JADT 2020 : 15es Journées internationales d'Analyse statistique des Données Textuelles

Suivez l'évolution de vos mondes lexicaux dans le temps !

Max Beligné[1,2], Isabelle Lefort[1], Sabine Loudcher[2]

[1] Université de Lyon, Lyon 2, EVS UMR 5600

[2] Université de Lyon, Lyon 2, ERIC EA 3083

max.beligne@univ-lyon2.fr ; isabelle.lefort@univ-lyon2.fr ; sabine.loudcher@univ-lyon2.fr

Abstract

To study the evolution of the content of thematic groups over time, a tool, Diachronic'Explorer, originally created to follow the evolution of clusters is presented, adapted and tested to be used on lexical worlds from the Reinert's method. The goal is to improve the visualization compared to what is currently easily achievable using lexical worlds but also to test another calculation method on this subject. An experiment was carried out and the results obtained validate the relevance of this methodological transfer.

Keywords: lexical worlds, diachronic analysis, methodological transfert

Résumé

Pour étudier l'évolution du contenu de thématiques dans le temps, un outil, Diachronic'Explorer, créé à l'origine pour suivre l'évolution de clusters, est présenté, adapté et testé pour servir sur les mondes lexicaux issus de la méthode Reinert. L'objectif est d'améliorer la visualisation par rapport à ce qui est actuellement facilement réalisable en utilisant des mondes lexicaux mais aussi de tester une autre méthode de calcul sur ce sujet. Une expérimentation a été réalisée et les résultats obtenus valident la pertinence de ce transfert méthodologique.

Mots clés : monde lexical, analyse diachronique, transfert méthodologique

1. Introduction

Dans le domaine de l'analyse des données textuelles, l'identification de thématiques et de leurs évolutions est un défi majeur pour comprendre tout corpus diachronique. Plusieurs techniques peuvent être utilisées pour mettre à jour des ensembles thématiques : méthodes de partitionnement de données (clustering) ou modèle thématique (topic model). Notre choix a été d'utiliser une technique très connue dans ce domaine, la méthode Reinert, mais en adaptant ensuite un outil élaboré à partir de résultats utilisant des algorithmes de clustering plus récents[1]. Ce choix prend tout son sens à partir de l'état de l'art réalisé.

Deux approches différentes doivent être en effet distinguées dans l'étude de l'évolution chronologique de thématiques textuelles. La première consiste à effectuer des calculs sur l'ensemble du corpus pour, ensuite, déterminer dans quelle proportion chaque thématique obtenue varie dans le temps. La deuxième, plus complexe, s'appuie au contraire sur des calculs effectués sur chaque époque séparément. Dans ce cas, il est nécessaire de déterminer ultérieurement s'il existe une ressemblance entre une thématique calculée à une époque « n » et une autre calculée à l'époque « n+1 » pour établir des continuités chronologiques. Si cette

[1] Jean-Charles Lamirel qui a pensé à l'origine Diachronic'Explorer utilise en amont un algorithme de gaz neuronal (« Growing Neural Gas ») pour déterminer des thématiques.

deuxième méthode offre l'avantage de donner à voir l'évolution du contenu d'une thématique, elle n'est pas sans poser quelques difficultés.

La première est de choisir l'indicateur établissant s'il y a une ressemblance ou non entre deux thématiques. Dans le cadre de la méthode Reinert, ce qui a été jusqu'ici utilisé est l'utilisation des distances de Labbé entre chaque classe obtenue (Ratinaud, 2015). Or, recourir à cet unique indicateur a été fortement critiqué (Brunet, 2004). Si l'objectif était logiquement dans un premier temps de proposer un processus de traitement, valable et fonctionnel, et non de multiplier les indicateurs, il est temps de pouvoir comparer ces résultats avec d'autres issus de méthodologies différentes. La deuxième difficulté importante concerne la visualisation. Toujours dans le cadre de la méthode Reinert, la solution trouvée (Ratinaud, 2015) passe par la production d'un graphique arboré obtenu grâce à une classification hiérarchique ascendante (méthode de Ward) sur la matrice des distances entre thématiques. Or, cette présentation ne rend pas la lecture des continuités chronologiques aisée (cf figure 7). De plus, l'appréhension de l'évolution du contenu des thématiques ne peut se faire qu'indirectement par des allers retours entre le graphique arboré et chaque thématique.

Face à ce problème, la recherche plus récente sur les clusters a permis d'expérimenter d'autres représentations comme l'utilisation des diagrammes alluviaux. Ces graphiques sont un type particulier de diagramme de Sankey adapté pour les représentations des dynamiques temporelles. Ils ont été utilisé dès 2010 par Rosvall et Bergstrom pour étudier l'évolution des clusters de citations scientifiques (Rosvall et Bergstrom, 2011). L'article Diachronic'Explorer : keep track of your clusters ! (Dugué et al., 2016a) les utilise en proposant plus largement une méthode permettant de déterminer s'il existe une continuité entre un cluster calculé à une époque « n » et un autre calculé à une époque « n+1 ». Cette proximité d'objectif, alliée au besoin de résultats issus d'autres méthodologies et à une représentation semblant plus adéquate, a conduit à essayer de transférer cet outil.

Ce travail a finalement mené plus loin que ce qui était initialement prévu, c'est-à-dire de simples adaptations d'un outil pour travailler sur les mondes lexicaux issus de la méthode Reinert puisqu'il a finalement abouti à une reformulation théorique du principe fondateur de Diachronic'Explorer. Une présentation de la méthode originelle sera effectuée pour faire comprendre comment elle fonctionne et la raison d'être des changements effectués. Ensuite, nous présenterons une expérimentation sur un corpus de 4000 articles de géographie issus de deux revues (Les *Annales de Géographie* et l'*Espace Géographique*) sur la période 1892-2000. Cet exemple permet de présenter les avantages de cette approche tout en interrogeant d'éventuelles différences par rapport aux résultats obtenus avec la méthodologie traditionnelle (Labbé + Ward)[2].

2. La méthodologie de Diachronic'Explorer appliquée aux mondes lexicaux

Cette partie vise à comprendre comment marche la méthode de Diachronic'Explorer et comment elle peut être appliquée aux mondes lexicaux. Une place importante a été donnée à la présence d'exemples pour faciliter la compréhension.

2.1. La sélection des variables par la F-mesure de trait

Diachronic'Explorer propose un calcul spécifique, appelé « F-mesure de trait », afin de sélectionner les variables les plus représentatives d'un cluster. Cette F-mesure est définie par la

[2]Cette abréviation renvoie à l'ensemble méthodologique explicité ci dessus (paragraphe 3 de l'introduction).

moyenne harmonique d'un « rappel de trait » et d'une « prépondérance de trait ». L'exemple de la figure1 illustre ces calculs à partir de deux classes étiquetées (M :Masculin / F :Féminin) et trois variables (taille des pieds, longueur de cheveux et taille du nez). Les différents rectangles entourent les chiffres dont les sommes ont été réalisées et correspondent dans les formules aux valeurs soulignées de la couleur correspondante.

FR(P,M) correspond à la mesure de rappel pour la variable taille des pieds chez les hommes. Le rappel de trait d'une variable pour un cluster est égal à la somme des valeurs de cette variable pour ce cluster divisé par la somme des valeurs de cette variable pour tous les clusters. FP(P,M) correspond à la mesure de prédominance pour la variable taille des pieds chez les hommes. La prédominance de trait d'une variable pour un cluster est égale à la somme des valeurs de la variable pour ce cluster divisé par la somme des valeurs de toutes les variables pour ce cluster. La moyenne harmonique calