtenus marquent une avancée importante. En effet, depuis plus d'une dizaine d'années, tous les systèmes créés pour résoudre ce problème, faisant l'objet d'une compétition[357], ne dépassaient pas un taux de réussite 40 %. La première version d'*AlphaFold* en 2018, qui s'appuyait sur un système d'apprentissage à base de réseaux neuronaux utilisant la convolution, approche les 60 %. En 2020, *AlphaFold 2*, qui s'appuie toujours sur un système d'apprentissage à base de réseaux neuronaux, mais utilisant une

[357] Cette compétition se nomme *CASP* (*Critical Assessment of protein Structure Prediction*) : https://predictioncenter.org/ (consulté le 15/03/2021).

autre méthode que la convolution, celle des transformeurs[358], dépasse les 80 %. La résolution n'est pas totale, mais il s'agit pour ce secteur d'une avancée considérable. En effet, quand il fallait auparavant plusieurs mois pour avoir la structure d'une protéine, quelques minutes ou heures suffisent maintenant pour l'obtenir avec des résultats suffisamment fiables et précis pour être exploitables.

Il est intéressant de souligner que la société *DeepMind* a révélé plusieurs points de son système d'apprentissage (Callaway 2021), dont certains sont contre-intuitifs pour des biologistes moléculaires. Par exemple, le programme casse la protéine en plusieurs morceaux pour prédire l'agencement de chacun sans tenir compte des liaisons entre eux dans un premier temps. Ce n'est que dans un second temps que les liaisons entre les différents morceaux sont effectuées. Les structures trouvées sont exactes dans environ 80 % des cas, mais personne ne sait précisément expliquer pourquoi. La situation actuelle se caractérise donc par des modèles qui trouvent des résultats bien meilleurs que tous ceux qui existaient auparavant, mais les règles de structuration des protéines ne sont pas vraiment connues. Il faut bien comprendre que les chercheurs qui se trouvent valorisés dans ce cadre travaillent dans l'intelligence artificielle et sont, donc différents, des chercheurs qui l'auraient été dans une recherche à base de construction de règles (donnant un rôle central aux connaissances en biologie). Toutefois, il est loin d'être certain que les erreurs *qu'AlphaFold 2* n'a pas résolues puissent l'être sur la base de simple amélioration de ce modèle. Dans ce cadre, les chercheurs en biologie ont encore toute leur place (ne serait-ce que pour comprendre les erreurs restantes), mais il est certain qu'ils doivent composer avec cette nouvelle situation.

Ce changement doit-il se lire comme un changement de paradigme au sens kuhnien du terme ? Il me semble ici que l'évolution est un peu trop contemporaine pour répondre à cette question. Je fais l'hypothèse qu'un temps de recul et de mise à distance est nécessaire pour savoir par exemple si les étudiants en biologie moléculaire sont entraînés sur des exercices-types qui n'existaient pas auparavant. Il faudrait également enquêter pour comprendre si les valeurs et la métaphysique de ce domaine, au sens où Thomas Kuhn les définit (2008, postface de 1969), sont en train de changer. Il existe sans aucun doute une effervescence liée au début de résolution de ce problème, mais cela ne signifie pas pour autant qu'il y a un changement de paradigme kuhnien. Ces précisions rendent compte de l'importance de revenir à un examen précis des éléments de la matrice originelle. Des changements

[358] Un transformeur est un modèle particulier d'apprentissage profond de type *seq2seq* (séquence à séquence) basé sur le mécanisme d'attention. Cette thèse n'a pas pour objectif d'expliciter cette méthode. Pour la situer par rapport aux méthodologies utilisées, *BERT* repose sur cette technologie dans le domaine du Traitement Automatique du Langage.

importants peuvent toucher des sciences naturelles sans être obligatoirement des changements de paradigme tel que Thomas Kuhn les a définis.

La section suivante présente un retour d'expériences partant des sciences formelles et expérimentales, avec l'ajout ensuite de données individuelles et *in fine* un essai d'utilisation de données massives sur de la prédiction politique.

2. Des sciences « dures » aux sciences « très dures »[359]

Ces retours d'expérience ont été décrits par Aurélie Jean, mathématicienne de formation, dans *De l'autre côté de la Machine : Voyage d'une scientifique au pays des algorithmes* (2019).

Tout d'abord, pendant son doctorat sur la morphologie des nanoparticules, cette chercheuse parvient, après avoir élaboré quelques hypothèses simplificatrices, à écrire un algorithme permettant de simuler des nanostructures numériques. Elle revient sur les décisions nécessaires vécues comme des sacrifices, notamment pour calibrer approximativement l'algorithme réalisé. Une étape de validation avec différents tests permet d'obtenir non pas une correspondance parfaite entre les simulations numériques et le « réel », mais des résultats tout de même « impressionnants » avec de « fortes similarités » (Jean 2019, 81).

Ensuite, Aurélie Jean travaille sur d'autres thèmes de recherche, notamment un, portant sur l'évaluation du risque, chez les participants d'un marathon, d'être victime d'un traumatisme crânien dû aux ondes de pression suite à l'explosion de deux bombes. Il faut prendre en compte la position des marathoniens et l'ensemble du mobilier urbain pour modéliser cette situation. Malgré un important travail réalisé et des résultats obtenus, le chef du projet décide de ne pas publier les résultats qui pourraient être utilisés par les assurances et avoir des effets induits non maitrisés. Cette situation illustre parfaitement ce que Ian Hacking appelle un « effet de boucle » (2008, 158), notion qui s'applique également aux concepts. Par exemple, la manière dont est conçu le concept d'« handicapé » a des effets sur le monde social. *A contrario*, un concept comme celui de « cuesta » ne peut pas modifier le monde physique qu'il décrit. Cette différence essentielle, explique pour Ian Hacking l'existence d'épistémologies et de styles de recherche différenciés.

Enfin, une dernière expérience que mentionne Aurélie Jean est son essai pour développer un outil numérique capable d'évaluer la stabilité politique d'un pays à partir de l'analyse de l'activité des réseaux sociaux. Le problème d'accès aux données, de nombreux biais et le

[359] En reprenant ici un jeu de mots d'Herbert Simon.

problème des conséquences que peuvent engendrer les résultats, l'ont conduit à abandonner cette recherche. Il n'est pas question de tirer ici des conclusions d'un tel abandon. Il est en effet possible de construire en SHS des modèles locaux qui ont des résultats satisfaisants. Toutefois, plus la complexité augmente, plus il est difficile de construire et de se servir de modèles adéquats. En imaginant qu'un tel modèle existe et soit reconnu, il est évident que les pays auraient tout intérêt à publier à l'aide de robots sur les réseaux sociaux des informations susceptibles d'influencer les résultats de l'algorithme. Si ce problème des « effets de boucle » préexiste à l'utilisation des données massives, il est possible de penser que le numérique l'a démultiplié.

Le prochain cas étudié s'inscrit dans une certaine continuité avec ce dernier retour d'expérience d'Aurélie Jean en traitant des sciences du texte.

3. De *Google Books* aux Grands Modèles de Langue

Historiquement, un projet emblématique ayant revendiqué un changement de paradigme[360] a été présenté par Jean-Baptiste Michel *et al* dans un article intitulé *Quantitative Analysis of Culture Using Millions of Digitized Books* (2011). Il s'agit de s'appuyer sur les 5,2 millions de livres numérisés par *Google* pour fournir des analyses quantitatives des évolutions culturelles mondiales. Ce travail est aujourd'hui incarné par l'outil *Google N-gram* qui permet d'avoir une vue quantifiée de l'évolution d'un terme ou d'une expression dans la littérature mondiale numérisée. Les réactions ont été assez vives, notamment par rapport à la vérification impossible des contextes des occurrences pour un chercheur extérieur (les numérisations appartiennent à *Google Books* et ne sont pas directement accessibles). Un point particulièrement intéressant est la présentation de ce travail qui a eu lieu à la réunion annuelle de *l'American Historical Association* (AHA). Le président de cette association, Anthony Grafton[361], a écrit à ce propos un compte-rendu intitulé *Loneliness and Freedom* (2011). Il raconte comment les présentateurs du projet ont été interpellés sur l'absence d'historien dans leur équipe. Leur réponse a été que les historiens et autres humanistes seront trouvés par la suite. Anthony Grafton écrit à ce propos :

> « Apparemment, les historiens n'ont pas établi, aux yeux de nombre de leurs collègues des sciences naturelles, qu'ils possèdent des connaissances spécialisées qui pourraient être précieuses, voire cruciales, même lorsqu'un projet scientifique porte sur la reconstitution d'une partie du passé humain » (Grafton 2011).

[360] La culturonomique étudie les évolutions culturelles et revendique les parallèles avec la génomique (d'où son nom) et les évolutions biologiques.

[361] Anthony Grafton est un historien américain, professeur à l'université de Princeton.

La tension est ici très marquée. Parmi les historiens qui se sont saisis ensuite de la notion de « très grand corpus », Jacques Guilhamou met en avant qu'il existe, selon lui, l'avantage dans ces corpus de pouvoir inclure des informations de contexte dans le corpus lui-même. Il déclare : « le temps du corpus limité, échantillonné, clos, est désormais bien révolu chez les historiens du discours » (Guilhaumou 2002, 39). Cette affirmation mérite d'être nuancée. Il existe évidemment encore des recherches (mon travail doctoral en faisant partie) qui limitent et clôturent leurs corpus.

De plus, l'idée basée sur l'inclusion du contexte dans le corpus est très discutable. Pour Jacques Guilhamou, la numérisation massive et les capacités de traitements permettent de construire des corpus permettant de prendre en compte sur certaines questions tous les éléments concernés. Dans la pratique, la question de l'exhaustivité est toujours délicate, même sur des sujets bien délimités, car les périmètres peuvent toujours être discutés. En outre, les données de contexte ne renvoient pas forcément à des textes existants. Ainsi, quelle que soit la taille d'un corpus, il ne contiendra jamais toutes les données de contexte nécessaire pour le comprendre.

Cela étant précisé, les *Grands Modèles de Langue* (*Large Language Model*) ont connu un essor considérable durant la période où j'ai réalisé ma thèse. Sans rentrer dans les détails, des techniques de plongement de mots sont appliquées à de très grands corpus. Ces méthodes qui étaient au début de ma thèse inconnues du grand public se sont retrouvées sur le devant de la scène avec la sortie de *ChatGPT* en novembre 2022 par la société *OpenAI*. L'augmentation des performances et des usages a été considérable (Dale 2021). Les discours révolutionnaires, marqués autant par les espoirs que les peurs, ont été nombreux (Menissier *et al*. 2023). Sur le plan scientifique, si les nouveaux enjeux sont manifestes (Birhane *et al*. 2023), la continuité avec les méthodes précédentes (Langlais 2023) ne plaide pas pour une révolution kuhnienne au sens strict du terme.

S'il m'est impossible de traiter dans le cadre de cette thèse de manière approfondie du sujet des *Grands Modèles de Langue*, j'aimerais toutefois ici développer une problématique à partir des réflexions de Stephen Wolfram, très marquées par une perspective modélisatrice :

> « L'ingénierie spécifique de ChatGPT l'a rendu très convaincant. Mais en fin de compte (au moins jusqu'à ce qu'il puisse utiliser des outils extérieurs), ChatGPT tire "simplement" un "fil de texte cohérent" des "statistiques de la sagesse conventionnelle" qu'il a accumulées. Mais il est étonnant de constater à quel point les résultats sont humains. Et comme je l'ai expliqué, cela suggère quelque chose qui est au moins scientifiquement très important : le langage humain et les schémas de pensée qui le sous-tendent sont en quelque sorte plus simples et plus "semblables à des lois" dans leur structure que nous ne le pensions. ChatGPT l'a implicitement découvert. » (Wolfram 2023).

Cette affirmation d'un langage humain et des schèmes de pensée qui le sous-tendent comme « semblables à des lois » est particulièrement discutable. Si les *Grands Modèles de Langage* arrivent à répondre correctement à un très grand nombre de questions et à compléter des textes de manière cohérente et vraisemblable (avec aussi parfois des erreurs appelées hallucinations), la seule loi sur laquelle ils reposent est celle d'une optimisation de leurs milliards de paramètres pour minimiser les erreurs de prédiction par rapport à ce qui est déjà connu. La production d'un espace latent, en partie représentatif des différences et des proximités de significations entre les termes d'un grand nombre de langages, ne veut pas dire que nous obtenons une carte permettant d'avoir une vue surplombante des espaces linguistiques, disciplinaires, sociaux et culturels. Il existe d'ores et déjà des milliers de *Grands Modèles de Langue,* et même ceux qui sont spécialisés sur un domaine n'offrent que des espaces latents partiels qui seront sûrement améliorés à l'avenir par d'autres modèles, sans jamais tendre vers une représentation unique et complète.

Pour finir ces mises en perspective, j'ai choisi de m'intéresser successivement aux positionnements de trois auteurs : Boris Beaude, chercheur issu de la géographie française ayant fait un virage vers la sociologie ; Rob Kitchin, géographe britannique et Bruno Bachimont, philosophe et informaticien. Il est vrai qu'il existe beaucoup d'autres chercheurs ayant écrit sur les données massives. Le critère qui a prévalu dans ces choix est l'intérêt de leurs réflexions par rapport aux problématiques de mon travail doctoral, notamment celle d'une critique des approches kuhniennes en sciences sociales.

III. Analyses de trois positionnements de chercheurs

1. Boris Beaude : quand perturbation rime avec changement de paradigme

Boris Beaude, dans un article intitulé *(re)Médiations numériques et perturbations des sciences sociales contemporaines* (2017), fait état de changements sur quatre plans :

- Les relations sociales sont de plus en plus marquées par les médiations numériques qui changent l'espace et l'ensemble de la société.
- L'observation de ces relations sociales peut s'appuyer sur les nombreuses traces numériques laissées par les individus. Toutefois, plusieurs difficultés existent :
    - la représentativité de ces traces est souvent faible ;
    - leur accès n'est pas aisé pour de nombreux sociologues ;

    - elles sont souvent produites dans d'autres contextes de production que celui de la science, impliquant des adaptations et des conversions pouvant être complexes.
- Concernant la production de connaissance, la physique et l'informatique présentent des atouts indéniables pour traiter de ces traces numériques. Les approches relevant d'une « physique sociale »[362] se développent avec comme objectif, non pas « de saisir les causes des faits sociaux singuliers, mais d'en identifier les relations stables, comme autant de déterminations mues par des causes générales et universelles » (Beaude 2017, 98).
- Enfin, les changements se jouent dans la production de sens avec la résurgence des tensions historiques entre « d'une part entre les sciences sociales et les sciences de la nature, d'autre part entre les sciences sociales "positivistes" et les sciences sociales "interprétatives" » (*Ibid* 2017, 84). Le besoin d'une culture hybride des sociologues est mis en avant pour ne pas être exclu des nouvelles voies computationnelles tout en préservant une prudence interprétative et critique.

Toutes ces analyses sont confirmées par mon travail doctoral. *A contrario*, quand Boris Beaude s'appuie sur ces changements pour déclarer que la sociologie est « à l'aube d'un changement de paradigme » (*Ibid* 2017, 93), une divergence dans nos positionnements ne peut qu'être soulignée. L'argumentation de Boris Beaude concernant ce nouveau paradigme reconnaît qu'il s'inscrit à la fois dans l'histoire fournie de la quantification et de l'analyse des réseaux en sciences sociales, mais ce paradigme devrait participer à associer toutes les « composantes mobilisées » de manière plus pleine et partagée. Il est évident que nous sommes ici loin d'une définition kuhnienne en termes de rupture entre deux mondes incommensurables.

Boris Beaude reprend également une grande partie de l'argumentaire de Dominique Boullier sur les sciences sociales de troisième génération (2017) puisque leurs analyses respectives réservent toutes deux une place centrale aux traces numériques. Pour Dominique Boullier, la première génération de sciences sociales est basée sur les recensements avec comme acteur fondateur Émile Durkheim autour du concept de « société(s) ». La deuxième génération s'appuie sur les sondages avec comme acteurs fondateurs Georges H. Callup et Paul Lazarsfeld autour du concept d'« opinion(s) ». Enfin, la troisième génération fait

[362] Cette expression, utilisée par Auguste Comte dans son *Cours de philosophie positive* (1830), est reprise par exemple par Alex Pentland (2014) pour désigner un courant de recherche centré sur des approches mathématiques du monde social.

référence au suivi des traces avec comme acteurs fondateurs Michel Callon, Bruno Latour et John Law. Toutefois, il est reconnu par Dominique Boullier que ces différentes générations co-existent. Au sens kuhnien du terme, il n'existe donc pas de changement de paradigme.

Pour revenir à l'article de Boris Beaude, alors que celui-ci est marqué par une réflexion épistémologique largement référencée, les travaux de Thomas Kuhn et de Jean-Claude Passeron, n'y sont pas abordés. Ceci est d'autant plus étonnant que les tensions mises en avant (sciences sociale / science de la nature, sciences sociales « positivistes » / « interprétatives ») sont au cœur des réflexions de ces auteurs. Mon travail doctoral me conduit à penser que l'usage du terme « paradigme », dans un sens faible comme synonyme de « changement », empêche le développement de toute une partie de la perspective critique, même si l'article de Boris Beaude fait une place importante aux illusions de l'« utopie computationnelle ».

Pour soutenir son argumentation, Boris Beaude s'appuie également sur les réflexions de Rob Kitchin, que la sous-section suivante explore plus en détail.

2. Rob Kitchin : vous avez dit paradigme ?

Rob Kitchin, dans son article, *Big Data, new epistemologies and paradigm shifts* (2014) prend en compte une partie de l'héritage kuhnien et de ses critiques. En effet, il affirme tout d'abord dans une perspective similaire à celle développée dans *La Structure des révolutions scientifiques* (2008, *1962*) que :

> « dans certains domaines académiques, il y a peu de preuves de l'existence de paradigmes, notamment dans certaines sciences sociales où l'on utilise un ensemble diversifié d'approches philosophiques (par exemple, la géographie humaine, la sociologie), alors que dans d'autres domaines, comme les sciences[363], il y a eu plus d'unité épistémologique sur la façon dont la science est menée, en utilisant une méthode scientifique bien définie, étayée par des tests d'hypothèses pour vérifier ou falsifier les théories »[364] (Kitchin 2014, 3).

Il prend d'ailleurs en compte également une partie des critiques formulées à l'encontre de Thomas Kuhn en affirmant que « les comptes-rendus paradigmatiques produisent des histoires trop aseptisées et linéaires sur la façon dont les disciplines évoluent, en aplanissant

[363] Ici, Rob Kitchin assimile "sciences" à sciences expérimentales.

[364] Traduction personnelle de : « within some academic domains there is little evidence of paradigms operating, notably in some social sciences where there is a diverse set of philosophical approaches employed (e.g. human geography, sociology), although in other domains, such as the sciences, there has been more epistemological unity around how science is conducted, using a well defined scientific method, underpinned by hypothesis testing to verify or falsify theories » (Kitchin 2014, 3).

les manières désordonnées, contestées et plurielles dont la science se déroule dans la pratique » (*Ibid* 2014, 3).

Mais, malgré ces positionnements, Rob Kitchin s'appuie sur la notion de paradigme pour analyser la production de connaissances dans le domaine des données massives. Préalablement, plusieurs rappels sont effectués, prenant le contre-pied d'affirmations courantes :

- Les données massives ne peuvent pas capturer un domaine entier et en fournir une résolution complète. Il peut exister une recherche d'exhaustivité, mais ce qui est obtenu (sauf dans les cas de domaines très spécialisés et bien délimités) n'est qu'un échantillon posant la question classique de sa représentativité. Ce rappel est confirmé par mon travail doctoral, et tout le questionnement initial sur le choix des revues (*cf.* section Chap2.I.1).
- Il n'existe pas de recherche sans théorie, modèle et hypothèse *a priori*. Même les algorithmes mettant en avant des méthodes globales plutôt « inductives » reposent sur des présupposés ou des choix assumés qui orientent les recherches.
- Les données ne parlent pas d'elles-mêmes. Il ne suffit pas d'avoir une formation en statistique pour lire et affirmer des résultats. Il existe des risques importants d'interprétations erronées si les chercheurs qui effectuent les analyses des résultats n'ont pas de connaissance des contextes dans le domaine étudié.

Ces rappels conduisent Rob Kitchin à mettre en avant le raisonnement abductif en se référant à C.S Peirce et par conséquent au pragmatisme. Ce positionnement global ne l'empêche pas de distinguer l'utilisation des données massives en sciences naturelles qui « est appelée à transformer l'approche de la recherche » (Kitchin 2014, 7) de leur trajectoire en SHS qu'il qualifie de « moins certaine » (*Ibid* 2014, 7). De plus, au sein des SHS, les usages des chercheurs positivistes et post-positivistes sont différenciés. Pour les premiers, les données massives peuvent être mobilisées pour développer des modèles plus sophistiqués, étendus et précis, prenant en compte davantage des contextes spécifiques. Néanmoins, Rob Kitchin ne cache pas l'existence de tensions dans ce domaine, citant plusieurs travaux de physiciens qui « ignorent souvent délibérément deux siècles de recherche en sciences sociales » (Kitchin 2014, 5) et qui aboutissent à des approches abusivement réductionnistes.

Pour les post-positivistes, Rob Kitchin renvoie au champ des humanités numériques avec des objectifs plus pluriels et de multiples controverses. Mes recherches me conduisent à penser qu'il y a une certaine simplification dans cette assimilation des humanités numériques

à un courant post-positiviste,[365] mais je partage totalement l'affirmation concernant des objectifs pluriels et de multiples controverses.

Face à la tension existante entre positivistes et post-positivistes, Rob Kitchin mobilise comme alternative les épistémologies employées dans les SIG critiques et les statistiques radicales. Ces approches utilisent « des techniques quantitatives, des statistiques inférentielles, la modélisation et la simulation tout en étant conscientes et ouvertes par rapport à leurs lacunes épistémologiques, en s'appuyant sur la théorie sociale critique pour encadrer la façon dont la recherche est menée, le sens donné aux résultats et les connaissances utilisées »[366] (Kitchin 2014, 9). La mobilisation des réflexions de Jean-Claude Passeron dans mon travail doctoral partage cette perspective sans la limiter à un domaine particulier (SIG critique ou autre). La perspective historique plus profonde et plus problématisée de cette tension conduit à transformer toute la prudence de Rob Kitchin en une interrogation profonde : y a-t-il vraiment un changement de paradigme, car il existe avant les SIG critiques tout un développement réflexif des méthodologies quantitatives (Desrosières 2016) ?

Bruno Bachimont, avec ses formations multiples (ingénieur, informaticien, philosophe), est beaucoup plus affirmatif que Rob Kitchin. Sa proposition du nominalisme comme changement de paradigme partage une proximité avec celle d'Olivier Orain, tout en présentant de nombreuses différences.

3. Bruno Bachimont : le retour d'un paradigme nominaliste

Bruno Bachimont développe le positionnement d'une herméneutique matérielle, comme « jeu entre calcul et interprétation, manipulation technique et constitution du sens » (Bachimont et Rastier 1998), positionnement qui le place résolument dans une situation de l'entre-deux, sans devoir nécessairement se réclamer plus d'un pôle que de l'autre. Mais, à la différence de l'entre-deux de Jean-Claude Passeron (2006, *1991*)**,** Bruno Bachimont développe une perspective historique sous forme de changement de paradigme. Pour cet auteur (2017), l'épistémologie de la mesure − Galilée, Descartes et Newton − a abouti à une première révolution, celle d'une nominalisation de la nature. Le langage mathématique a

[365] L'exemple de Tim Hitchcock (*cf.* section Chap12.II.6) montre plus une co-existence de différents positionnements au sein des différents projets.

[366] Traduction personnelle de : « These approaches employ quantitative techniques, inferential statistics, modelling and simulation whilst being mindful and open with respect to their epistemological short-comings, drawing on critical social theory to frame how the research is conducted, how sense is made of the findings, and the knowledge employed » (Kitchin 2014, 9).

rompu l'évidence du lien entre la nature et le langage courant. Nous vivons, selon lui, actuellement une seconde révolution avec une science des données devant aboutir à une nominalisation de la culture. Les arguments fournis font valoir plusieurs ruptures :

> « Une rupture des données par rapport à leur origine et leur nature, le mode de collecte les rassemblant en dépit de leur hétérogénéité ; une rupture du traitement par rapport aux données ; une rupture de ce qui est montré par rapport à ce qui est calculé » (Bachimont 2017, 392).

Cette triple rupture constitue pour Bruno Bachimont un nouveau paradigme reposant sur l'acception d'une dose d'arbitraire au niveau de la collecte, du traitement et de la visualisation. L'arbitraire, vu classiquement comme une faiblesse, permet de réaliser des rapprochements inédits, ainsi que de remettre en cause des interdits et des clivages. Pour Bruno Bachimont, ces changements sont essentiellement dus aux données massives.

Mon travail doctoral me conduit à penser que ce que met en avant cet auteur peut s'appliquer également à de plus petits corpus. Il peut en effet exister aussi des ruptures par rapport à l'origine des données, au traitement et à la visualisation dans des corpus réduits. En allant plus loin, il est possible de se demander si l'épistémologie des SHS, telle qu'elle est présentée par Jean-Claude Passeron (avec en amont la nécessité d'une transformation des données et en aval l'obligation d'un travail interprétatif), ne constitue pas déjà, en grande partie, ce que Bruno Bachimont présente comme une révolution. Ce renversement de perspective permet de penser l'« arbitraire » qui est mis en avant par cet auteur de façon plus fine, plutôt comme un bricolage pour s'adapter à des données et à des problématiques spécifiques.

La proposition de Bruno Bachimont est aussi intéressante par le fait qu'elle utilise le même argument qu'Olivier Orain pour soutenir une perspective paradigmatique, à savoir une sortie par ou vers le nominalisme. Cette thèse a déjà montré que cette porte de sortie était problématique (*cf.* Chap9.VI) et cachait une tension forte qui se trouve confirmée dans le cas de Bruno Bachimont. En effet, dans un débat avec François Rastier, Bruno Bachimont reconnaît un réalisme local. Son argumentation repose sur l'exemple d'un chirurgien qui doit opérer et dont « les connaissances motivant la décision reposent sur un principe de réalité » (Bachimont et Rastier 1998). Quand les données massives sont utilisées de manière appliquée, nul doute que ce principe de réalité se retrouve aussi. Par conséquent, même si Olivier Orain et Bruno Bachimont placent le nominalisme à des endroits différents pour affirmer ce qu'ils pensent, tout deux, être un nouveau paradigme, ils sont confrontés à la même limite de son incomplétude et d'une résistance du réalisme.

Pour finir, les rapprochements, les mises en perspectives et les mises en dialogues que j'ai esquissés démontrent toute l'actualité de la tension – philosophique, scientifique, épistémologique - entre réalisme et nominalisme, permettant de mieux saisir plusieurs enjeux, bien au-delà d'une controverse autour de l'histoire de la géographie française. Malgré toutes les critiques déjà existantes, l'attraction du prisme kuhnien reste grande pour les SHS.

Épilogue

Et pistez mots et derniers tours de piste…

Entre le début et la fin de ma thèse, je suis passé, en reprenant la distinction effectuée par Christophe Lejeune (2019), d'une conception « séquentielle » à une conception « parallèle » de la recherche :

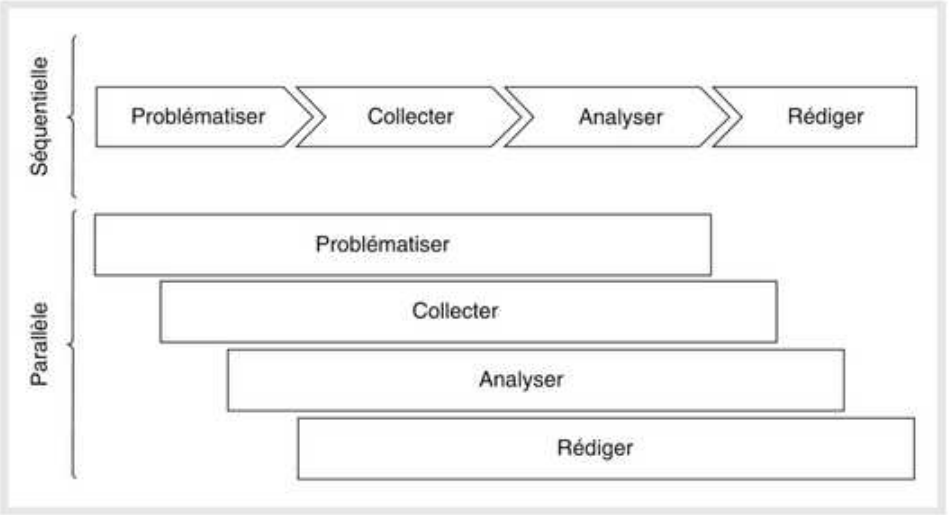

Figure n°27: Conception séquentielle et parallèle de la recherche
(Lejeune 2019, 24).

Il est intéressant de noter que cette figure est tirée d'un manuel d'analyse dit « qualitative ». Sans avoir fait de recherche exhaustive à ce sujet, mes lectures de manuels « quantitatifs » me conduisent à penser que ce domaine privilégie le plus souvent une présentation séquentielle. Rendre compte d'une organisation parallèle est particulièrement complexe, car il n'est pas forcément facile (ni même utile) de présenter tous les allers-retours d'une recherche. Pour autant, ne pas donner l'illusion d'une organisation séquentielle, parfaitement logique, mais totalement reconstruite *a posteriori*, a été une volonté assumée de ce travail doctoral. Cet épilogue, en miroir du prologue, fait place à quelques éléments plus développés qu'une conclusion classique.

Il revient tout d'abord sur les différentes étapes de cette thèse en explicitant à chaque fois quelques éléments synthétiques et réflexifs. Ensuite, des ouvertures sont réalisées dans plusieurs directions. Dans un premier temps, une actualité scientifique en lien avec la problématique de ce travail est développée. Dans un second temps, j'ai fait le choix de revenir sur plusieurs expériences et réflexions provenant des travaux dirigés (TD) que j'ai donnés pendant les deux dernières années de mon doctorat dans le département de sociologie à l'Université de Grenoble. Plus que des ouvertures à proprement parler, il s'agit d'éclairages apportés sur mon sujet de thèse, et ce, à partir d'autres sources de réflexions.

Chapitre 14 :
Retour sur cinq grandes étapes de cette thèse

Ce chapitre reprend successivement les cinq grandes étapes de cette thèse. L'objectif ici est de fournir des derniers moments de synthèses et de réflexions par rapport au travail effectué.

I. L'élaboration de la problématique

La première étape de ce travail a été l'élaboration de la problématique. Cette étape a été déjà largement explicitée et développée dans ce manuscrit (*cf.* Chap2). Cette période intense d'explorations, de développements, d'abandons de certaines pistes, de décantations et de prise de décisions, a été un temps tout à la fois excitant et préoccupant, car il existe à ce moment-là une grande part d'inconnu dans ce qui reste à construire. La difficulté n'a pas été seulement de trouver une réponse adaptée à des attentes universitaires, mais aussi de répondre à mes propres attentes en essayant de trouver une piste assez pertinente, fructueuse et originale. Ces adjectifs témoignent, à mon sens, de l'ampleur de la difficulté, surtout dans un monde de la recherche où en approfondissant son champ, il est assez fréquent de découvrir qu'une voie que l'on croyait originale a déjà été largement travaillée.

Après avoir trouvé la piste qui j'allais suivre, je me souviens d'avoir lu l'ouvrage de Ian Hacking (2008), *Entre science et réalité, la construction sociale de quoi ?*, et d'y avoir trouvé une similitude forte entre ma situation et la présentation qu'il faisait du travail de laboratoire. Il le décrit comme la recherche d' « un ajustement robuste entre l'appareillage, nos croyances concernant l'appareillage, les analyses, les interprétations des données et les théories » (Hacking 2008, 104). Face à cette recherche, pour reprendre les termes de cet auteur, « le monde résiste » (*Ibid* 2008, 104). Si cette réflexion a retenu alors toute mon attention, c'est sans nul doute parce que je venais d'éprouver assez intensément la résistance du monde, mais aussi parce que l'idée d'un « ajustement » entre une question de recherche, un choix de corpus et une orientation méthodologique correspondait bien à ce que j'avais fini par trouver.

Toutefois, concernant mon travail, j'étais et je reste sceptique par rapport à l'adjectif « robuste » que Ian Hacking a choisi pour qualifier l'« ajustement » recherché. Nul doute que certains verront dans cette réflexion une reconnaissance problématique d'une faiblesse des SHS par rapport aux sciences de laboratoire analysées par Ian Hacking. Il me tient à cœur de récuser une telle idée qui évalue le modèle de scientificité des SHS à l'aune des « sciences exactes »[367]. Si j'ai développé dans le corps de ma thèse, une approche sous forme

[367] Sur ce sujet, le livre d'Alain Testart intitulé *Essai d'épistémologie* (1991) est emblématique de cette position. Gerard Lenclud en a fait un bon compte-rendu critique intitulé *La statue du commandeur* (1993) publié dans les *Annales, Sociétés, Civilisation*.

de gradient entre sciences exactes et sciences historiques herméneutiques, plutôt qu'une disjonction des modèles de scientificité comme Jean-Claude Passeron *(cf.* section Chap8.II), il n'existe aucune prééminence d'un pôle en termes de qualité scientifique. Si les SHS se rattachent *in fine* plus à l'histoire qu'à la physique, c'est que la société ne se laisse pas mettre en équations ou alors de manière très imparfaite pour reprendre le titre d'un ouvrage de Pablo Jensen (2018).

Dans les SHS, il me semble toujours nécessaire, comme l'exprime Jean-Claude Passeron, de recontextualiser les concepts et les résultats par rapport aux situations spécifiques qui ont permis de les penser. Cette recontextualisation est aussi possible et intéressante dans les « sciences dures », mais elle n'a pas le même statut. Il est en effet possible de s'intéresser au contexte de formalisation du modèle de la gravité, mais cet historique change *in fine* très peu la compréhension que l'on peut avoir du modèle et de son champ d'application. Les concepts en SHS sont plus intimement liés aux moments et aux lieux de leurs formations et de leurs évolutions.

Au niveau du concept de « paradigme », Ilya Prigogine et Isabelle Stengers (1986) explicitent son caractère partiel et historique. Thomas Kuhn, ayant étudié la physique à l'Université Harvard, a connu cet apprentissage par répétition d'exercices-types qui permet d'assimiler les théories utilisées dans une communauté, mais aussi qui finit par déterminer quel problème est intéressant et quelle solution est acceptable. De plus, c'est une génération qui a été très marquée par la révolution einsteinienne, dynamique qui ne s'est pas ensuite reproduite avec une telle ampleur.

Au niveau de la géographie française, il est nécessaire de prendre en compte l'héritage historique bien développé par Rémy Knafou et Mathis Stock dans le dictionnaire écrit sous la direction de Jacques Lévy et Michel Lussault à la section « Épistémologie de la géographie » :

> « Bien que la production de l'espace ait été au centre de la géographie "science de l'espace", les procédures étaient davantage du ressort des sciences de la nature : recherche de lois, modélisations mathématiques, raisonnement hypothético-déductif, etc. » (Knafou et Stock dans Lévy et Lussault 2003, 324).

Pour reprendre une expression d'Olivier Soubeyran (1997), mon travail doctoral, dans sa critique, touche à un « imaginaire disciplinaire ». Cette affirmation permet de comprendre l'aspect potentiellement vif des controverses entourant ce sujet. En effet, si les polémiques à cet endroit se sont largement atténuées, elles ne sont en rien éteintes[368]. Même si de l'eau

[368] Ma présentation lors du séminaire EHGO en septembre 2019 en est un exemple vécu.

a coulé sous les ponts depuis l'acmé concomitante de l'épistémologie kuhnienne et de la « nouvelle » géographie (décennies 1960 et 1970 en se plaçant à un niveau international), le « modernisme » incarné par la « nouvelle géographie » reste une référence importante, et cela d'autant plus que le « post-modernisme » est loin de convenir à tout le monde. L'alternative réflexive offerte par les développements de Jean-Claude Passeron reste très méconnue d'une grande partie des géographes.

Cette situation va de pair avec une conception de la géographie historiquement très marquée par la dualité géographie classique / « nouvelle géographie ». Même s'il existe de nombreux discours pour nuancer cette dualité, celle-ci se maintient fortement par la sémantique (le « nouveau » s'oppose à l'« ancien »). À cela s'ajoute une histoire de la géographie impulsée par une partie de l'équipe *Épistémologie et Histoire de la GéOgraphie* qui redouble le poids de cette vision traditionnelle avec le maintien, voire le renforcement de cette dualité (même si elle est nuancée sur certains points) du fait d'un choix de lecture qui privilégie le prisme kuhnien et le concept de « révolution scientifique » (Kuhn 2008, *1962*).

La réflexion précédemment engagée sur l'adjectif « robuste » est, à mon avis, un aspect crucial à comprendre. Il est ici important de préciser pour ceux qui connaissent mal les travaux de Jean-Claude Passeron qu'ils ne visent aucunement à enlever les garde-fous de rigueur et de volonté de mise à l'épreuve empirique. Il ne s'agit en rien d'un affaiblissement, mais bien plutôt d'un décalage du modèle de scientificité. Une « moindre robustesse » n'implique pas un « moindre intérêt » dans le modèle de scientificité des SHS. Pour illustrer ce point, il est possible de revenir sur l'hypothèse socio-linguistique développée par Olivier Orain et reprise comme point de départ dans cette thèse.

Cette hypothèse affirme qu'un changement de paradigme s'incarne dans des changements sémantiques majeurs, ce qui est totalement pertinent et en accord avec les réflexions originelles de Thomas Kuhn (*cf.* section Chap2.II.1). Toutefois, il est évident qu'en montrant qu'un changement sémantique majeur s'est produit, on ne démontre pas pour autant qu'un changement de paradigme a eu lieu. Cette assertion est soutenue par la règle logique : « "A implique B" n'implique pas que "B implique A" ». Ainsi, il est possible de rejeter au nom de cette règle logique tout intérêt d'étudier la sémantique pour essayer de saisir un changement de paradigme. Il s'agit de mon point de vue d'une position radicale liée à un raisonnement formel. *A contrario*, les sciences sociales offrent la possibilité d'une telle exploration. Évidemment, une fois qu'un changement sémantique majeur est mis en avant (ou récusé), il existe, pour affirmer l'existence (ou non) d'un changement de paradigme, une zone d'indétermination qui nécessite des suppléments d'investigation non quantitatifs. Cette

réflexion est, à mon avis, un point clé pour comprendre ce travail de thèse, mais aussi plus globalement les incessantes remises en question qui selon Thomas Kuhn marquent les sciences sociales.

Par rapport à cette moindre robustesse, il n'est pas pertinent de penser qu'elle doive être supprimée. Il ne s'agit pas de revendiquer le maintien des zones de flous et d'ambiguïtés, mais, comme le développe Jean-Claude Passeron, ce qui sous le prisme des « sciences dures » apparait comme des erreurs ou des approximations liées à la plasticité du langage naturel, est inexpugnable et nécessaire dans les SHS pour effectuer des rapprochements. Répétons qu'il ne s'agit pas ici de nier l'intérêt des efforts de clarification et de conceptualisation effectués par certains chercheurs. Il ne s'agit pas non plus de dénier la possibilité de construire des simulations géographiques à partir de prémisses formalisées. Il s'agit simplement d'affirmer que ces démarches de conceptualisation et de formalisation mathématique ne sont pas supérieures en termes de pertinence scientifique. Elles ne remplacent pas des approches plus discursives et moins mathématisées.

Ces réflexions font partie de ma problématique qui, même arrivée à ce stade de la conclusion, reste difficile à synthétiser en une ou deux phrases. La présentation la plus efficace que j'ai trouvée est une division en trois sous-problématiques :

- La première est relative à l'hypothèse socio-linguistique d'Olivier Orain. Il s'agit d'étudier le changement sémantique pour donner du poids ou réfuter la pertinence d'une lecture kuhnienne de la géographie française. Si les résultats obtenus sont intéressants sur de nombreux points, il m'a également fallu reconnaitre leurs insuffisances par rapport à ma volonté d'administration de la preuve (concernant ici la présence ou l'absence de paradigme). Cette situation a abouti à ma mobilisation des propositions passeroniennes comme positionnement épistémologique opératoire pour comprendre les difficultés rencontrées. Dans cette perspective, il était cohérent de compléter mon travail quantitatif par des examens non quantitatifs.
- La deuxième sous-problématique examine ainsi de manière critique les analyses de la thèse d'Olivier Orain mobilisées pour justifier la pertinence de sa lecture kuhnienne de la géographie française. Du fait du travail précédent, un examen en parallèle des propositions passeroniennes a été réalisé. Cette étude, prenant en compte une épaisseur historique significative, est à mon sens d'autant plus intéressante qu'elle est reliée à des enjeux contemporains.
- La troisième sous-problématique s'est ainsi attachée à mettre en perspective ce travail par rapport au champ des humanités numériques et aux domaines de recherche liés aux données massives. Ces deux extensions se comprennent par

l'inscription de mon travail dans le champ des humanités numériques et de l'interrogation du modèle de scientificité des SHS par les données massives. Dans ces deux domaines − humanités numériques et données massives − des auteurs revendiquent une ou de(s) révolution(s) scientifique(s) avec des perspectives kuhniennes. Il est donc particulièrement intéressant d'interroger leurs travaux et de les mettre en réflexion avec les analyses précédemment effectuées.

Cette formulation déclinée en trois sous-problématiques est, certes, une façon de contourner la difficulté rencontrée pour les subsumer en une seule problématique. S'il s'agit tout de même de s'astreindre à cet exercice, il me semble que cette thèse explore globalement la question du modèle de scientificité le plus adapté pour effectuer et penser les SHS passées et contemporaines. Pour cela, elle s'interroge sur sa propre trajectoire de scientificité et sur celles, passées et actuelles, empruntées par les chercheurs en sciences sociales.

Cela étant dit, cette formulation rend très partiellement compte du travail concrètement réalisé. Une des raisons de ce décalage provient du fait que ma construction doctorale a pris en charge les points me semblant les plus importants au fur et à mesure du cheminement. Un des risques est de perdre le lecteur qui n'est pas spécialiste de chacun de ces sujets. J'ai bien conscience que certaines parties sont assez massives et difficiles à suivre dans tous leurs détails pour des non-spécialistes. Ce qui pose ici un problème, ce n'est pas l'existence d'une échelle fine, voire très fine d'informations (échelle nécessaire à la majorité des thèses), mais bien la diversité et l'hétérogénéité des entrées. Il y a là une des difficultés à l'exercice de la pluridisciplinarité, complexe à gérer tant dans la production des connaissances que dans leurs restitutions.

De plus, sur certains aspects comme ceux de la préparation des données, il existe, de mon point de vue, dans nombre de travaux beaucoup (trop) d'ellipses par rapport à ce qui a été effectivement effectué. Cette thèse prend délibérément le contre-pied de cette manière de faire.

II. La préparation des données textuelles

La préparation des données constitue la deuxième grande étape de cette thèse. Sa présentation peut être divisée en deux sous-étapes :

- La première a consisté à créer les algorithmes pour améliorer le contenu textuel des articles océrisés de deux revues (Les *Annales de Géographie* et *L'Espace Géographique*) préalablement choisies comme les plus pertinentes par rapport à la problématique retenue. En amont, ces deux revues avaient été numérisées et

documentées par l'UAR *Persée* dans un objectif de consultation Web et non d'analyse textuelle. Le problème principal − améliorer le contenu textuel en vue des traitements envisagés − a été décomposé en de nombreux sous-problèmes (traitement des titres, résumés, mots-clés, figures, notes de bas de page…) qui ont chacun fait l'objet d'explorations, de traitements automatiques et de corrections manuelles les plus systématiques possibles.

- La seconde sous-étape concerne le passage de ces « scripts maison » en un code plus élégant, intégré à une application, permettant de reproduire les différentes étapes de la recherche effectuée. Ce passage m'a obligé à une montée en compétences importante en matière de programmation. Sur le fond, la construction d'une application répondait à un double objectif convergent : me frotter aux exigences concrètes d'une recherche reproductible et expérimenter les conséquences épistémologiques de ce processus.

Cette dynamique s'inscrit donc tout autant dans une préoccupation contemporaine[369] que dans un questionnement historique. En effet, Olivier Orain met bien en avant le fait que la dimension constructible des recherches – caractéristique pour lui importante des postures scientifiques « post-crise » – est allée de pair avec le développement d'un positionnement épistémologique spécifique : le nominalisme. J'ai profité de ma recherche doctorale pour expérimenter de manière effective cette affirmation d'un changement épistémologique lié au fait de rendre une recherche constructible. Ce que j'en retire n'est pas une conclusion univoque. Il est évident qu'en donnant l'accès aux données, aux méthodes, aux résultats et/ou aux questionnements qui caractérisent un travail de recherche, sa déconstruction et sa critique sont également rendues plus aisées. Si cette dynamique peut engendrer chez certains chercheurs des évolutions épistémologiques, celles-ci ne sont ni obligatoires ni uniformes. Trois stades différents d'évolution peuvent être caractérisés :

- Un premier stade est la reconnaissance d'un traitement non-exhaustif et imparfait d'un sujet de recherche. Les limites du travail effectué sont plus ou moins explicitées selon les situations et les chercheurs. Parfois, il est affirmé que la recherche suivante va pouvoir remédier aux principales limites explicitées. Parfois, l'expression et la reconnaissance de limites plus fondamentales sont développées, ce qui marque une évolution épistémologique non négligeable. Au niveau de la préparation des données de cette recherche, il s'agit par exemple de reconnaitre et d'assumer des choix dans

[369] Préoccupation se retrouvant par exemple dans les principes du FAIR : Facile à trouver, Accessible, Interopérable, Réutilisable.

la construction (intégration ou non des comptes-rendus, des figures…) relevant d'une part de subjectivité.

- Un deuxième stade est celui de la reconnaissance qu'il n'existe jamais qu'un seul modèle explicatif, mais toujours plusieurs compréhensions possibles qui dépendent des façons dont les chercheurs interrogent les mondes et construisent leurs recherches. Ce positionnement donne lieu à plusieurs difficultés :
    - La cohérence théorique : tous les modèles d'intelligibilité ne peuvent pas être tenus ensemble, car ils sont construits sur des prémices et des raisonnements différents, et parfois même opposés.
    - L'articulation théorie / données construites : les latitudes liées à l'adaptation des théories et la fabrication des données ont des limites et, à ce stade (à la différence du suivant) toutes les articulations ne se valent pas.
    - Les importances respectives des pôles expérimental et herméneutique.

- Un troisième stade est de penser qu'il n'est pas possible d'affirmer, sur n'importe quel sujet, qu'un modèle d'intelligibilité est plus pertinent qu'un autre, ce qui fait basculer le chercheur dans une posture relativiste.

Si rendre une recherche plus constructible peut favoriser le passage d'un stade à l'autre, ce processus n'a rien d'automatique. En effet, le rendu de la construction d'une recherche peut être très réaliste (avec une restitution de recherche très affirmative sur son déroulé logique, ses résultats et sa correspondance avec le monde effectif) et inversement, le rendu d'un travail de facture « réaliste » (au sens où les processus de construction ne sont pas mis en avant) ne signifie pas que l'auteur n'a pas conscience de sa dimension construite et de ses limites. En rapport avec cette affirmation, la thèse d'un réalisme naïf généralisé à toutes les générations de géographie post-vidalienne me semble très difficile à croire et à soutenir. S'il est certain que les géographes classiques ont élaboré et répondu à des attentes universitaires qui ne mettaient pas en avant les processus de construction et les limites de leurs recherches, ils n'ignoraient pas tous que leurs écrits étaient des construits limités. Croire qu'à l'exception de deux ou trois noms, ils pensaient tous que leurs productions scientifiques rendaient parfaitement compte du monde, peut être certes pensé et affirmé, mais n'est-ce pas là une fiction bien commode et rassurante pour les suivants ?

Par rapport à la critique possible d'un continuisme exagéré, mon travail doctoral ne remet pas en cause que la « nouvelle géographie », en insistant sur l'importance de l'explicitation en amont des cadres théoriques d'une recherche, a marqué une évolution scientifique importante. Néanmoins, le prisme kuhnien qui conduit Olivier Orain à s'appuyer sur les

analyses développées par Ian Hacking d'un point de blocage entre « réalisme » et « nominalisme » (reprenant en partie la célèbre querelle médiévale des universaux) n'est pas adapté, d'après mes analyses, pour lire l'évolution de la géographie française de cette époque. Le fait d'ailleurs que Ian Hacking préfère le terme « structurisme-inhérent » à celui de « réalisme » est loin d'être un détail : la « nouvelle géographie » étant profondément structuraliste, il n'y a pas eu de rupture profonde et affirmée à ce niveau dans la géographie française sur la période étudiée par Olivier Orain. Les exemples de Franck Auriac et Claude Raffestin développés dans la thèse du plain-pied ne fournissent pas sur le fond des éléments d'« incommensurabilité » justifiant la pertinence d'un recours au modèle kuhnien.

Dans une visée plus contemporaine, la mise en avant détaillée du caractère construit des données, des méthodes et de l'interprétation des résultats, s'accommode mal avec une optique kuhnienne. Paul Girard dans sa présentation du projet RICardo[370] montre que le modèle centre-périphérie censé être « universel » (tout comme le modèle kuhnien) présente des limites importantes dès qu'un ensemble de situations est effectivement examiné en détail. Si, en amont, les données, les méthodes et les résultats sont explicités de manière critique, ces grands schémas sont très souvent déconstruits, comme le confirme également mon travail doctoral. Il existe dans ce cadre un certain décalage générationnel à un attachement et une sauvegarde à tout prix du modèle kuhnien.

Pour revenir plus spécifiquement à la préparation des données, le point, le plus important à mon sens, est un changement épistémologique personnel : à la suite d'une présentation dans un colloque[371], j'ai pris conscience qu'au-delà de l'intérêt de détailler finement l'ensemble des opérations effectuées pour rendre mes données pertinentes à traiter, il y avait un grand intérêt à inscrire ce processus dans une démarche plus large de « critique des sources ». Des opérations techniques acquièrent alors une fonction heuristique beaucoup plus grande en fournissant un recul critique sur le matériel empirique mobilisé. De plus, ce processus s'intègre alors dans la tradition d'un exercice fondateur de la discipline historique renouvelé par les contraintes et les potentialités numériques. Cette perspective m'a ouvert un champ de réflexion que je n'avais pas envisagé au départ et qui reste, pour moi, à développer avec d'autres travaux. Au lieu de passer sous silence et d'invisibiliser une série de transformations en amont des données, il s'agit *a contrario* de les mettre en valeur pour mieux comprendre leur construction initiale et leur nouvel usage.

---

[370] L'ensemble de la réflexion présentée est tiré du séminaire en ligne : https://medialab.sciencespo.fr/en/news/transnum-en-video-un-tournant-numerique-pour-lhistoire-economique/ (consulté le 18/09/2023). Le projet RICardo (Research on International Commerce) étudie le commerce international des débuts de la révolution industrielle à la veille de la seconde guerre mondiale : https://www.sciencespo.fr/recherche/fr/content/le-projet-ricardo.html (consulté le 18/09/2023).

[371] https://diachronie.org/2018/11/22/programme-colloque-international-histoire-langues-et-textometrie/ (consulté le 18/09/2023).

Pour finir sur cette étape de prétraitement, je dois reconnaitre *in fine* deux déceptions par rapport à l'application Web construite : la première relève du design et de l'ergonomie qui n'ont pas été assez travaillés pour mettre vraiment en valeur tout ce qui a été réalisé. La seconde est la conscience que la phase de test de cette application n'a pas été assez développée. Une telle application laisse toujours apparaitre des bugs qui doivent être corrigés au fur et à mesure de leurs advenues. Si j'ai corrigé beaucoup d'erreurs que j'ai identifiées au fil de mon utilisation, cette application Web n'a toutefois pas été assez utilisée par des utilisateurs extérieurs pour corriger des problèmes non anticipés.

III. Le traitement des données

La troisième étape a été celle du traitement des données. S'il ne s'agit pas de développer ici l'ensemble des opérations effectuées, il me semble important de revenir sur la phase préalable d'évaluation visant à choisir la méthode et les paramètres. En l'absence de référentiel objectif, j'ai dû adopter une méthode comparative pour évaluer les résultats obtenus de proximité sémantique. Ce moment a été identifié comme crucial, car il renvoie à l'absence *in fine* de base d'évaluation externe au langage : les mots se définissent par les mots dans un système qui n'est pas totalement stabilisé, logique et jamais exempt d'ambiguïtés. Il est vrai qu'une telle reconnaissance conduit sur le fond à s'éloigner d'une forme de réalisme.

Il est nécessaire ici de saisir toute la tension entre, d'un côté, une volonté d'objectivation et, de l'autre, cette reconnaissance d'ambiguïtés sémantiques persistantes et de limites méthodologiques décisives. Il ne fait aucun doute que l'évolution rapide du champ de recherche des plongements de mot ouvre la possibilité d'un nouveau chantier empirique. Je reste persuadé que, si des gains d'objectivation avec des améliorations dans la détection des proximités sémantiques sont évidemment possibles, les ambiguïtés sémantiques ne seront pas totalement levées et le travail d'interprétation aura toujours une place centrale, à moins d'un réel changement de paradigme.

La notion de paradigme est ici entendue dans un sens fort, bien illustré par le modèle de la chute des corps. Si une personne teste empiriquement ce modèle et qu'il trouve un résultat inattendu, il remet en cause totalement le modèle. Un tel cas est très rare dans les SHS. Par exemple, cette thèse a beau mettre à jour de nombreuses limites fortes et fondamentales des interprétations réalisées par Olivier Orain à partir de son application du prisme kuhnien à la géographie française, il est certain que sa lecture ne sera pas pour autant remise totalement en cause et abandonnée. Il n'existe pas d'élément comme dans le cas de la chute de corps qui provoque en soi ce renversement total du modèle. Il est toujours possible de jouer sur le

sens de certains termes pour essayer d'adapter le modèle ou de relire des éléments historiques pour qu'ils correspondent un peu mieux au modèle et ne le discréditent pas. Il y a là une différence forte de situations épistémiques et c'est aussi pourquoi on touche avec l'objectivation de la sémantique un élément crucial.

Il est possible ici d'objecter, face à cet argument poppérien, que les réflexions kuhniennes doivent être différenciées de celles de Karl Popper (1972). En effet, dans le modèle kuhnien, ce n'est pas la contradiction d'un test empirique qui provoque le changement de paradigme mais la conversion non totalement rationnelle d'un ensemble de chercheurs. Toutefois, il est nécessaire ici de rappeler que Thomas Kuhn définit la cause du changement de paradigme − l'« anomalie »−comme « l'impression que la nature, d'une manière ou d'une autre, contredit les résultats attendus dans le cadre du paradigme qui gouverne la science normale » (Kuhn 2008, 83). Cet aspect est d'autant plus important quand une lecture internaliste du modèle kuhnien est privilégiée, ce qui est le cas de la thèse d'Olivier Orain (*cf.* sections Chap8.III et Chap9.VI.3). S'il existe évidemment de multiples interprétations des propositions kuhniennes dont certaines, externalistes, sont forcément moins concernées par la force de l'argument poppérien précédemment développé, ce dernier doit être pleinement considéré dans le cas de la thèse du plain-pied.

Pour finir avec cet argument poppérien, il ne s'agit pas de nier le fait qu'il existe parfois, dans quelques cas, des éléments qui peuvent se révéler déterminants pour des communautés au sein même des SHS dans le rejet ou l'adoption d'une nouvelle théorie. Il n'en reste pas moins que la dynamique globale des SHS n'est pas celle d'une succession d'ensembles théoriques bien stabilisés qui périment à chaque fois ceux qui l'ont précédé. Les SHS sont bien plus marquées par des co-existences, des effets parfois de domination de certains courants, mais aussi avec des *revivals* d'anciennes optiques qui n'ont jamais été complètement abandonnées. Il existe, de manière générale, de plus ou moins fortes tensions entre des voies qui ne s'excluent pas et qui ne peuvent pas être strictement réfutées.

Pour revenir plus directement aux résultats obtenus de proximités sémantiques, sans être donc déterminants, ils n'en sont pas moins intéressants. Au niveau des termes étudiés, « espace » et « milieu », sur la période analysée par Olivier Orain (1910-1985), il n'a pas existé de passage net d'un ensemble conceptuel à un autre, avec effacement total de l'ancien ensemble par le nouveau. Un élément mis à jour est la montée d'une géographie utilisant le terme « espace » au singulier, et non plus au pluriel, avec le développement de nouveaux univers sémantiques. Cette observation est plus en adéquation avec les affirmations d'Olivier Orain sans pour autant valider la pertinence d'utiliser un modèle kuhnien pour lire l'histoire de la géographie française.

Mon travail quantitatif, du fait de ses conclusions non déterminantes, a été complété par de nombreuses analyses non quantitatives.

IV. Des analyses non quantitatives

La quatrième étape a commencé par des retours dans les textes. Il est assez étonnant de constater à quel point la critique du tournant quantitatif a été rapide aux États-Unis. Alors qu'il publie *Explanation in Geography* en 1969, texte de référence de la « new géographie », David Harvey écrit en 1972 :

> « La révolution quantitative [de la géographie] est arrivée à son terme et les rendements marginaux décroissants sont apparemment en train de s'installer, car… [elle] sert à nous renseigner de moins en moins sur les choses les plus pertinentes… Il existe une disparité évidente entre le cadre théorique et méthodologique sophistiqué que nous utilisons et notre capacité à dire quoi que ce soit de vraiment significatif sur les évènements qui se déroulent autour de nous… En bref, notre paradigme ne s'en sort pas bien. Il est mûr pour être renversé. »[372] (Harvey 1972, 6).

De telles critiques internes ont mis plus de temps à advenir en France au sein du mouvement spatialiste. Elles ont été moins radicales, mais vives pour autant. Certaines ont déjà été explicitées dans cette thèse (*cf.* section Chap2.II.2.b). Le texte d'Henri Chamussy, *d'amour et d'impuissance*, publié dans les brouillons *Dupont* 1978 en est également un bel exemple. Les limites de la démarche hypothético-déductives y sont explicitées avec beaucoup de justesse et de franchise :

> « Il ne suffit pas de dire "voici ma problématique, et je pose mes hypothèses" pour que tout marche bien. En général, on ne trouve pas du tout les données qui permettent de confirmer ou d'infirmer les hypothèses, et, à la fin de la démonstration, malgré tout le mal qu'on s'est donné pour prouver le contraire, les hypothèses sont restées… des hypothèses. Ou bien, quand on a l'impression qu'un géographe a bien démontré ses hypothèses, c'est souvent qu'il les a posées… après coup, en fonction des données dont il disposait. N'est-ce pas vrai ? Que celui qui n'a jamais commis ce péché me jette la première pierre » (Chamussy 1978, 74).

À la suite de ce paragraphe, il reconnaît qu'il n'existe pas de « paradigme enveloppant […] Il n'y a pas de passe-partout pour ouvrir les serrures multiples de la Recherche. Il y a un jeu, infini peut-être, de clés. » (Chamussy 1978, 77). Par rapport à ces remarques, plusieurs interprétations sont possibles. Tout d'abord, dans une perspective kuhnienne, il

[372] Traduction personnelle de : "[Geography's] quantitative revolution has run its course and diminishing marginal returns are apparently setting in as … [it] serve[s] to tell us less and less about anything of great relevance. … There is a clear disparity between the sophisticated theoretical and methodological framework which we are using and our ability to say anything really meaningful about events as they unfold around us. … In short, our paradigm is not coping well." (Harvey 1972, 6).

peut être revendiqué que ces citations renvoient à un temps de « crise » dans le système kuhnien. Dans ce cadre, il est légitime que les chercheurs soient quelque peu perdus, ce que le texte d'Henri Chamussy montre parfaitement bien. Toutefois, comme le reconnaît Olivier Orain dans sa thèse, le nouveau paradigme n'est jamais complètement advenu.

On peut aussi arguer que le discours d'Henri Chamussy, en particulier cette métaphore du jeu infini de clés, illustre parfaitement le constructivo-nominalisme que défend Olivier Orain. Cette objection me semble valide jusqu'à un certain point. En partant toujours des réflexions qu'Henri Chamussy développe dans cet article, notamment celles sur la froideur des mathématiques et la défense d'une géographie incarnée[373], il me semble fondamental de considérer toutes les implications de cette citation de Paul Valéry : *« On ne s'enivre ni ne se désaltère avec des étiquettes de bouteilles* » (2020, 35). L'argument ressemble à un trait d'humour, mais il touche à une autre dimension d'un réalisme profond[374] qui caractérise fortement le discours d'Henri Chamussy et me semble encore partagé par de nombreux géographes. Il y a un ancrage des recherches dans l'empirique et dans ce qu'il provoque et évoque au-delà d'un nominalisme abstrait marqué par l'affirmation que tout n'est qu'étiquette.

Il n'est pas question d'ignorer que parallèlement Henri Chamussy a été et est resté un fort promoteur d'une lecture kuhnienne de la géographie française. Le fort investissement dans l'apprentissage de nouvelles approches, tout autant que le bousculement d'une hiérarchie dans la géographie française, permettent de comprendre pourquoi Henri Chamussy oscille entre ces deux positions incompatibles. D'un côté, l'affirmation d'avoir vécu un changement de paradigme ; de l'autre, cette réflexion que la géographie ne se caractérise pas par un paradigme.

Si je développe ici ces propos alors que j'ai déjà en grande partie réalisé des analyses détaillées de certains positionnements d'Henri Chamussy dans le corps de cette thèse (*cf.* section Chap10.IV.2.a), c'est que je pense que nous sommes ici sur le nœud du problème : d'un côté, une revendication d'un changement scientifique important, couplé à des évènements sociaux (mai 68) marqués par une teneur révolutionnaire indéniable ; de l'autre, le maintien et le développement d'une lecture kuhnienne qui conduisent à des ambiguïtés et des difficultés notables dès qu'il s'agit de la détailler et de développer sa justification. Olivier

[373] « J'ai peur des discours secs et formalisés. La mathématique a certainement sa beauté silencieuse, rigoureuse, dépouillée (la beauté du diable tentateur ?), mais il est bien difficile d'y pénétrer, et puis elle est le reflet d'un monde abstrait, absolument abstrait. Nous ne pouvons guère parler avec les mêmes mots des montagnes et des plaines, des eaux courantes et des villes, des déserts et des océans. Toutes les analyses factorielles du monde sur les exploitations agricoles des Alpes ne me feront pas oublier l'odeur des foins, au soir d'été » (Chamussy 1978, 79).

[374] Réalisme profond qui est moins lié à cette volonté expérimentale de réaliser des affirmations en accord avec l'empirique, mais plus directement lié à la reconnaissance de l'importance des expériences sensibles.

Orain ne reconnaît qu'à demi-mot ces difficultés et sa construction contribue au fait qu'une partie des géographes français reste dans cette ambiguïté problématique sans tirer les conséquences épistémologiques d'un abandon d'une lecture kuhnienne de leur discipline.

Cette situation explique pourquoi ce travail ne s'est pas contenté d'un simple « retour dans les textes », mais a éprouvé la nécessité d'un travail beaucoup plus conséquent, avec des examens détaillés et critiques menés lors de cette étape sur trois fronts :

- Le premier concerne l'histoire de la géographie par rapport aux différentes étapes du modèle kuhnien : de la proto-science à la science, du paradigme à l'anomalie, de l'anomalie à la crise et de la crise au nouveau paradigme. Sans revenir sur toutes les limites mises à jour sur ce plan dans les développements d'Olivier Orain, le point le plus critique est à mon sens que la discontinuité scientifique est *in fine* toujours remise à plus tard dans la thèse du plain-pied. Alors que Thomas Kuhn affirme que seul un paradigme peut succéder à un paradigme, Olivier Orain rend compte d'une situation en contradiction avec cette dynamique puisque le nouveau paradigme ne s'affirme jamais complètement.
- Le deuxième front concerne les composantes d'une matrice disciplinaire détaillées par Thomas Kuhn. Sur ce plan, dans les géographies post-vidaliennes, il n'existe pas ou pratiquement pas de généralisation symbolique au sens où les entend Thomas Kuhn. De plus, les valeurs de prédiction (notamment quantitative) sont absentes et les exercices-types dans leur sens originel de « résolution de puzzle » sont très réduits alors que ces éléments sont fondamentaux dans le modèle kuhnien. Les nuances sémantiques et subtilités d'écriture de la thèse du plain-pied pour essayer tout de même d'appliquer ce modèle expliquent une certaine complexité dans la déconstruction entreprise car les points de discussion sont multiples, souvent très spécifiques et relèvent parfois de subtilités du fait de leur complexité initiale.
- Le troisième front concerne les auteurs hors de la sphère géographique mobilisés par Olivier Orain. Dans les trois cas examinés (Ian Hacking, Hilary Putnam et Jean-Michel Berthelot), ce travail montre que la thèse du plain-pied a laissé de côté de nombreux points problématiques. Chaque lecteur pourra se positionner entre l'admission d'un processus de sélection inhérent à toute construction scientifique et la reconnaissance de coupes plus problématique : Hilary Putnam reste un réaliste malgré toutes ses évolutions intellectuelles ; Ian Hacking propose un point de blocage qui n'est pas vraiment adapté par rapport à la situation étudiée par Olivier Orain ; les programmes de recherche d'Irme Lakatos (repris par Jean-Michel Berthelot) ne sont pas des sous-paradigmes, mais une proposition épistémologique concurrente de celle de Thomas Kuhn.

Par rapport à toutes ces limites, il peut être soutenu qu'il est normal qu'un modèle ne décrive pas parfaitement une situation empirique. Or, dans un tel cas, il est reconnu comme particulièrement important de préciser les limites du modèle par rapport aux situations examinées. C'est ce travail que réalise cette thèse en réexaminant une partie des analyses effectuées par Olivier Orain. Face aux nombreuses limites, l'alternative est soit de continuer d'aménager le modèle kuhnien, soit de l'abandonner pour penser à l'aide d'une autre grille de lecture.

Si les propositions passeroniennes m'ont beaucoup apporté sur ce sujet, j'ai le sentiment que mon utilisation dans ce rendu doctoral s'est trop limité à un schéma mobilisant l'argument de l' "entre-deux", entre pôle expérimental et pôle herméneutique. Si ce schéma est pratique, il pose aussi des approximations sémantiques qui mériteraient d'être levées, permettant ainsi de mieux caractériser, sous la continuité, des évolutions épistémologiques de la géographie française. Une autre perspective envisageable est de tenter de donner une place plus secondaire à ce schéma. Ces réflexions me conduisent à préciser que ce travail doctoral s'est centré délibérément surtout sur l'étape de critique du modèle kuhnien. Les lectures passeroniennes proposées doivent être considérées comme de premières étapes d'un travail à approfondir.

V. Les mises en perspective contemporaines

La dernière partie s'est attelée à développer des mises en perspectives contemporaines du sujet traité dans le champ des humanités numériques et dans la sphère des données massives. L'élaboration de cette partie a été largement aidée par les nombreuses réflexions d'auteurs visant à affirmer et argumenter en faveur d'une révolution scientifique. Un processus inverse à celui revendiqué par la géographie française des années 1970-1980 est souvent mis en avant avec le développement d'approches, cette fois-ci, inductives. Si le mythe d'une inductivité radicale (Schmitt 2018) marquant une rupture incommensurable avec les pratiques précédentes a bien été remis en cause, il est vrai que les machines peuvent réaliser des rapprochements inattendus au sein de données existantes. La recherche n'est pas devenue pour autant pilotée entièrement par les données (« data-driven ») comme certains tentent de l'affirmer (Hey 2012).

Si une partie non négligeable des discours paradigmatiques peut être facilement déconstruite, ces derniers génèrent également beaucoup d'idées tout à fait pertinentes qu'il serait dommage de rejeter dans la foulée. Par exemple, l'idée développée par Boris Beaude (2017) d'une réactivation de tensions historiques liées à ce développement quantitatif

contemporain me semble tout à fait convaincante. Plus qu'une réapparition du spectre de la « physique sociale », c'est la question de l'objectivation, avec en toile de fond un travail scientifique de plus en plus marqué par des médiations informatiques (au niveau de l'obtention, de la préparation et du traitement des données) qui interroge. La question de l'analyse et du degré de véridicité des résultats obtenus finit toujours par advenir : quel est le degré de correspondance avec ce qui existe, ce qui a pu être observé, ce qui a pu être mesuré ? *In fine*, c'est la question du réalisme qui est alors posée et souvent débattue.

Ainsi, quand Bruno Bachimont affirme l'avènement « d'un nominalisme de la culture [qui] remplace la compréhension linguistique par la donnée collectée et la corrélation statistique » (2017), il me semble illusoire de penser que les chercheurs vont en rester à ce premier stade des données collectées et des corrélations statistiques trouvées par les algorithmes. Il y aura toujours des controverses sur ce que signifient ces corrélations et à quoi elles correspondent effectivement. Ces controverses interpréteront les résultats en passant par la compréhension linguistique et plus globalement s'inscriront dans une forme de raisonnement sociologique telle que développée par Jean-Claude Passeron (2006, *1991*). Il est ici intéressant d'observer comment cet argument d'une rupture paradigmatique par le « nominalisme » ressort sous une nouvelle forme après la proposition d'Olivier Orain.

Par rapport à cet argument nominalisme, il convient d'apporter quelques précisions. Il est évident que les noms donnés aux entités et aux phénomènes géographiques sont arbitraires et peuvent être considérés comme des étiquettes, mais il faut bien comprendre les implications plus profondes du nominalisme. Comme le soulignait le mathématicien Hector José Sussmann :

> « Libre à chacun d'appeler un opérateur auto-adjoint un "éléphant" et une décomposition spectrale "une trompe". On peut alors démontrer un théorème suivant lequel "tout éléphant a une trompe". Mais on n'a pas le droit de laisser croire que ce résultat a quelque chose à voir avec de gros animaux gris » (cité par Ekeland 2015, 123).

Ce trait d'humour pose question par rapport à la géographie, qui à l'inverse des mathématiques, n'a pas un langage auto-suffisant détaché du monde. Il reste possible à partir de propositions formalisées de simuler des mondes virtuels, mais dans le domaine de la géographie, c'est à partir de la comparaison avec le monde actuel que la valeur de ces simulations peut être évaluée. Cela marque une différence notable avec les mathématiques qui n'ont pas besoin de ce retour au monde.

Dans toute cette réflexion, je n'ignore pas non plus qu'il existe des géographies qui ont réalisé plus d'efforts que d'autres dans la clarification de leurs systèmes conceptuels. Si ma formation et mon attention aux ambiguïtés sémantiques me rendent assez sensible à l'importance de ce travail, je ne pense nullement que ces différences de formalisation rendent

en soi certaines géographies supérieures à d'autres. Il n'est pas question de critiquer sur ce point la thèse d'Olivier Orain qui est assez subtile pour ne pas tomber dans l'expression directe de ce travers de la « supériorité d'une branche de la géographie ». Toutefois, le prisme kuhnien, par sa succession de paradigmes qui se remplacent les uns les autres, tend au fond à entretenir cette idée sous-jacente.

S'il est compréhensible qu'il soit difficile pour certains d'abandonner cette idée d'une supériorité des systèmes les plus formalisés et mathématisés, la diversité et le maintien d'une forte diversité d'approches ne peuvent qu'être reconnus par les géographes contemporains. Une fois accepté que les dynamiques de connaissance dans cette discipline ne se caractérisent pas par des changements de paradigme, il est très étrange de croire à une exception post-vidalienne. L'idée d'une école dominante est sûrement à la base d'une telle proposition, mais ce travail de thèse montre qu'un tel passage de l'« école » au « paradigme » apporte *in fine* plus d'ambiguïtés que d'apports heuristiques.

L'approfondissement que j'ai effectué dans les humanités numériques, outre le fait qu'il m'a permis d'éprouver la pertinence d'un prisme de lecture centrée sur les complémentarités et les tensions entre approches modélisantes et herméneutiques (*cf.* section Chap12.I.3), a été marqué par la découverte du travail de Julianne Nyhan et Andrew Flin (2016). Leurs réflexions, à l'aval d'interviews de nombreux pionniers dans le domaine des humanités numériques, sont particulièrement stimulantes. La pluralité sémantique autour de l'utilisation du terme de « révolution » les a conduits à s'éloigner d'une recherche en termes de véracité. Leur étude s'est ainsi orientée vers la fonction potentielle et le symbolisme de ce terme pour le groupe qui le manie. Comme le montrent aussi Julianne Nyhan et Andrew Flin, cette étiquette révolutionnaire, souvent soutenue initialement par le prisme kuhnien, fait, dans un second temps, courir le risque pour le groupe qui l'a initialement utilisée, de se retourner contre lui. Par rapport aux grandes prétentions souvent originellement affichées, les résultats plus modestes poussent à adapter le positionnement. Cela explique l'apparition et le développement d'une méfiance, voire d'un rejet de ce terme de « révolution » au sein même du collectif qui l'avait initialement mis en avant. Ce processus correspond assez bien à ce qui s'est passé globalement dans la géographie française.

Alors qu'une grande partie des « nouveaux géographes » prenaient de la distance avec le prisme kuhnien, c'est une reprise de flambeau qui a été effectuée par une partie de l'équipe *Épistémologie et Histoire de la GéOgraphie* (EHGO). La thèse d'Olivier Orain marque à mon sens un véritable cap, car la tentative de légitimer de manière forte le cadre kuhnien est poussée à son terme. Il est compréhensible que deux décennies plus tard, cette perspective soit datée, mais l'affirmation suivante formulée dans la thèse du plain-pied reste à mon sens d'une grande actualité :

> « en matière d'histoire des sciences, mettre en scène un évènement est tout sauf un geste insignifiant : le registre gradué qui va du "tournant" à la "rupture", de la "coupure épistémologique" à la "révolution scientifique", quel que soit son niveau de validation argumentaire, a une très grande performativité » (Orain 2003, 131).

La façon dont une recherche nous invite à penser l'histoire d'une discipline se doit par conséquent d'être considérée avec une grande attention. Il est nécessaire de la réfléchir en prenant en compte les différentes motivations historiques qui se sont succédé, mais aussi de la considérer à travers ses éventuelles rénovations et les résonances qu'elle crée dans nos pratiques personnelles. C'est l'objet du dernier chapitre de cette conclusion.

Chapitre 15 :

Actualités et ouvertures de cette thèse

Une actualité concomitante de la rédaction de cette conclusion doit être signalée du fait du lien étroit avec la problématique de ce travail doctoral.

I. Quel recalibrage de la « révolution quantitative » française ?

Il s'agit de la sortie de l'ouvrage : *Recalibrating the quantitative revolution in Géography : Travels, Networks, Translations* (Gyuris, Michel, et Paulus 2022). Ce livre rédigé par plusieurs auteurs rend compte de diverses situations nationales. Son objectif premier est d'apporter et de synthétiser des éclairages alternatifs par rapport à la perspective la plus développée que les auteurs de cet ouvrage jugent comme étant trop anglo-centrée. Un autre objectif complémentaire de ce livre est de complexifier ce qui est considéré comme :

> « des éléments de l'ancien récit homogénéisant et totalisant de la révolution quantitative [qui] ont survécu dans la critique radicale, par exemple, les frontières fermes revendiquées entre les géographies pré-quantitatives et les géographies quantitatives, ainsi que celles opposant les "quantificateurs" aux "non-quantificateurs" »[375] (Gyuris, Michel, et Paulus 2022, 16).

Après la lecture de cette thèse, il pourra paraître quelque peu étonnant, voire paradoxal, que ce soit Olivier Orain, fer de lance d'une lecture en termes de révolution kuhnienne, qui traite, dans le livre, de ce sujet pour la situation française. Le chapitre qu'il a écrit, *Une histoire sociale de la géographie quantitative en France des années 1970 aux années 1990*, permet ici d'apporter quelques précisions sur les évolutions de son positionnement. Depuis sa thèse soutenue en 2003, Olivier Orain s'est nettement tourné vers une histoire plus sociale et politique. Son récit est nettement moins prisonnier d'une volonté de validation du modèle kuhnien et d'une perspective internaliste.

Par rapport à cette évolution, il est important de souligner qu'il existe une tension entre l'approche structuraliste et une lecture centrée sur les acteurs (Touraine 2014). Toutefois, cette tension n'apparait pas dans l'écriture d'Olivier Orain, alors même que son approche délibérément plus sociale s'accompagne conjointement d'une lecture kuhnienne persistante, et même, sur certains points, plus affirmée. En effet, une précaution importante prise dans sa thèse avec la reconnaissance d'un paradigme spatialiste incomplet est passée sous silence dans cet ouvrage. Il est ainsi affirmé à propos de la géographie théorique et quantitative :

[375] Traduction personnelle de « many éléments of the former homogenizing and totalizing narrative of the quantitative revolution survived in the radical critique, e.g., claimed firm boundaries between prequantitatives geographies and quantitatives geographies, and "quantifiers" versus "nonquantifiers" » (Gyuris, Michel, et Paulus 2022, 16).

> « Au niveau des études supérieures et postuniversitaires, cela a donné l'occasion de définir des solutions concrètes aux problèmes et d'élaborer un apprentissage progressif pour les étudiants, en donnant dans le même parcours des valeurs partagées, des généralisations symboliques, des métaphysiques et, bien sûr, des techniques »[376] (Orain 2022, 124).

En citant ainsi tous les éléments de la matrice kuhnienne, l'idée d'un paradigme spatialiste est avalisée. Le terme d'« exercice-type » est stratégiquement évité. De manière plus globale, il faut souligner que ce choix d'Olivier Orain de réaffirmer sa lecture kuhnienne s'effectue dans un ouvrage dont le sous-titre de l'ouvrage « *Travels, Networks, Translations* » traduit une orientation latourienne en mettant en œuvre des analyses en termes de réseaux et de traductions… Or, ce positionnement épistémologique est sur le fond en tension avec une perspective kuhnienne.

Derrière ces problématiques de cadres théoriques se pose alors la question du niveau de « recalibrage » de la révolution quantitative et celle de l'utilisation de ces récits historiques dans les contextes contemporains. Cet enjeu est bien mis en valeur par les auteurs de cet ouvrage avec le choix de terminer par une série de questions posée à l'ensemble des auteurs, dont la dernière est : « Comment décririez-vous la pertinence de l'héritage de cette histoire pour les nouvelles méthodologies quantitatives, par exemple pour les discussions sur les géographies numériques ou les big data ? ». Sur ce sujet, Olivier Orain répond rapidement qu'il n'a pas travaillé cette question. Le fait d'avoir déjà apporté de nombreux éléments de réflexions sur cette thématique (*cf.* Chap13) me conduit à proposer en ouverture une perspective différente : celle d'une réflexion des apports de cette thèse ancrée dans mon actualité de fin de thèse.

II. Analyses à partir de mon actualité d'enseignant et d'ingénieur en appui à la recherche

Cette ouverture s'appuie tout d'abord sur les travaux dirigés (TD) que j'ai préparés et donnés pendant les deux dernières années de mon doctorat dans le département de sociologie de l'Université Grenoble Alpes. Ces TD ont été l'occasion de me frotter effectivement à différentes pratiques sociologiques, d'expérimenter par la pédagogie et de prolonger mon travail de recherche par de nouvelles approches et questions. Ils m'ont permis de penser et de réfléchir sur mon sujet à partir d'un autre lieu. Ce qui suit ne constitue pas à proprement

[376] Traduction personnelle de : « At a graduate and postgraduate level, it gave an opportunity to define concrete problem solutions and to elaborate a progressive apprenticeship for students, donating in the same journey shared values, symbolic generalizations, metaphysics and, of course, techniques » (Orain 2022, 124).

parler un retour d'expérience. C'est bien plutôt un ensemble de trois réflexions développées à partir de ces TD et intimement reliées à mon sujet de thèse.

1. À partir d'un manuel : la méthode hypothético-déductive

En préparant un TD d'analyse qualitative de contenu, j'ai tout d'abord été marqué par la lecture du manuel écrit par Jean de Bonville en 2006 et consacré à l'analyse de contenu des médias. Cet ouvrage est charpenté selon les dires même de l'auteur par une « approche hypothético-déductive d'inspiration poppérienne ». Les chapitres de cet ouvrage traitent successivement de la construction de la problématique, de la constitution du corpus, du codage, du traitement des données et de la présentation des résultats. Dans la conclusion, Jean de Bonville affirme :

> « Dans la réalité quotidienne du commun des chercheurs, pour un gramme de génie il faut plus d'un kilo de savoir-faire technique. Pour le chercheur « ordinaire », au sens où Thomas Kuhn parle de « science ordinaire », la recherche se résume à des opérations solidement fondées sur le plan conceptuel, bien menées sur le plan technique, mais d'un certain point de vue prévisibles et routinières. » (de Bonville 2006, 395).

Si ma thèse ne remet pas en cause les grandes étapes que décrit Jean de Bonneville, cette conclusion rend mal compte, à mon sens, de la diversité des approches et des différents arrangements propres à chaque recherche. De mon point de vue, il n'existe pas une ou des recette(s) qui marche(nt) à tous les coups. Il est bien entendu possible de suivre les grandes étapes proposées par Jean de Bonville, mais la probabilité est forte de se retrouver face à une situation imprévue. Le chercheur est alors contraint de revenir aux étapes précédentes et de trouver des arrangements originaux pour répondre à ses difficultés.

Ayant sûrement conscience que sa conclusion portait à controverse, Jean de Bonville a pris le soin d'y apporter quelques précisions :

> Cette insistance sur la méthode et sur l'intelligence du travail de terrain aura sans doute éveillé les soupçons des esprits prompts à déceler du positivisme dans toute recherche empirique. L'approche hypothético-déductive d'inspiration popperienne dont l'ouvrage adopte la démarche aura renforcé ces soupçons. Or ces choix découlent des objectifs pédagogiques de l'ouvrage et non de postulats théoriques. En effet, l'utilisation systématique de la logique et du vocabulaire de la sociologie empirique constitue la *lingua franca* méthodologique de la communauté scientifique. Pourquoi inventer un nouvel *espéranto* ? (de Bonville 2006, 395).

Cette question finale de Jean de Bonneville est à mon sens mal posée. En effet, il ne s'agit pas de créer un nouvel *espéranto*, mais d'affirmer qu'il n'existe pas de langue universelle. La sociologie, avec ses divers courants et controverses, illustre bien l'absence d'une *lingua*

franca théorique et méthodologique. Le choix effectué par Jean de Bonville n'est pas seulement pédagogique, il est aussi scientifique et politique. Par exemple, lors de mes expériences pédagogiques en sociologie à Grenoble, j'ai vu intervenir un responsable de section sur un support partagé pour rappeler qu'il serait bien de présenter d'autres plans que le modèle IMRAD (intiment liée à une présentation hypothético-déductive), car ce dernier n'était pas celui privilégié par les enseignants du département dans toute la suite du cursus. À l'inverse, dans une réunion du comité de rédaction de la revue *Géocarrefour*, certains acteurs ont exprimé la volonté de privilégier des articles construits autour d'une démarche d'inspiration hypothético-déductive pour répondre aux standards internationaux et être mieux reconnus.

Ma pratique d'ingénieur en appui à la recherche à la Plateforme Universitaire de Données Grenoble-Alpes (PUD-GA) durant ma dernière année de thèse m'a permis certes d'observer des dynamiques de procéduralisation de la recherche. Les déclarations de traitement de données aux Délégués à la Protection des Données (DPD) et les rédactions de Plan de Gestion de Données (PGD) conduisent à définir à l'amont de nombreux points. Alix Levain *et al.* (2023) montrent l'impact de ces dynamiques dans le domaine de l'ethnologie. Les transformations des critères de reconnaissance de la crédibilité des savoirs conduisent à des résistances des chercheurs. L'épistémologie passeronienne est avancée comme une des raisons de ces résistances renvoyant à :

> « l'indissociabilité entre la forme des données et l'épistémologie dont elles relèvent – c'est-à-dire, ici, sa dimension interprétative et non popperienne (Passeron, 2011). Ainsi, se déploient simultanément, au sein des communautés qui pratiquent l'ethnographie, une critique de la bureaucratisation de la recherche et des formes de mise à distance que les protagonistes de l'institutionnalisation de l'ORD [*Open Research Data* : données ouvertes de la recherche] s'efforcent, avec peine, d'analyser. » (Levain *et al*. 2023).

Si ces auteurs analysent la situation de l'ethnographie comme un « cas limite », le raisonnement passeronien s'applique en fait à l'ensemble des SHS. Dans la pratique, il n'existe aucune incompatibilité entre le mouvement de la science ouverte et le raisonnement passeronien. Les plans de gestion de données peuvent être adaptés au fur et à mesure des recherches et ils n'empêchent nullement les interprétations. Toutefois, la tension mise en avant par ces auteurs doit être, à mon sens, pleinement considérée, et cela d'autant plus qu'une partie des ingénieurs accompagnant la recherche n'a pas forcément conscience de ce point de vigilance.

2. À partir d'un article : le constructivisme contemporain

La deuxième réflexion ici présentée s'inscrit dans des lectures effectuées pour préparer un TD intitulé « La quantification des phénomènes sociaux ». Ces lectures m'ont permis de découvrir les réflexions de Cyril Lemieux, notamment son article publié en 2012 : *Peut-on ne pas être constructiviste ?*. Son constat dans cette publication est le suivant :

> *«* L'approche constructiviste, tout au long des années 1990, s'est généralisée. Sa puissance théorique, ses formules rhétoriques et la légitimité de ses procédures d'enquête se sont imposées, au fur et à mesure que ses tenants, au prix de luttes indissociablement intellectuelles et institutionnelles, gagnaient des positions dominantes dans nombre de départements de science politique et de sociologie et parvenaient à faire de l'idée que la réalité est socialement construite le fondement de la formation transmise aux étudiants. » (Lemieux 2012, 170).

Il est évident que ces idées d'une bataille gagnée des « positions dominantes » et d'un « fondement de la formation transmise aux étudiants » (même s'il n'est pas question d'exercice-type) peuvent être mobilisées en faveur d'une utilisation du terme paradigme. Par rapport à la thèse d'Olivier Orain, il faut toutefois préciser que le constructivisme auquel fait référence Cyril Lemieux, n'est pas similaire au réalisme teinté de constructivisme développé par la géographie des années 1970-1980 (*cf.* section Chap9.VI). En acceptant dans un premier temps de faire fi de cette différence, un rapprochement avec la thèse du plain-pied peut être effectué quand Cyril Lemieux s'interroge sur le fait de savoir

> « s'il est envisageable, aujourd'hui, de prendre ouvertement ses distances à l'égard de la position constructiviste sans s'attirer immédiatement le soupçon de vouloir restaurer la vieille épistémologie du "réalisme naïf " ou de chercher, sur le plan politique, à réhabiliter des positions réactionnaires favorables à la naturalisation du monde social » (Lemieux 2012, 171).

Nul doute que cette réflexion peut alimenter la thèse du plain-pied d'Olivier Orain. En effet, selon cette citation, l'alternative est soit de se reconnaitre et d'être reconnu comme « constructiviste », soit de courir le risque d'être assimilé à « la vieille épistémologie du "réalisme naïf" ». Cette alternative est ici volontairement dramatisée par Cyril Lemieux pour problématiser son article, car de nombreuses ambiguïtés de ce constructivisme contemporain sont ensuite développées. Par exemple, de nombreux chercheurs revendiquent que toute connaissance est socialement construite tout en se pensant eux-mêmes un peu plus près que les autres d'une connaissance objective. Cette position comme l'affirme Cyril Lemieux n'est pas en accord avec un constructivisme totalement assumé et relève d'un réalisme masqué.

De plus, l'exemple problématique que développe cet auteur à propos du fait de considérer que le développement d'une maladie virale n'est rien d'autre qu'une construction sociale, conduit à souligner que la géographie, prise dans l'interface nature / société, a peut-être

stratégiquement intérêt dans certaines de ses (dé)monstrations à éviter un constructivisme fort. En effet, s'il n'est pas question d'ignorer que les facteurs sociaux contribuent à des diffusions plus ou moins rapides des maladies, il est évident qu'il existe aussi des facteurs physiques qui limitent ou favorisent l'émergence et la multiplication des virus. La géographie, même si elle est conduite souvent à remettre en question l'appellation de « facteurs naturels » (Pigeon 2002), a tout intérêt à conserver dans ses possibilités explicatives l'idée que la diffusion d'une maladie virale ne relève pas que d'une construction sociale. Derrière cette affirmation se pose également la question des responsabilités, ce qui reconduit évidemment aussi à la problématique de l'objectivation.

L'actualité pandémique qui a marqué une grande partie de l'effectuation de cette thèse montre que les discours restent rarement limités à des faits prouvés. Certains peuvent sans aucun doute regretter cette absence d'une objectivité totale et « jeter le bébé avec l'eau du bain » en déclarant la recherche de toute objectivation vaine. Ma thèse insiste *a contrario* sur le besoin de ne pas abandonner cette recherche et met en avant une autre conception : il existe dans le travail interprétatif qui entoure des faits (plus ou moins solidement établis) des extensions qui sont certes plus fragiles, mais précieuses car elles confèrent du sens au-delà d'une simple accumulation de faits. Il est souvent particulièrement difficile de tracer des limites qui distingueraient expressément les faits prouvés[377] des interprétations. Il existe alors une « interminabilité » de nombreuses analyses qui « est à l'opposé de la stratégie de recherche qui vise obstinément à la clôture du questionnement dans un protocole falsificateur » (Passeron 2006, 591).

Il peut exister bien entendu de manière locale la production de preuves qui rendent caduques certains discours, mais comme le soulignent Silvio O. Funtowicz et Jérôme R Ravetz (1993), la période contemporaine est plutôt marquée par la formule : « faits incertains, valeurs contestées, enjeux élevés et décisions urgentes » (*Ibid* 1993). Ces auteurs développent l'idée d'une science à l'« âge post-normal » (*Ibid* 1993). Si je comprends l'idée d'une accélération liée à des « enjeux élevés » et des « décisions urgentes », la formule me semble maladroite, car elle sous-entend l'idée pour les SHS d'une période antérieure de science normale.

3. À partir d'un compte-rendu : la géographie post-vidalienne

La troisième et dernière réflexion ici présentée provient originellement d'un exercice de fiche de lecture. Le choix de textes proposés aux étudiants incluait un article écrit par

[377] Les preuves renvoyant toujours à des protocoles qu'il convient d'examiner et de comparer.

William Foote Whyte (1949) : *The Social Structure of the Restaurant*. En lisant ce texte, j'ai été frappé par l'utilisation précoce de sociogrammes me rappelant des propositions intellectuelles de la « nouvelle géographie » plus de vingt ans après[378]. À la suite de ces réflexions, j'ai effectué quelques recherches pour savoir si la géographie française avait tenu compte des recherches de l'École de Chicago du début du XX^{ème} siècle aux années 1950. Un compte-rendu abordant ce sujet, intitulé *Géographie humaine et sociologie* et publié par Jean-Charles Falardeau (1950) dans la revue *Géocarrefour*, montre que la géographie française a effectivement réfléchi ces propositions de la sociologie américaine. Ce compte-rendu montre également que la géographie française n'était pas un monde coupé des productions extérieures, contrairement à ce qui est parfois soutenu.

Ce choix de développer ici ce compte-rendu provient du fait qu'il me semble important dans ce dernier temps conclusif de revenir sur la géographie post-vidalienne qui constitue le cœur de l'argumentation d'Olivier Orain. Dans ce compte-rendu, Jean-Charles Falardeau, après avoir rappelé que l'observation constitue la plus grande part du travail du sociologue, fait un parallèle avec la géographie en rappelant que l'« A B C de tout reste l'observation » (Falardeau 1950, 346). Si cet élément corrobore la thèse d'Olivier Orain, d'autres points de ce compte-rendu viennent aussi la nuancer, voire la contredire.

Tout d'abord, l'observation n'est pas une finalité comme le souligne J-C Falardeau :

> « On peut reprocher à certains géographes humains de ne pas aller au-delà du plan de l'observation : ils ne font que de la – graphie. Notre ambition à tous géographes et sociologues, est d'aller au-delà, pour un ensemble cohérent et plausible d'explications » (Falardeau 1950, 346).

Cette posture, surtout quand elle est soutenue par le projet initial de Paul Vidal de la Blache d'éviter tout déterminisme en mettant à jour toute la complexité des causalités, est difficilement compatible avec une posture réduite à un simple plain-pied. L'arpentage du terrain ne suffit pas. Il n'existe pas d'écriture transparente qui rend le monde tel quel, sans médiation. S'il est évident que certaines écritures rendent très peu compte de leurs processus de construction, faut-il pour autant entretenir cette illusion en pensant (ou en faisant penser) que leurs auteurs relevaient d'un réalisme naïf ? Ma thèse répond négativement à cette

[378] La découverte des différentes structures sociales de restaurants présentées par William Foote Whyte m'a notamment provoqué un effet similaire à celui que j'avais connu lors de ma découverte bien antérieure des différents types de stations de ski répertoriés par Rémy Knafou (1978). Dans les deux cas, une sorte de jeu intérieur s'est mis en place se basant sur la remémoration d'un ensemble d'expériences vécues et d'essais de mise en correspondance avec les types schématiques proposés. Il est évidemment possible de faire remonter cette tradition intellectuelle jusqu'à l'outil de « l'idéal-type » proposé par Max Weber (2005). Toutefois, il faut préciser qu'il existe entre certaines propositions de la « nouvelle géographie » et les sociogrammes une plus grande proximité du fait des représentations sagittales utilisées pour décrire et comprendre les systèmes étudiés.

question. L'hypothèse d'un réalisme naïf généralisé me semble quelque peu réductrice, voire offensante, pour plusieurs générations de géographes.

Par rapport au travail critique mené, une des difficultés que ma thèse a dû affronter est de veiller à ne pas se laisser emporter dans des affirmations contraires à celles de la thèse du plain-pied qui constitueraient une exagération à front renversé. Par exemple, dans la discussion qui suit le compte-rendu précédemment cité sur la sociologie américaine, M. A Gibert développe l'idée « qu'il n'y a pas de faits géographiques en eux-mêmes, mais des enchaînements de faits, des rapports chronologiques ou spatiaux » (Falardeau 1950, 344). Il est évident que cette précision peut être utilisée pour revendiquer une ontologie non référentielle : les faits géographiques n'existent pas « en eux-mêmes ». Toutefois, il me semble qu'une telle utilisation aboutit à affirmer de manière abusive un positionnement théorique qui n'est pas confirmé par d'autres éléments, notamment, comme le rappelle ce même compte-rendu, qu'à l'époque « le point de départ de l'observation de la géographie humaine, c'est le sol » (*Ibid* 1950, 342). Les développements des géographies post-vidaliennes restent ancrés dans le matériel, le concret, l'observable et, donc, dans un réalisme non négligeable.

Il faut aussi bien entendu rappeler qu'il existe au sein des géographies post-vidaliennes une certaine faiblesse théorique en ce qui concerne la géographie humaine. Cette faiblesse théorique ne veut pas dire pour autant absence de cadres d'appréhension du monde. Toutefois, il a existé de la souplesse dans ces cadres, ce qui a contribué à une logique non-paradigmatique avec beaucoup de place laissée pour saisir les spécificités et les disparités des cas étudiés. La prise en compte des anomalies au sens d' « "omalos", inégal, irrégulier, fait d'aspérité » (Lefort 2015) ne crée pas d'anormalité conduisant à des changements de paradigmes.

Enfin, il faut reconnaitre qu'il existe dans les géographies post-vidaliennes globalement une méfiance par rapport aux généralisations trop hâtives, notamment vis-à-vis du risque de lire des situations empiriques par le prisme de cadres préconstruits. Cet aspect, bien mis en avant par Olivier Orain pour soutenir ses analyses, est développé par Maurice Le Lannou dans le compte-rendu précédemment cité à propos de l'école de Chicago :

> « Mais la sociologie américaine travaille-t-elle vraiment sur du concret ? Pas tout à fait. Elle recherche sinon des lois, du moins des normes de la vie générale dans des cadres urbains. Elle s'intéresse à la ville théorique, à la ville splendide établie sur une surface plate, balayée par des vents convergents ou divergents, en tous cas géométrique, peuplée d'hommes nés dans le même « environnement ». Cette ville serait, nous dit-on, composée de cinq anneaux concentriques. N'est-ce pas une vue assez éloignée de la réalité ?

> Il reste que ces recherches sont passionnantes. Je les assimilerais volontiers à la géographie générale. La géographie humaine générale est proche de la sociologie et la sociologie sera de la géographie générale chaque fois qu'elle nous fera connaître les normes de l'organisation des hommes en groupes, abstraction faite de toute contingence. La géographie tout court pourrait être une confrontation entre ces idées abstraites et les obstacles géographiques au sens très large du terme : ceux du milieu physique (site) comme ceux d'une société humaine particulière » (Falardeau 1950, 345).

Ces réflexions de Maurice Le Lannou ne se caractérisent pas par un rejet des propositions théoriques de l'École de Chicago. *A contrario*, elles sont reconnues comme « passionnantes » et rentrant dans le cadre de « la géographie humaine ». Il est certes vrai que ces modèles de géographie générale ne sont pas alors utilisés et que leur manipulation effective a bousculé la géographie classique. Cependant, comme le montrent les propos de Maurice Le Lannou, il n'existe pas un « rejet de principe » ni une « anomalie » fondamentale au sens kuhnien du terme. Maurice Le Lannou peut théoriquement et techniquement penser cette possibilité comme voie de recherche. Cette dernière ne fait pas partie d'un autre monde qui serait inacceptable et/ou insensé par rapport à ses propres univers de recherche. De plus, les travaux produits ultérieurement par la « nouvelle géographie » ne périmeront pas les résultats des études précédemment menées, contrairement à ce qui se passe dans un changement de paradigme tel que Thomas Kuhn le décrit.

III. *In fine…*

Pour conclure, mon travail doctoral et celui d'Olivier Orain (2003) montrent, tous les deux, que l'usage du modèle kuhnien appliqué à une relecture épistémologique d'un moment historique de la géographie française continue de présenter de nombreuses complexités et ambiguïtés. S'il est toujours possible de jouer sur la plasticité sémantique de l'expression « de plain-pied » et sur celle des termes « réalisme » et « constructivisme » pour nourrir l'idée d'une révolution, il est bien plus simple de sortir de cette perspective paradigmatique. Il n'existe pas d'incommensurabilité mais une coexistence de positionnements entre réalisme et nominalisme : ceux-ci se déploient sur de multiples plans (linguistique, scientifique, ontologique…) et selon de subtiles différences à la fois pour chaque plan et entre eux. La géographie française a évidemment connu des évolutions épistémologiques importantes dans les années 1960-1970, mais l'application du modèle kuhnien conduit, quand elle est poussée à son terme, soit à l'adoption de conceptions distantes et clivantes, soit au report de la révolution scientifique à plus tard.

Pour s'extraire des approches polarisantes, il est utile de rappeler, comme l'effectue Dominique Raynaud (2019), que les sciences physiques ne raisonnent pas toujours dans un

cadre hypothético-déductif et qu'elles ne recourent pas non plus systématiquement à un raisonnement fondé sur des prédictions. Symétriquement, il existe bien, aussi, des modélisations et des prédictions au sein des SHS. Cependant, je suis loin de partager la conclusion de Dominique Raynaud qui se sert de ces constats pour s'opposer aux propositions de Jean-Claude Passeron.

Si ce dernier affirme qu'« on ne peut marier langue naturelle et langue artificielle sans désigner au couple un chef de ménage » (Passeron 2006, 568) et soutient l'idée que la langue naturelle est pour l'instant le chef de ménage des SHS, ce positionnement n'exclut en rien la possibilité d'expérimentations et de modélisations. Ainsi, quand Dominique Raynaud (2019) présente en détail une modélisation ayant donné lieu à une prédiction s'étant avérée exacte[379] dans le domaine des sciences historiques, il ne s'agit aucunement, à mon avis et contrairement à ce qu'il affirme, d'un élément de discrédit des réflexions passeroniennes (2006, *1991*). Toutefois, la proposition que j'ai réalisée − une conception privilégiant un gradient entre deux pôles, expérimental et herméneutique *(cf.* section Chap8.V) − permet, je pense, d'éviter ce problème en affirmant plus clairement la possibilité et l'existence d'un pôle expérimental quelles que soient les disciplines. Il n'en reste pas moins que ce pôle prend des formes et des importances très variables suivant chaque discipline et recherche. Conjointement, mon travail doctoral réaffirme, comme Jean-Claude Passeron, que la langue naturelle et le pôle interprétatif restent prééminents dans les SHS.

Pour finir, et cela, quel que soit le positionnement épistémologique de mes lecteurs, j'espère que mon travail doctoral a été, est et sera une occasion de réfléchir sur les modalités de production des savoirs dans les sciences sociales. Les réflexions épistémologiques doivent être couplées autant que possible aux dynamiques mêmes de la production scientifique. Ces réflexions me semblent d'autant plus signifiantes qu'elles s'accompagnent d'interrogations sur certaines catégories d'analyse employées usuellement : « inductif » / « déductif », « réalisme » / « constructivisme »… Le syntagme daté de « nouvelle géographie », utilisé comme tel dans ce manuscrit, mériterait également d'être pris en charge réflexivement, en particulier pour sa performativité : en effet, tant que le prisme kuhnien reste aussi latent dans l'histoire de la géographie française (avec des critiques, soit à demi-mot, soit réduites à un ou deux auteurs), cette expression « nouvelle géographie » continuera de faire circuler une mémoire, une représentation, une culture disciplinaire et une forme d'impensé épistémologique.

---

[379] En l'occurrence, il s'agit de la relation entre le cosmopolitisme dans les universités médiévales par rapport à leur position dans le réseau. Les compositions de quelques universités n'ayant été découvertes que tardivement, il a pu être observé que la loi établie n'était pas remise en cause par ces nouvelles données, processus que Domique Raynaud qualifie et met en avant comme « raisonnement expérimental » (2009).

Au défi fort d'articuler des questionnements historiques avec des problématiques contemporaines, il n'est pas inutile d'ajouter les problématiques et les enjeux des différents horizons d'attente. Sur ce plan, la critique des discours révolutionnaires entourant les innovations technologiques n'empêche en rien de les utiliser, de manière intensive, et de s'engager dans les nombreuses possibilités d'explorations qu'elles permettent. Ce manuscrit en présente une. Au plaisir de lire les vôtres…

Bibliographie :

Une version électronique de la bibliographie de cette thèse est disponible à cette adresse : **@BiblioComplete**. Il est alors possible d'afficher des sous-bibliographies par thème et par langue, de revenir dans le texte de l'application et d'accéder directement aux ressources quand elles sont disponibles sur le Web.

La bibliographie qui suit est celle de l'ensemble du manuscrit sans distinction thématique ou linguistique :

Alvarado, Rafael. « Digital Humanities and the Great Project: Why We Should Operationalize Everything ». In *Debates in the Digital Humanities 2019*, University of Minnesota Press, (2019): 75-82.

Anderson, Chris. « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete ». *Wired*, 2008.

Auriac, Franck. « Analyse spatiale et matérialisme : introspection ». *Géocarrefour* 78, nº 1 (2003): 7-11.

Bachelard, Gaston. *Le nouvel esprit scientifique*. Presses Universitaires de France, 2020, 1ère ed : 1934, 232 p.

Bachimon, Philippe. « Physiologie d'un langage. L'organicisme aux débuts de la géographie humaine ». *Espace Temps* 13, nº 1 (1979): 75-103.

Bachimont, Bruno. « Le numérique comme milieu : enjeux épistémologiques et phénoménologiques. : Principes pour une science des données ». *Interfaces numériques* 4, nº 3 (2017): 385-402.

Bachimont, Bruno et François Rastier. « Herméneutique matérielle et artéfacture. Échange entre François Rastier et Bruno Bachimont sur sa thèse », *Texto !,* http://www.revue-texto.net/1996-2007/Dialogues/Rastier-Bachimont.html, (1998): consulté le 21 sept 2021

Bailly, Antoine S et Jean-Bernard Racine. « La géographie et l'espace géographique : à la recherche d'une épistémologie de la géographie ». *L'Espace géographique*, nº 4 (1979): 283-91.

———. « Les géographes ont-ils jamais trouvé le nord ? Questions à la géographie ». *L'Espace géographique* 7, nº 1 (1978): 5-14.

Baker, Monya. « 1,500 Scientists Lift the Lid on Reproducibility. » *Nature* 533, nº 7604 (2016): 452-54.

Barber, Bernard. « Review of TS Kuhn. The Structure of Scientific Revolutions ». *American Sociological Review* 28 (1963): 298-99.

Bastin, Gilles et Paola Tubaro. « Le moment big data des sciences sociales ». *Revue française de sociologie* 59, nº 3 (2018): 375-94.

Bataillon, Claude. *Géographes : Génération 1930*. Presses Universitaires de Rennes, 2009, 226 p.

Beaude, Boris. « (re)Médiations numériques et perturbations des sciences sociales contemporaines ». *Sociologie et sociétés* 49, nº 2 (2017): 83-111.

Beaujeu-Garnier, Jacqueline. *La géographie : méthodes et perspectives*. Masson, 1971, 141 p

Beligné, Max, « Répondre au défi de la reproductibilité d'une recherche en humanité numérique par la création d'une interface web », Video canal-U, https://www.canal-u.tv/145775, (2023), consulté le 8 sept 2023.

Beligné, Max, Campar Aleksandra, Jean-Hugues Chauchat, Isabelle Lefort, Sabine Loudcher et Julien Velcin, « Détection automatique de métaphores dans des textes de Géographie : une étude prospective », Article court. Acte colloque Traitement Automatique Langage Naturel (TALN), 2017

Beligné, Max, Isabelle Lefort et Sabine Loudcher. « Suivez l'évolution de vos mondes lexicaux dans le temps ! », 15ème Journées internationales d'Analyse statistique des Données Textuelles, http://lexicometrica.univ-paris3.fr/jadt/, 2020a.

Beligné, Max, Isabelle Lefort et Sabine Loudcher. « Une épistémologie numérique des disciplines est-elle possible ? » *Humanités numériques,* vol 1, 2020b.

———. « Du nettoyage des données à la critique des sources » dans *Histoire de mots,* Ed de la Sorbonne, 2023a.

———. « Analyse critique d'un changement de paradigme en géographie à partir de réflexion entre discontinuités scientifiques, lexicales et sémantiques ». dans *Penser avec les discontinuités en géographie*. Presses universitaires de Rennes. 2023b.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, et Christian Jauvin. « A Neural Probabilistic Language Model ». *Journal of Machine Learning Research* 3 (2003): 1137-55.

Bennafla, Karine et Claude Kergomard. « Commentaire de documents géographiques », https://www.ens.psl.eu/IMG/file/concours/2006/AL/geographie_epreuve_a_option_oral.pdf, (2005), consulté le 1 sept 2023.

Berdoulay, Vincent. « Géographie : lieux de discours ». *Cahiers de géographie du Québec* 32, nº 87 (1988): 245-52.

———. « La métaphore organiciste ». *Annales de géographie* 91, nº 507 (1982): 573-86.

———. *La formation de l'école française de géographie (1870-1914).* Bibliotheque nationale. Paris, 1981. 245 p.

Berdoulay, Vincent et Olivier Soubeyran. « Lamarck, Darwin et Vidal : aux fondements naturalistes de la géographie ». *Annales de géographie* 100, nº 561 (1991): 617-34.

Bergson, Henri. *La pensée et le mouvant: Essais et conférences (1903-1923)*. Felix Alcan. Paris, 1934.

Berque, Augustin. *Écoumène, introduction à l'étude des milieux humains*. Belin., 2014, 446 p.

Berra, Aurélien. « Faire des humanités numériques ». Dans *Read/Write Book 2 : Une introduction aux humanités numériques*, OpenEdition Press, (2012): 25-43.

———. *Connaître aujourd'hui. L'épistémologie problématique des humanités numériques*. MESHS. DHNord ,https://publi.meshs.fr/page/connaitre_aujourd_hui_epistemologie_problematique_des_humanites, (2014) consulté le 10 mai 2018.

Berry, David. « Digital Humanities: First, Second and Third Wave ». *Stunlaw*, article de blog, http://stunlaw.blogspot.com/2011/01/digital-humanities-first-second-and.html (2011), consulté le 8 mars 2023.

———. « Subjectivités computationnelles ». Traduit par Yves Citton et Anthony Masure. *Multitudes* 59 (2015).

Berthelot, Jean-Michel. *Épistémologie des sciences sociales*. Presses Universitaires de France, 2018, 1094 p.

Bertrand, Romain. *Le Détail du monde*. Seuil, 2019. 288 p.

Besse, Jean-Marc, Pascal Clerc, Marie-Claire Robic, Wolf Feuerhahn et Olivier Orain. « Qu'est-ce que le « spatial turn » ?. Table ronde ». *Revue d'histoire des sciences humaines*, nº 30 (2017): 207-38.

Bigot, Jean-Edouard, Virginie Julliard et Clément Mabi. « Humanités numériques et analyse des controverses au regard des SIC ». *Revue française des sciences de l'information et de la communication*, nº 8 (2016).

Birhane, Abeba, Atoosa Kasirzadeh, David Leslie et Sandra Wachter. « Science in the Age of Large Language Models ». *Nature Reviews Physics* 5, nº 5 (2023): 277-80.

Bloor, David. *Knowledge and Social Imagery*. Routledge & Kegan Paul. Londres, 1976, 209 p.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, et Tomas Mikolov. « Enriching Word Vectors with Subword Information ». *Transactions of the Association for Computational Linguistics* 5 (2017): 135-46.

Bonnamour, Jacqueline. « La Géographie et ses dictionnaires ». *L'Information Géographique* 68, nº 1 (2004): 76-86.

Bonville, Jean de. *L'analyse de contenu des médias: De la problématique au traitement statistique*. Bruxelles: De boeck sup, 2006, 451 p.

Borgman, Christine L. *Qu'est-ce que le travail scientifique des données ? : Big data, little data, no data*. Traduit par Charlotte Matoussowsky. Encyclopédie numérique. Marseille, OpenEdition Press, 2020.

Bougnoux, Daniel. *La crise de la représentation*. Paris, La Découverte, 2019, 240 p.

Boulineau, Emmanuelle. « Un géographe traceur de frontières : Emmanuel de Martonne et la Roumanie ». *L'Espace géographique* 30, nº 4 (2001): 358-69.

Boullier, Dominique. « Pour des sciences sociales de troisième génération (SS3G) ». In *Big data et traçabilité numérique*, édité par Pierre-Michel Menger et Simon Paye, 163-84.

Bourdeloie, Hélène. « Ce que le numérique fait aux sciences humaines et sociales. » *tic & société* 7, nº 2 (2013).

Boure, Robert. « Sociologie des revues de sciences sociales et humaines ». *Réseaux. Communication - Technologie - Société* 11, nº 58 (1993): 91-105.

Bourgeat, Serge. « La thèse d'Etat de géographie (1960-1984) : la diffusion de l'innovation au risque des contraintes disciplinaires ». Thèse, Université Joseph-Fourier - Grenoble I, 2007.

Boussidan, Armelle. « Dynamics of Semantic Change : Detecting, Analyzing and Modeling Semantic Change in Corpus in Short Diachrony ». These, Lyon 2, 2013.

Boyd, Danah et Kate Crawford. « Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon ». *Information, Communication & Society* 15, nº 5 (2012): 662-79.

Braam, Robert R et Henk F Moed. « Mapping of Science by Combined Co-Citation and Word Analysis. I. Structural Aspects ». *Journal of the American Society for Information Science* 42, nº 4 (1991): 233-51.

Braudel, Fernand. « La géographie face aux sciences humaines ». *Annales* 6, nº 4 (1951): 485-92.

Broc, Numa. « École de Grenoble contre école de Paris : les Alpes enjeu scientifique ». *Revue de Géographie Alpine* 89, nº 4 (2001): 95-105.

———. « Roger Dion (1896-1981) ». *Annales de géographie* 91, nº 504 (1982): 205-17.

Broca, Sébastien. « Épistémologie du code et imaginaire des « SHS 2.0 » ». *Variations. Revue internationale de théorie critique*, nº 19 (2016).

Bruneau, Michel. « Pierre Gourou (1900-1999) ». *L'Homme. Revue française d'anthropologie*, nº 153 : (2000):7-26.

Brunet, Étienne. « Qui lemmatise dilemme attise », *Scolia*, vol. 13, (1999): 7-32

Brunet, Roger. *Le déchiffrement du Monde : Théorie et pratique de la géographie*, Humensis, 2017, 359 p.

———. « Raisons et saisons de géographe ». *Géocarrefour* 78, nº 1 (2003): 13-18.

———. (dir). *Géographie universelle*. 10 volumes. Belin-Reclus, 1995.

Brunet, Roger, Robert Ferras et Hervé Théry (dir). *Les mots de la géographie: dictionnaire critique*. RECLUS, 1992, 520 p.

Brunhes, Jean. *La Géographie humaine*. 3 volumes, Librairie Félix Alcan, 1925.

Bunge, William. *Theoretical Geography*. C. W. K. Gleerup, 1962

Calbérac, Yann. « Terrains de géographes, géographes de terrain. Communauté et imaginaire disciplinaires au miroir des pratiques de terrain des géographes français du XXe siècle »., Thèse, Université Lumière - Lyon II, 2010.

Callaway, Ewen. « DeepMind's AI Predicts Structures for a Vast Trove of Proteins ». *Nature* 595, nº 7869 (2021): 635-635.

Callon, Michel, Jean-Pierre. Courtial et F. Laville. « Co-Word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemsitry ». *Scientometrics* 22, nº 1 (1991): 155-205.

Callon, Michel. « The Sociology of an Actor-Network: The Case of the Electric Vehicle ». In *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*,. London: Palgrave Macmillan UK, (1986): 19-34

Callon, Michel, Jean-Pierre Courtial, William A. Turner et Serge Bauin. « From Translations to Problematic Networks: An Introduction to Co-Word Analysis ». *Social Science Information* 22, nº 2 (1983): 191-235.

Camacho-Collados, Jose et Mohammad Taher Pilehvar. « On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis ». *arXiv*, 2017, consulté le 21 janv 2021

Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, *et al*. « Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015 ». *Nature Human Behaviour* 2, nº 9 (2018): 637-44.

Canguilhem, Georges. *Le normal et le pathologique*. Presses universitaires de France, 1972, 1ère ed : 1943, 224 p.

Carayol, Valérie et Franck Morandi, éd. *Le tournant numérique des sciences humaines et sociales*. Médias et TIC. Pessac, MSH d'Aquitaine, 2019. 132 p.

Cardon, Dominique, Jean-Philippe Cointet et Antoine Mazières. « La revanche des neurones. L'invention des machines inductives et la controverse de l'intelligence artificielle ». *Réseaux* 211, nº 5 (2018): 173-220.

Chamussy, Henri. « D'amour et d'impuissance », Brouillons Dupont, 3:(1978): 67-81

———. « Réflexions sur la formation des concepts en géographie ». *Géographes associés* 1, nº 1 (1996): 51-62.

Chateau-Smith, Carmela. « Changements de paradigme et prosodie sémantique : étude d'un corpus diachronique d'anglais géologique ». Thèse, Lorient, 2012.

Chen, Chaomei. « Visualizing Scientific Paradigms ». *Journal of the American Society for Information Science and Technology* 54, nº 5 (2003): 392-468.

Chen, Chaomei, Timothy Cribbin, Robert Macredie et Sonali Morar. « Visualizing and Tracking the Growth of Competing Paradigms: Two Case Studies ». *Journal of the American Society for Information Science and Technology* 53, nº 8 (2002): 678-89.

Chen, Xiuwen, Jianming Chen, Dengsheng Wu, Yongjia Xie et Jing Li. « Mapping the Research Trends by Co-Word Analysis Based on Keywords from Funded Project ». *Procedia Computer Science* 91 (2016): 547-55.

Claval, Paul. « La réflexion théorique en géographie et les méthodes d'analyse ». *L'Espace géographique* 1, nº 1 (1972): 7-22.

———. « Des essais sur l'épistémologie de la géographie ». *L'Espace géographique* 10, nº 4 (1981): 305.

———. « Quelques orientations actuelles de la réflexion épistemologique en géographie : systemes, structures et metaphores ». *Paralelo 37* 8 (1985): 173-82.

———. *Histoire de la géographie française, de 1870 à nos jours*. Nathan Université, 1998, 543 p.

Claval, Paul et André-Louis Sanguin. *La géographie française à l'époque classique (1918-1968)*. Editions L'Harmattan, 1996, 354 p.

Clavert, Frédéric et Valérie Schafer. « Les humanités numériques, un enjeu historique ». *Quaderni. Communication, technologies, pouvoir*, nº 98 (2019): 33-49.

Clerc, Pascal. « Pour une épistémologie des controverses en géographie ». *Bulletin de l'association de géographes français. Géographies* 97, nº 1/2 (2020): 17-25.

Clerc, Pascal, Florence Deprest, Guilhem Labinal et Didier Mendibil. *Géographies - Épistémologie et histoire des savoirs sur l'espace - 2e éd.*. Armand Colin, 2019. 352 p.

Clivaz, Claire. « Lost in Translation? The Odyssey of "Digital Humanities" in French ». *Studia Universitatis Babes-Bolyai Digitalia*, Digitising the Humanities, 62, nº 1 (2017): 26-41.

Cointet, Jean-Philippe. « La cartographie des traces textuelles comme méthodologie d'enquête en sciences sociales ». HDR, ENS, 2017.

Collectif. « Manifeste des Digital Humanities ». *Journal des anthropologues*, Association française des anthropologues, nº 122-123 (2010): 447-52.

Collectif, Débat. « Le postmodernisme en géographie ». *L'espace géographique* 33, nº 1 (2004): 6-37.

Comte, Auguste. *Cours de philosophie positive*. Bachelier, 1830.739 p.

Cribier, Françoise. « La géographie de la récréation en Amérique anglo-saxonne ». *Annales de géographie* 80, nº 442 (1971): 644-65.

Crompton, Constance. *Doing More Digital Humanities*. London. New York, Routledge, 2019, 350 p.

Cuyala Sylvain, « Analyse spatio-temporelle d'un mouvement scientifique : l'exemple de la géographie théorique et quantitative européenne francophone ». Thèse. Paris 1, 2014

Dacos, Marin et Pierre Mounier. « Humanités numériques : États des lieux et positionnement de la recherche française dans le contexte internationale ». Research Report. Institut français, 2015.

Dale, Robert. « GPT-3: What's It Good for ? » *Natural Language Engineering* 27, nº 1 (2021):113-118.

Daley, Robert, John Sinclair, Susan Jones et Ramesh Krishnamurthy. *English Collocation Studies: The OSTI Report*. Continuum, 2004. 224 p.

Daviet, Sylvie. « Trente ans de géographie industrielle dans les Annales de géographie (1970-1999) ». *Annales de géographie* 114, nº 641 (2005): 73-92.

De Angelis, Rossana. « De l'herméneutique matérielle à l'herméneutique digitale ou numérique ». *Texto ! Textes & Cultures* 25, nº 4, hal-03926958, consulté le 26 janvier 2021, 2020

Demangeon, Albert. *La Plaine picarde : Picardie, Artois, Cambrésis, Beauvaisis, étude de géographie sur les plaines de craie du nord de la France*. Armand Colin, 1905, 496 p.

———. « Le Val de Loire, d'après l'ouvrage de R. Dion ». *Annales de géographie* 43, nº 243 (1934): 315-19.

Denizot, Nathalie. « Les humanités, la culture humaniste et la culture scolaire ». *Tréma*, nº 43 (2015): 42-51.

Descartes, René. *Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences*. Bibliothèque nationale. 1888, 1ère ed : 1637

Desrosières, Alain. *La politique des grands nombres : Histoire de la raison statistique*. La Découverte, 2016, 666 p.

Deuff, Olivier le, et Frédéric Clavert. « Petite histoire des humanités digitales ». In *Les temps des humanités digitales. La mutation des sciences humaines et sociales*. Fyp éditions, (2014): 15-31

Devlin, Jacob, Ming-Wei Chang, Kenton Lee et Kristina Toutanova. « BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding ». *arXiv,* 2018, consulté le 7 mars 2019.

Dipanjan, Sarkar, « Implementing Deep Learning Methods and Feature Engineering for Text Data: The GloVe Model », billet de blog, https://www.kdnuggets.com/implementing-deep-learning-methods-and-feature-engineering-for-text-data-the-glove-model.html/, consulté en sept 2019, 2018

Dodge, Martin et Rob Kitchin. « Codes of Life: Identification Codes and the Machine-Readable World ». *Environment and Planning D: Society and Space* 23, nº 6 (2005): 851-81.

Domingos, Pedro. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* . Basic Books, 2015, 352 p.

Drozd, Aleksandr, Anna Gladkova, et Satoshi Matsuoka. « Word Embeddings, Analogies, and Machine Learning: Beyond King - Man + Woman = Queen ». In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, (2016): 3519-30.

Drucker, Johanna. *Visualisation. L'interprétation modélisante*. B42 éd., 2020, 200 p.

Duboscq, Pierre. « La mobilité rurale en Aquitaine, Essai d'analyse logique ». *L'Espace géographique* 1, nº 1 (1972), 23-42.

Dubossarsky, Haim, Daphna Weinshall et Eitan Grossman. « Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models ». In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, (2017): 1136-45.

Dugué, Nicolas, Jean Charles Lamirel et Pascal Cuxac. « Visualisation pour la détection d'évolutions dans des corpus de publications scientifiques. Indexation, classification et analyse diachronique pour la visualisation ». *Les Cahiers du numérique* 12, nº 4 (2016a): 157-84.

Dugué, Nicolas, Jean-Charles Lamirel et Pascal Cuxac. « Diachronic'Explorer: Keep Track of Your Clusters ». In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, Grenoble, (2016b):1-2

Ekeland, Ivar. *Le Calcul, l'Imprévu. Les figures du temps de Kepler à Thom*. Média Diffusion, 2015, 131 p

Fabiani, Jean-Louis. « Du chaos des disciplines à la fin de l'ordre disciplinaire ? » *Pratiques. Linguistique, littérature, didactique*, nº 153-154 (2012): 129-40.

Falardeau, Jean-Charles. « Géographie humaine et sociologie ». *Géocarrefour* 25, nº 4 (1950): 342-46.

Farge, Arlette et Sean Takats, *Le goût de l'archive à l'ère numérique*, vidéo, https://gout-numerique.net/645, 2018, consulté le 2 sept 2022

Fares, Murhaf, Andrey Kutuzov, Stephan Oepen et Erik Velldal. « Word Vectors, Reuse, and Replicability: Towards a Community Repository of Large-Text Resources ». In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, Gothenburg, (2017): 271-76.

Febvre Lucien. *La Terre et l'évolution humaine*, Paris, 1922, 471 p.

Fel, André. « Deux géographies humaines ? » *L'Espace géographique* 1, nº 2 (1972): 107-112.

Ferret, Olivier. « Similarité sémantique et extraction de synonymes à partir de corpus ». dans *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles*, (2010): 31-40

Ferrier, Jean-Paul, Jean-Bernard Racine et Claude Raffestin. « Vers un paradigme critique : matériaux pour un projet géographique ». *L'Espace géographique* 7, nº 4 (1978): 291-97.

Filleron, Jean-Charles. « "Paysage", pérennité du sens et diversité des pratiques ». *Actes Sémiotiques*, https://www.unilim.fr/actes-semiotiques/1265, consulté en sept 2021, 2008

Firth, John Rupert. *Papers in Linguistics, 1934-1951*. Oxford University Press, 1957, 233 p.

Fleck, Ludwik. *Genèse et développement d'un fait scientifique*. Paris, Flammarion, 2008, 280 p.

Foucault, Michel. *Les mots et les choses. Une archéologie des sciences humaines*. Editions Gallimard, 1990, 1ère ed: 1966, 407 p.

Frermann, Lea, et Mirella Lapata. « A Bayesian Model of Diachronic Meaning Change ». *Transactions of the Association for Computational Linguistics* 4 (2016): 31-45.

Funtowicz, Silvio O., et Jerome R. Ravetz. « Science for the Post-Normal Age ». *Futures* 25, nº 7 (1993): 739-55.

Gabriel, Markus. *Pourquoi le monde n'existe pas*. Jean-Claude Lattès, 2014,118 p.

Gadamer, Hans-Georg. *La Philosophie herméneutique*. Paris: PUF, 1996, 264 p.

Gale, Stephen. « On the Heterodoxy of Explanation: A Review of David Harvey's Explanation in Geography ». *Geographical Analysis* 4, nº 3 (1972): 285-322.

García Robles, Ana, Sonja Zillner, Wolfgang Gerteis, Gabriella Cattaneo, Andreas Metzger, Daniel Alonso, Martina Barbero, Ernestina Menasalvas, et Edward Curry. « Achievements and Impact of the Big Data Value Public-Private Partnership: The Story so Far ». In *The Elements of Big Data Value: Foundations of the Research and Innovation Ecosystem*, (2021): 63-93.

Gaudin, Solène. « Le temps de l'engagement, enjeux et développement d'une géographie appliquée (1970-1980) ». *Bulletin de l'association de géographes français. Géographies* 92, nº 1 (2015): 111-25.

Geertz, Clifford. « Thick Description: Towards an Interpretive Theory of Culture ». In *The Cultural Geography Reader*, Routledge, (2008): 41-51.

George, Pierre. « Existe-t-il une géographie appliquée ? » *Annales de géographie* 70, nº 380 (1961): 337-46.

———. « Géographie et urbanisme ». *Annales de géographie* 74, nº 406 (1965): 641-59.

George, Pierre, et Fernand Verger. *Dictionnaire de la géographie*. 4e éd. poche mise à jour. Quadrige. Paris : PUF. 2013

Ginsburger, Nicolas. « Des îles grecques à la géographie coloniale : Marcel Dubois à la conquête de la Sorbonne (1876-1895) », revue *Cybergeo,* https://journals.openedition.org/cybergeo/28368, mis en ligne en 2017, consulté le 5 nov2020.

———. « Théodore Lefebvre, un bon géographe pour Poitiers ? Identités sociale, professionnelle et disciplinaire entre stratégies, rivalités et affinités dans la géographie française de l'entre-deux-guerres (1933-1934) ». *Norois* 230, nº 1 (2014): 7-19.

Ginzburg, Carlo. *Mythes, emblèmes, traces : Morphologie et histoire*. Verdier, 2010. 376 p.

———. « Signes, traces, pistes. Racines d'un paradigme de l'indice ». *Le Débat*, nº 6 (1980): 3-44.

Giusti, Christian. « Géologues et géographes français face à la théorie davisienne (1896-1909) : retour sur "l'intrusion" de la géomorphologie dans la géographie ». *Géomorphologie : relief, processus, environnement* 10, nº 3 (2004): 241-254.

Gladkova, Anna, Aleksandr Drozd, et Satoshi Matsuoka. « Analogy-Based Detection of Morphological and Semantic Relations with Word Embeddings: What Works and What Doesn't. » In *Proceedings of the NAACL Student Research Workshop*, San Diego, (2016) : 8-15.

Gold, Matthew K. *Debates in the Digital Humanities*. University of Minnesota Press, 2012. 533 p.

Gonen, Hila, Ganesh Jawahar, Djamé Seddah, et Yoav Goldberg. « Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora ». In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (2020):538-55.

Gould, Peter. « The New Geography: Where the Action Is ». *Harper's Magazine*, 1968.

Grafton, Anthony. « Loneliness and Freedom ». *Perspectives on history* 49 (2011).

Grandjean Bernard, « Coup d'œil sur l'école française de géographie au milieu du XX^{ème} siècle », *Geographica Helvetica*, (1957): 45-56.

Granger, Gilles Gaston. *Sciences et Réalité*. Odile Jacob, 2001, 232 p.

Graves, Norman J. « Can Geographical Studies Be Subsumed under One Paradigm Or Are a Plurality of Paradigms Inevitable? » *Terra* 93, nº 3 (1981): 85-90.

Groulx, Lionel-H. « Querelles autour des méthodes ». *Socio-anthropologie*, nº 2 (1997), http://journals.openedition.org/socio-anthropologie/30, consulté le 2 mars 2022

Guérin-Pace, France, Thérèse Saint-Julien et Anita W. Lau-Bignon. « Une analyse lexicale des titres et mots-clés de 1972 à 2010 ». *L'Espace Géographique* Tome 41, nº 1 (2012): 4-30.

Guichard, Éric. « L'internet et les épistémologies des sciences humaines et sociales ». *Revue Sciences/Lettres*, nº 2, (2014), http://journals.openedition.org/rsl/389, consulté le 8 janv 2018.

Guilhaumou, Jacques. « Le corpus en analyse de discours : perspective historique ». *Corpus*, nº 1 (2002), https://journals.openedition.org/corpus/8, consulté le 18 mars 2019.

Gyuris, Ferenc, Boris Michel, et Katharina Paulus. *Recalibrating the Quantitative Revolution in Geography: Travels, Networks, Translations*. Routledge, 2022, 245 p.

Hacking, Ian. « Work in a New World : The Taxonomic Solution ». Dans *World Changes. Thomas Kuhn and the Nature of Science*. MIT Press. (1993): 275-310

———. *Entre science et réalité*. Traduit par Baudoin Jurdant. Paris : La Découverte. 2008, 308 p

Hamilton, William L., Jure Leskovec, et Dan Jurafsky. « Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change ». In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, (2016): 1489-1501.

Harris, Zellig Sabbettai. « Analyse du discours ». Traduit par Françoise Dubois-Charlier. *Langages* 4 13 (1969): 8-45.

Harvey, David. *Explanation in Geography*. London : Hodder & Stoughton Educational. 1969

———. « Revolutionary and Counter Revolutionary Theory in Geography and the Problem of Ghetto Formation ». *Antipode* 4, nº 2 (1972): 1-13.

Hey, Tony. « The Fourth Paradigm – Data-Intensive Scientific Discovery ». In *E-Science and Information Management*. Communications in Computer and Information Science. Berlin, Springer, (2012):1334-1337

Hitchcock, Tim. « Historyonics: Big Data for Dead People: Digital Readings and the Conundrums of Positivism »,http://historyonics.blogspot.com/2013/12/big-data-for-dead-people-digital.html 2013, consulté le 15 avril 2021.

Jean, Aurélie. *De l'autre côté de la machine : Voyage d'une scientifique au pays des algorithmes*. Humensis, 2019. 145 p.

Jensen, Pablo. *Pourquoi la société ne se laisse pas mettre en équations*. Média Diffusion, 2018. 289 p.

Johnston, R. J. « Paradigms and Revolutions or Evolution ? : Observations on Human Geography since the Second World War ». *Progress in Human Geography*, (1978): 189-206.

Jones, Steven E. *The Emergence of the Digital Humanities*. New York: Routledge, 2013, 224 p.

Juillard, Étienne. « Henri Baulig (1877-1962) ». *Annales de géographie* 71, nº 388 (1962a): 561-66.

———. « La région : essai de définition ». *Annales de géographie* 71, nº 387 (1962b): 483-99.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, et Slav Petrov. « Temporal Analysis of Language through Neural Language Models », Baltimore, (2014): 61-65

Kirschenbaum, Matthew G. « What Is Digital Humanities and What's It Doing in English Departments? » In *Debates in the Digital Humanities*, Routledge, (2012): 211-220

Kitchin, Rob. « Big Data, New Epistemologies and Paradigm Shifts ». *Big Data & Society* 1, nº 1, https://journals.sagepub.com/doi/10.1177/2053951714528481, (2014), consulté le 25 mars 2020.

Kitchin, Rob et Gavin McArdle. « What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets ». *Big Data & Society* 3, https://journals.sagepub.com/doi/full/10.1177/2053951716631130, (2016), consulté le 25 janvier 2021.

Kleymann, Rabea, Andreas Niekler et Manuel Burghardt. « Conceptual Forays: A Corpus-Based Study of "Theory" in Digital Humanities Journals ». *Journal of Cultural Analytics* 7, nº 4, (2022) https://culturalanalytics.org/article/55507 consulté le 8 septembre 2023.

Knafou, Rémy. « Les stations intégrées de sports d'hiver des Alpes françaises : l'aménagement de la montagne à la "française" ». Masson, 1978, 319 p.

Kuhn, Thomas S. *La structure des révolutions scientifiques*. Traduit par Laure Meyer. Paris:,Flammarion, 2008, 1ère ed : 1962, 284 p.

———. « Commensurability, Comparability, Communicability ». *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, (1982): 669-88.

———. « Metaphor in Science ». In *Metaphor and Thought*, édité par Andrew Ortony, 2e éd., Cambridge University Press, (1993): 533-542

———. *The Road since Structure: Philosophical Essays, 1970-1993*. Édité par James Conant et John Haugeland, University of Chicago Press, 2000. 336 p.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi, et Steven Skiena. « Statistically Significant Detection of Linguistic Change ». In *Proceedings of the 24th International Conference on World Wide Web*, (2015): 625-635.

Labasse, Jean. *Les capitaux et la région : étude géographique : Essai sur le commerce et la circulation des capitaux dans la région lyonnaise*. Presses de Sciences Po, 2012, 562 p.

Lacoste, Yves. *De la géopolitique aux paysage s: Dictionnaire de la géographie*. Armand Colin. 2003

Ladrière, Jean. *Les limitations internes des formalismes. Etude sur la signification du théorème de Gödel et des théorèmes apparentés dans la théorie des fondements des mathématiques*. Gauthier-Villars. Paris, 1957.

Lafon, Pierre. *Dépouillements et statistiques en lexicométrie*. Slatkine, 1984, 268 p.

Lakatos, Imre. *Histoire et Méthodologie des sciences*. Presses Universitaires de France, 1994, 272 p.

Lakatos, Imre et Alan Musgrave, éd. *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science*. Vol. 4. Cambridge: Cambridge University Press, 1970. 145 p.

Lakmal, Dimuthu, Surangika Ranathunga, Saman Peramuna et Indu Herath. « Word Embedding Evaluation for Sinhala », (2020): 1874-81

Lamirel, J.-C. et S. Al Shehabi. « MultiSOM: A Multiview Neural Model for Accurately Analyzing and Mining Complex Data ». In *Fourth International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'06)*, London IEEE, (2006): 42-54.

Laney, Doug. « 3d Data Management Controlling Data Volume Velocity And Variety ». Meta group research, vol. 6, no 70, (2001):1

Langlais, Pierre Carl. « #DHIHA8 Humanités numériques : et si nous avions créé une nouvelle discipline ? » Dans *Sciences communes*, https://scoms.hypotheses.org/998. Consulté en janv 2020. 2019

———. « La recherche en crise de reproductibilité ? » *EPRIST Analyse I/IST*, nº 30. https://www.eprist.fr/wp-content/uploads/2020/04/EPRIST_I-IST_Recherche-en-crise-de-reproductibilite_Avril2020.pdf. Consulté le 18 mars 2021. 2020

———. « ChatGPT : comment ça marche ? » Billet de blog ; *Sciences communes*. https://scoms.hypotheses.org/1059. Consulté et publié le 7 fev 2023.

Langlois, Charles-Victor et Charles Seignobos. *Introduction aux études historiques*. ENS Éditions. 1ère édition : 1897. 2014. 180 p.

Larivière, Vincent, Éric Archambault, Yves Gingras et Étienne Vignola-Gagné. « The Place of Serials in Referencing Practices: Comparing Natural Sciences and Engineering with Social Sciences and Humanities ». *Journal of the American Society for Information Science and Technology* 57, nº 8 (2006): 997-1004.

Latour, Bruno. *La science en action : Introduction à la sociologie des sciences*. Traduit par Michel Biezunski. Paris : La Découverte. 2005. 662 p.

Latour, Bruno, *Aramis ou l'amour des techniques*. La Découverte. 2020. 448 p.

Latour, Bruno, Steve Woolgar et Michel Biezunski. *La vie de laboratoire : La production des faits scientifiques*. Paris, La Découverte. 2006. 308 p.

Le Roy Ladurie, Emmanuel. 1968. « La fin des érudits ». *Le Nouvel Observateur*.

Leduc, Pierre. « La notion d'incommensurabilité chez Thomas Kuhn ». Thèse. Université du Québec. 2007. 429p.

Lefort, Isabelle. « Références scientifiques et préférences littéraires. Pour un déchiffrement brunetien ». *Géocarrefour*, vol. 78/1. (2003): 79-88.

———. « Olivier Orain, De plain–pied dans le monde ». *Géocarrefour*, vol. 86/3-4. (2011): 234-36.

———. « Préface ». Dans *Principes de géographie humaine* par Paul Vidal de la Blache. Bibliothèque idéale des sciences sociales. Lyon : ENS http://books.openedition.org/enseditions/3694 consulté le 10 sept 2018, 2015

Leininger-Frézal, Caroline. « Olivier Orain, De plain-pied dans le monde ». *Géocarrefour*, vol. 86/3-4. (2011): 233-34.

Lejeune, Christophe. *Manuel d'analyse qualitative*. De Boeck Supérieur. 2019. 162 p.

Lemercier, Claire « TransNum - Une deuxième vague de quantification en histoire économique ? - Discussion et conclusion ? », Vidéo Siences Po, https://www.youtube.com/watch?v=refDXCkYkN8. consulté en janv 2021, 2020.

Lemieux, Cyril. « Peut-on ne pas être constructiviste ? » *Politix* 100 /4, (2012): 169-87.

Lenclud, Gérard. « La statue du commandeur (note critique) ». *Annales* 48/5, (1993): 1221-30.

Lepetit, Bernard. « Une logique du raisonnement historique (note critique) ». *Annales* 48/5, (1993) : 1209-19.

Levain, Alix, Florence Revelin, Anne-Gaëlle Beurier et Marianne Noël. « La crédibilité des matériaux ethnographiques face au mouvement d'ouverture des données de la recherche ». *Revue d'anthropologie des connaissances* 17/2, mis en ligne le 7 avril 2023, https://journals.openedition.org/rac/30291, consulté le 10 mai 2023.

Lévi-Strauss, Claude. *Les Structures élémentaires de la parenté*. Berlin : De Gruyter Mouton. 2002, 1ère ed : 1949, 624 p.

Lévy, Jacques, et Michel Lussault. *Dictionnaire de la géographie et l'espace des sociétés*. Paris : Belin. 2003.

Libera, Alain de. *La querelle des universaux : de Platon à la fin du Moyen Age*. Seuil. 1996. 522 p.

Liu, Alan. « Digital Humanities and Academic Change ». *English Language Notes* 47, (2009): 17-35.

Liu, Yong, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, et Vassilis Kostakos. « CHI 1994-2013 : mapping two decades of intellectual progress through co-word analysis ». Dans *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. Toronto, Ontario, Canada : ACM Press. (2014): 3553-62.

Lohr, Steve. « For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights ». *The New York Times* 24/2, (2014)

Longhi, Julien. « Contours, perspectives et tensions des "humanités numériques" ». *Sens-Dessous* 24/2, (2019): 123-35.

Longhi, Julien, et André Salem. « Approche textométrique des variations du sens ». Dans *Journées d'Analyse de Données Textuelles*. (2018): 452-58.

Mair, Andrew. « Thomas Kuhn and Understanding Geography ». *Progress in Human Geography* 10/3, (1986): 345-69.

Margner, Volker, Umapada Pal et Apostolos Antonacopoulos. *Document Analysis And Text Recognition: Benchmarking State-of-the-art Systems*. World Scientific. 2018, 303 p.

Marr, Bernard. « Big Data: The 5 Vs Everyone Must Know ». LinkedIn Pulse, 6, 2014.

Marz, Nathan et James Warren. *Big Data : Principles and best practices of scalable realtime data systems.* Shelter Island, NY : Manning Publications. 2015. 328 p.

Massardier, Gilles. « Les savants les plus "demandés". Expertise, compétences et multipositionnalité. Le cas des géographes dans la politique d'aménagement du territoire ». *Politix. Revue des sciences sociales du politique* 9/36, (1996) : 163-80.

Masterman, Margaret. « The Nature of a Paradigm ». Dans *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. Sous la direction de Alan Musgrave et Imre Lakatos, 4 : 59-90. Cambridge : Cambridge University Press. (1970):59-90.

Mayaffre, Damon. « Les corpus réflexifs : entre architextualité et hypertextualité ». *Corpus*, nº 1, https://journals.openedition.org/corpus/11, mis en ligne en 2002, consulté le 7 mars 2017.

Mayer-Schönberger, Viktor. *Big Data: A Revolution That Will Transform How We Live*. London, GBR : John Murray Publishers. 2013

———. « La révolution Big Data ». *Politique étrangère* Hiver 4 : (2014):69-81.

McCarty, Willard. « Humanities Computing ». Dans *Encyclopedia of Library and Information Science* , (2003):1124-35.

Ménissier, Thierry, Pierre-Carl Langlais, Didier Schwab et Max Beligné, *ChatGPT : Données, méthodes et enjeux*. Canal-U vidéo, https://www.canal-u.tv/chaines/progedo/chatgpt-donnees-methodes-et-enjeux, consulté le 8 sept 2023

Merleau-Ponty, Maurice. *L'oeil et l'esprit*. Gallimard. 1964. 106 p.

Meunier, Jean-Guy. « Humanités numériques ou computationnelles : enjeux herméneutiques ». *Sens public*. (2014):2-25

———. « Le paradoxe des humanités numériques ». *Quaderni. Communication, technologies, pouvoir*, nº 98. (2019): 19-31.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, *et al*. « Quantitative Analysis of Culture Using Millions of Digitized Books ». *Science* 331. (2011) : 176-82.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado et Jeff Dean. « Distributed Representations of Words and Phrases and their Compositionality ». *Advances in neural information processing systems* 26. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf mis en ligne en 2013, consulté le 18 fev 2018

Montariol, Syrielle, et Alexandre Allauzen. 2019. « Learning dynamic word embeddings with drift regularisation ». http://arxiv.org/abs/1907.09169 mis en ligne en 2019, consulté le 20 mars 2020

Morandi, Franc, Claude Baltz, et Éric Delamotte.. *Humanités numériques : Regards épistémologiques et critiques*. ISTE Group. 2021, 334 p.

Morgan, Mary, et Margaret Morrison. « Models as Mediators: Perspectives on Natural and Social Science », Cambridge University Press, 1999

Morin, Edgar. « Pour une crisologie ». *Communications* 25/1. (1976): 149-63.

Mounier, Pierre, dir. *Read/Write Book 2 : Une introduction aux humanités numériques*. Marseille : OpenEdition Press. 2012. 264 p.

———. *Les humanités numériques : Une histoire critique*. Paris, MSH. 2018. 176 p.

Naylor, Simon, James Ryan, Ian Cook, David Crouch, James Ryan, Ian Cook et David Crouch. 2018. *Cultural Turns/Geographical Turns : Perspectives on Cultural Geography*. Routledge. 2018. 404 p.

Nowviskie, Bethany. « On the Origin of "Hack" and "Yack" ». Dans *Debates in the Digital Humanitie* , University of Minnesota Press, (2016):66-70.

Nyhan, Julianne et Andrew Flinn. *Computation and the Humanities. Towards an Oral History of Digital Humanities*. Springer Nature. 2016, 285 p.

Orain, Olivier. « Le plain-pied du monde : postures épistémologiques et pratiques d'écriture dans la géographie française au vingtième siècle ». Thèse, https://theses.hal.science/tel-00082408/file/Orain_these.pdf , mise en ligne en 2006, consulté le 7 janv 2017, Paris 1. 2003

———. « Misère du possibilisme ». *Le blog d'Olivier Orain*. http://www.esprit-critique.net/article-10489023.html , consulté le 20 mai 2020, mis en ligne en 2007a

———. « Olivier Orain : itinéraire de recherche ». *Le blog d'Olivier Orain*. http://www.esprit-critique.net/article-10089272.html , consulté le 20 mai 2020, mis en ligne en 2007b

———. *De plain-pied dans le monde : Ecriture et réalisme dans la géographie française au XXe siècle*. Editions L'Harmattan. 2009. 435 p

———. « La fabrique d'un livre : réponse et discussion ». *Géocarrefour*, vol. 86/3-4 : (2011): 237-40.

———.. « Mai 68 et ses suites en géographie française ». *Revue d'histoire des sciences humaines*, nº 26 (2015): 209-42..

———. « Une histoire de la géographie au prisme des sciences humaines ». HDR, Université Paris 1 Panthéon-Sorbonne, 2021

———. « A social history of quantitative geography in France from the 1970s to the 1990s: an overview of the blossoming of a multifaceted tradition ». Recalibrating the Quantitative Revolution in Geography. Travels, Networks, Translations. London : Routledge. (2022): 102-17

Pailhé, Joël. « Sur l'objet de la géographie de la population ». *L'Espace géographique* 1/1 (1972): 54-62.

———. « Pierre George, la géographie et le marxisme ». *Espace Temps* 18/1, (1981): 19-29.

Pashler, Harold, et Eric-Jan Wagenmakers..« Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? » *Perspectives on Psychological Science*. vol 7/6, (2012): 528-530

Passeron, Jean-Claude. *Le Raisonnement sociologique : Un espace non poppérien de l'argumentation*. Albin Michel. 1ère édition :1991, 2006, 674 p.

———. « La forme des preuves dans les sciences historiques ». *Revue européenne des sciences sociales. European Journal of Social Sciences*, (2001): 31-76

———. « Logique formelle, schématique et rhétorique ». Dans *L'argumentation : Preuve et persuasion*. Enquête. Paris : Éditions de l'École des hautes études en sciences sociales. (2020): 149-81

Pélissier, Paul. « Pierre Gourou, 1900-1999 ». *Annales de géographie.* vol 109/612. (2000): 212-17.

Pennington, Jeffrey, Richard Socher, et Christopher Manning. « Glove: Global Vectors for Word Representation ». Dans *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Association for Computational Linguistics. (2014): 1532-43

Pentland, Alex. *Social Physics: how good ideas spread - the lessons from a new science*. New edition. Melbourne : Scribe Publications. 2014. 320 p.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, et Luke Zettlemoyer. « Deep contextualized word representations ». Dans *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. (2018): 2227-37.

Philipponneau, Michel. « La Commission de géographie appliquée et le développement des applications de la géographie ». Dans *Géographes face au monde*. L'Harmattan. (1996): 271-283.

Pigeon, Patrick. « Réflexions sur les notions et les méthodes en géographie des risques dits naturels ». *Annales de géographie.* vol 111/627. (2002): 452-70.

Pimentel, Joao Felipe, Leonardo Murta, Vanessa Braganholo, et Juliana Freire. « A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks ». Dans *IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*.. Montreal, (2019): 507-17.

Pincemin, Bénédicte. « Diffusion ciblée automatique d'informations : conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents ». Thèse. Université Paris IV. 1999a.

———. « Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ? » *Sémiotiques* 17. (1999b): 71-120.

Pinchemel, Philippe. Lire la face de la Terre : de la géographie à la géonomie Entretien réalisé par Nicole Mathieu et Jean-Louis Tissier. *Natures Sciences Sociétés*, vol5/4. (1997): 47-54

Pinchemel, Philippe, et Geneviève Pinchemel. *La face de la Terre: éléments de géographie*. Armand Colin. 1988: 519 p.

Pluet-Despatin, Jacqueline. « Une contribution à l'histoire des intellectuels : les revues ». Dans *La belle époque des revues*. vol 20/1, (1992):125-136

Poncet, Patrick. *Intelligence spatiale*. Presses Universitaires de Rennes. 2017. 320 p.

Popper, Karl. *La connaissance objective : Une approche évolutionniste*. Paris. 1972. 580 p.

Pumain, Denise et Marie-Claire Robic, « Le rôle des mathématiques dans une "révolution" théorique et quantitative : la géographie française depuis les années 1970 », *Revue d'Histoire des Sciences Humaines*. 6/1. (2002):123-144.

Putnam, Hilary. *Reason, Truth and History*. Cambridge University Press. 1981. 244 p.

———. *Realism with a Human Face*. Harvard University Press. 1992. 426 p.

Quine, Willard Van Orman. *Word and Object, new edition*. 1960. 312 p.

Rabouin, David. *Mathesis universalis : L'idée de « mathématique universelle » d'Aristote à Descartes*. Humensis. 2015. 397 p.

Racine, Jean-Bernard. « Le modèle urbain américain. Les mots et les choses ». *Annales de géographie*, vol 80/440, (1971): 397-427.

———.. « Entre pluralisme et complexité : le rôle des valeurs dans la pratique et l'apport de la géographie humaine. Chronique d'une écriture erranter ». *Revue européenne des sciences sociales. European Journal of Social Sciences*, nº XLIV-134, (2006): 231-45.

Racine, Jean-Bernard, et Claude Raffestin. « Des directions (encore) nouvelles pour la Géographie moderne ». *Annales de géographie* 87/480. (1978): 182-94.

Raffestin, Claude. « Territorialité : concept ou paradigme de la géographie sociale ? » *Geographica Helvetica*, nº 41/2, (1986): 91-96

Raffestin, Claude, et Bernard Elissalde. « Une géographie buissonnière ». *Espace Temps.* 64 /1, (1997): 87-93.

Rastier, François. « Le terme : entre ontologie et linguistique ». *La banque des mots*, nº 7. http://www.revue-texto.net/Inedits/Rastier/Rastier_Terme.html, mis en ligne en 1995, consulté le 9 fev 2020

———. « Sémiotique et sciences de la culture ». *Linx. Revue des linguistes de l'université Paris X Nanterre*, nº 44, (2001): 149-68.

———. *La mesure et le grain. Sémantique de corpus*. Lettres numériques. 2011. 280 p.

———. « Corpus numériques et accès à la culture ». *Questions Vives. Recherches en éducation*, nº 28. https://journals.openedition.org/questionsvives/2421#quotation, mis en ligne en 2018, consulté le 6 fev 2021, 2017.

Raynaud, Dominique. « Le raisonnement expérimental en sociologie ». *Philosophia Scientiæ* 23-2 (2019): 19-46.

Recchia, Gabriel, Ewan Jones, Paul Nulty, John Regan, et Peter de Bolla. 2017. « Tracing Shifting Conceptual Vocabularies Through Time ». Dans *Knowledge Engineering and Knowledge Management*. Springer International Publishing. (2017):19-28

Reclus, Élisée. 1869. *Histoire d'un ruisseau*. J. Hetzel & Cie. Paris. 329 p

Resweber, Jean-Paul. « Les enjeux de l'interdisciplinarité ». *Questions de communication*, nº 19, (2011):171-200.

Reutenauer, Coralie. « Vers un traitement automatique de la néosémie : approche textuelle et statistique ». Thèse, Université de Lorraine. 2012.

Ricoeur, Paul. *Temps et Récit. Le Temps raconté:* Seuil, 1983, 432 p.

Robic, Marie-Claire. « Les petits mondes de l'eau : le fluide et le fixe dans la méthode de Jean Brunhes ». *L'Espace géographique* 17/1 (1988): 31-42.

———. « La stratégie épistémologique du mixte : le dossier vidalien ». *Espace Temps*. 47/1 (1991): 53-66.

———. « L'exemplarité du Tableau de la géographie de la France de Paul Vidal de la Blache ». Dans *Figures du texte scientifique*. Presse Universitaire de France. (2003): 81-107.

Rosset, Clément. *Le Réel. Traité de l'idiotie*. Minuit. 2012. 180 p.

Roth, Camille. « Coévolution des auteurs et des concepts dans les réseaux épistémiques : le cas de la communauté « zebrafish» ». *Revue française de sociologie* 49/3 (2008): 523-58.

Rubenstein, Herbert et John B. Goodenough. « Contextual correlates of synonymy ». *Communications of the ACM* 8/10 (1965): 627-33.

Ruiz, Émilien. 2019. « #DHIHA8 Nous sommes à la croisée des chemins ! » Billet. *Devenir historien-ne*. https://devhist.hypotheses.org/3692. mis en ligne en juin 2019, consulté en juillet 2019.

Rule, Adam, Aurélien Tabard et James D. Hollan. « Exploration and Explanation in Computational Notebooks ». Dans *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, Canada : ACM Press. (2018): 1-12

Rule, Alix, Jean-Philippe Cointet et Peter S. Bearman. « Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014 ». *Proceedings of the National Academy of Sciences* 112/35. (2015): 10837-44.

Ryle, Gilbert. *Collected papers. Volume II collected essays, 1929-1968*. Hutchinson. London. 1971. 560 p

Sabah, Gérard. « Sens et traitements automatiques des langues ». Dans *Ingénierie des langues*. , Hermès science publications. Paris. (2000):77-108

Saussure, Ferdinand de. *Cours de linguistique générale*. Paris : Payot. 1995. 520 p.

Schmitt, Eglantine. « Explorer, visualiser, décider : un paradigme méthodologique pour la production de connaissances à partir des big data ». Thèse, Université de technologie de Compiègne. 2018. 289 p.

Schnapp, Jeffrey, Todd Presner, Peter Lunenfeld et Johanna Drucker. « Manifeste pour des humanités numériques 2.0 ». Traduit par Quentin Julien-Saavedra et Yves Citton. *Multitudes* 59/2 (2015): 181-95.

Schreibman, Susan, Ray Siemens et John Unsworth. *A Companion to Digital Humanities*. Malden, MA : Wiley–Blackwell. 2004. 640 p.

Sen, Procheta, Debasis Ganguly et Gareth Jones. « Word-Node2Vec: Improving Word Embedding with Document-Level Non-Local Word Co-occurrences ». Dans *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. (2019): 1041-1105.

Serres, Michel, Gilles Dowek et Serge Abiteboul. Conversation avec Michel Serres : les mutations du cognitif dans *The conversation*. http://theconversation.com/conversation-avec-michel-serres-les-mutations-du-cognitif-93214. publié le 27 mars 2018, consulté le 12 avril 2018.

Sharrock, Wes et Rupert Read. *Kuhn : Philosopher of Scientific Revolutions*. Malden.. 2002. 248 p.

Simiand, François. « Géographie humaine et sociologie ». Dans *Méthode historique et sciences sociales*. (1909): 243-53.

Soja, Edward W. *Postmodern Geographies: The Reassertion of Space in Critical Social Theory*. Verso. 1989. 276 p.

Soubeyran, Olivier. « Vingt ans déjà : un retour à la case départ ? » *Cahiers de géographie du Québec* 32/87. (1988) : 231-44.

———. « La Géographie Coloniale : Un élément structurant dans la naissance de l'Ecole française de géographie ». *Les enjeux de la tropicalité*, (1989):82-90.

———. *Imaginaire, science et discipline*. L'Harmattan. 1997. 486 p.

Steinberg, Steve. « The ontogeny of RISC ». *Intertek,* 3/5 (1994): 1-10.

Stengers, Isabelle et Ilya Prigogine. *La nouvelle alliance - Métamorphose de la science,* Folio essais. Paris. 1986. 439 p.

Stoddart, D. R. « The paradigm concept and the history of geography ». Dans *Geography, ideology and social concern* , Basil Blackwell, Oxford, (1981): 70-80

Szondi, Peter. *Einführung in die literarische Hermeneutik*. Suhrkamp. 1975. 468 p.

Szymanski, Terrence. « Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings ». Dans *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada. (2017): 448-53.

Testart, Alain. *Essai d'épistémologie*. Christian Bourgois. Paris. 1991. 176 p.

Tiercelin, Claudine. *Hilary Putnam, l'héritage pragmatiste*. Philosophie de la connaissance. Collège de France. Paris. 2013. 74 p.

———. « Pourquoi le pragmatisme implique le réalisme ». *Cahiers philosophiques* 150/3 (2017): 11-34.

Touraine, Alain. 2014. *Le Retour de l'acteur : Essai de sociologie*. Fayard. 124 p.

Tutin, Agnès. « Dans cet article, nous souhaitons montrer que… Lexique verbal et positionnement de l'auteur dans les articles en sciences humaines ». *Lidil*, nº 41, 2010: 15-40.

Union International Geographical. *Comptes Rendus Du Congrés International de Géographie, Amsterdam*. E.J. Brill. 1938: 322 p.

Valéry, Paul. *Mauvaises pensées et autres (édition annotée et indexée)*. BoD, 2020, 286 p.

Vidal de la Blache, Paul. « Leçon d'ouverture du cours de géographie ». *Annales de géographie* 8/38 (1899): 97-109.

———. « Tableau de la géographie de la France ». Dans *Histoire de France depuis les origines jusqu'à la Révolution*. , Hachette et Cie, 1903, 483 p.

———. *Principes de géographie humaine*. ENS Éditions. Bibliothèque idéale des sciences sociales. http://books.openedition.org/enseditions/328, 1ère édition : 1922, consulté en janv 2018, mis en ligne 2015

Volvey, Anne. 2014. « Entre l'art et la géographie, une question (d')esthétique ». *Belgeo. Revue belge de géographie*, nº 3. Consulté le 11 sept 2023, mis en ligne en 2014.

Wallace, Alfred Russel. *The Malay Archipelago : the land of the orang-utan and the bird of paradise : a narrative of travel, with studies of man and nature*. London : Macmillan and Co. 1869. 550 p.

Weber, Max. *Essais sur la théorie de la science. Recueil d'articles publiés entre 1904 et 1917*. Les classiques des sciences sociales. http://classiques.uqac.ca/classiques/Weber/essais_theorie_science/essais_theorie_science.html. 1ère édition : 1904-1917, 2005

Wevers, Melvin, Tom Kenter et Pim Huijnen. « Concepts Through Time: Tracing Concepts in Dutch Newspaper Discourse (1890-1990) using Word Embeddings », 7. 2015

Wheeler, P. B. « Revolutions, Research Programmes and Human Geography ». *Area* 14/1 (1982): 1-6.

Whyte, William Foote. « The social structure of the restaurant ». *American Journal of Sociology* 54/4 (1949): 302-10.

Wilde, Max De, Florence Gillet, Simon Hengchen et Steph van Hooland. 2016. *Introduction aux humanités numériques : méthodes et pratiques,* De Boeck Superieur, 210 p.

Wolff, Denis. 2013. « La pratique de terrain d'un géographe moderne, Albert Demangeon (1872-1940) ». *Belgeo. Revue belge de géographie*. https://journals.openedition.org/belgeo/10791 consulté le 11 sept 2023, mis en ligne en 2013.

Wolfram, Stephen. « What Is ChatGPT Doing … and Why Does It Work? », billet de blog, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/. mis en ligne le 14 fev 2023

Zighed, Djamel Abdelkader. « Les Humanités Numériques : la révolution en Sciences Humaines et Sociales ». *Revue des Nouvelles Technologies de l'Information* RNTI-SHS-2. (2014):1-28.

Zuckerman, Matan, et Mark Last. « Using Graphs for Word Embedding with Enhanced Semantic Relations ». Dans *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing* (2019): 32-41.

Index :

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

V

W

Z

Répertoire des géographes cités

Ce répertoire est décliné en deux présentations :

- Une alphabétique pour chercher rapidement un géographe. Un tout petit nombre n'a pas été intégré du fait de l'impossibilité de trouver facilement leur date de naissance. Quelques entrées correspondent à des auteurs qui se revendiquent historien mais qui ont écrit sur l'histoire de la géographie.
- Une chronologique qui permet d'avoir une vue d'ensemble pour mieux voir la succession et les différentes générations de géographes.

Ce répertoire a été établi début octobre 2023.

Par ordre alphabétique

Auriac Franck (1935-2017)
Bachimon Philippe : 1955
Bailly Antoine (1944-2021)
Bataillon Claude : 1931
Baulig Henri (1877-1962)
Beaude Boris : 1973
Beaujeu-Garnier Jacqueline (1917-1995)
Beligné Max : 1982
Bennafla Karine : 1970
Besse Jean-Marc : 1956
Blanchard Raoul (1877-1965)
Bonnamour Jacqueline (1924-2020)
Boulineau Emmanuelle : 1973
Broc Numa (1934-2017)
Brunhes Jean (1869-1930)
Bruneau Michel : 1940
Brunet Roger : 1931
Calbérac Yann : 1980
Chamussy Henri (1935-2015)
Cholley André (1886-1968)
Christaller Walter (1893-1969)
Claval Paul : 1932
Clerc Pascal : 1960
Cribier Françoise : 1930
Cuyala Sylvain : 1985
Dardel Éric (1899-1967)
Daviet Sylvie : 1958
Demangeon Albert (1872-1940)
Dion Roger (1896-1981)
Diry Jean-Paul : 1947
Dubois Marcel (1856-1916)
Duboscq Pierre : 1937
Dugrand Raymond (1925-2017)
Durand-Dastès François (1931-2021)
Elissalde Bernard : 1951
Falardeau Jean-Charles (1917-1989)
Febvre Lucien (1878-1956)
Fel André (1926-2009)
Ferras Robert (1935-2013)
Ferrier Jean-Paul (1937-2016)
Filleron Jean-Charles : 1947
Gallois Lucien (1857-1941)
George Pierre (1909-2006)
Gibert André (1893-1985)
Ginsburger Nicolas : 1976
Giusti Christian : 1955
Gottmann Jean (1915-1994)
Gourou Pierre (1900-1999)
Guérin-Pace France : 1960
Hartshone Richard (1899-1992)
Harvey David : 1935
Humboldt Alexander von (1769-1859)
Isnard Hildebert (1904-1983)
Johnston Ronald John (1941-2020)
Juillard Étienne (1914-2006)
Kayser Bernard (1926-2001)
Kitchin Rob : 1970
Knafou Rémy : 1948
Labasse Jean (1918-2002)
Lacoste Yves : 1929
Lannou Maurice le (1906-1992)
Lebeau René (1914-1999)
Lefebvre Théodore (1889-1943)
Lefort Isabelle : 1957
Leininger-Frézal Caroline : 1981
Lévy Jacques : 1952
Lussault Michel : 1960
Martonne Emmanuel de (1873-1955)
Méo Guy Di : 1945
Meynier André (1901-1983)
Musset René (1881-1977)
Orain Olivier : 1968
Pailhé Joël : 1943
Pélissier Paul (1921-2010)
Pelletier Philippe : 1956
Philipponneau Michel (1921-2008)
Pigeon Patrick : 1959
Pinchemel Philippe (1923-2008)
Planhol Xavier de (1926-2016)
Poncet Patrick : 1974
Pumain Denise : 1946
Racine Jean-Bernard : 1940
Raffestin Claude : 1936
Reclus Elisée (1830-1905)
Reynaud Alain : 1942
Robic Marie-Claire : 1946
Rochefort Michel (1927-2015)
Sanguin André-Louis : 1945
Sion Jules (1879-1940)
Soja Edward (1940-2015)
Sorre Maximilien (1880-1962)
Soubeyran Olivier : 1952
Stock Mathis : 1970
Stoddart David (1937-2014)
Vallaux Camille (1870-1945)
Verger Fernand (1929-2018)
Vidal de la Blache Paul (1845-1918)
Volvey Anne : 1964
Wolff Denis : 1956

Par ordre chronologique

Alexander von Humboldt (1769-1859)
Elisée Reclus (1830-1905)
Paul Vidal de la Blache (1845-1918)
Marcel Dubois (1856-1916)
Lucien Gallois (1857-1941)
Jean Bruhnes (1869-1930)
Camille Vallaux (1870-1945)
Albert Demangeon (1872-1940)
Emmanuel de Martonne (1873-1955)
Raoul Blanchard (1877-1965)
Henri Baulig (1877-1962)
Lucien Febvre (1878-1956)
Jules Sion (1879-1940)
Maximilien Sorre (1880-1962)
René Musset (1881-1977)
André Cholley (1886-1968)
Théodore Lefebvre (1889-1943)
Walter Christaller (1893-1969)
André Gibert (1893-1985)
Roger Dion (1896-1981)
Éric Dardel (1899-1967)
Richard Hartshone (1899-1992)
Pierre Gourou (1900-1999)
André Meynier (1901-1983)
Hildebert Isnard (1904-1983)
Maurice Le Lannou (1906-1992)
Pierre George (1909-2006)
René Lebeau (1914-1999)
Étienne Juillard (1914-2006)
Jean Gottmann (1915-1994)
Jean-Charles Falardeau (1917-1989)
Jacqueline Beaujeu-Garnier (1917-1995)
Jean Labasse (1918-2002)
Philipponneau Michel (1921-2008)
Paul Pélissier (1921-2010)
Philippe Pinchemel (1923-2008)
Jacqueline Bonnamour (1924-2020)
Raymond Dugrand (1925-2017)
Xavier de Planhol (1926-2016)
Bernard Kayser (1926-2001)
André Fel (1926-2009)
Michel Rochefort (1927-2015)
Fernand Verger (1929-2018)
Yves Lacoste : 1929
Françoise Cribier : 1930
François Durand-Dastès (1931-2021)
Roger Brunet : 1931
Claude Bataillon : 1931
Peter Gould (1932-2000)
Paul Claval : 1932
Numa Broc (1934-2017)
David Harvey : 1935
Robert Ferras (1935-2013)
Henri Chamussy (1935-2015)
Franck Auriac (1935-2017)
Claude Raffestin: 1936
Jean-Paul Ferrier (1937-2016)
David Stoddart (1937-2014)
Pierre Duboscq : 1937
Michel Bruneau : 1940
Jean-Bernard Racine : 1940
Edward Soja (1940-2015)
Ronald John Johnston (1941-2020)
Alain Reynaud : 1942
Augustin Berque : 1942
Joël Pailhé : 1943
Antoine Bailly (1944-2021)
Guy Di Méo : 1945
André-Louis Sanguin: 1945
Marie-Claire Robic : 1946
Denise Pumain : 1946
Vincent Berdoulay : 1947
Jean-Charles Filleron : 1947
Jean-Paul Diry : 1947
Rémy Knafou : 1948
Bernard Elissalde : 1951
Hervé Théry : 1951
Olivier Soubeyran : 1952
Jacques Lévy : 1952
Christian Giusti: 1955
Philippe Bachimon : 1955
Jean-Marc Besse : 1956
Philippe Pelletier : 1956
Denis Wolff : 1956
Isabelle Lefort : 1957
Sylvie Daviet : 1958
Patrick Pigeon : 1959
Michel Lussault : 1960
Pascal Clerc : 1960
Anne Volvey : 1964
Olivier Orain : 1968
Rob Kitchin : 1970
Karine Bennafla : 1970
Mathis Stock : 1970
Emmanuelle Boulineau : 1973
Boris Beaude : 1973
Patrick Poncet : 1974
Nicolas Ginsburger : 1976
Yann Calbérac : 1980
Caroline Leininger-Frézal : 1981
Max Beligné : 1982
Sylvain Cuyala : 1985

Liste des figures :

Liste des tableaux :

Table des matières :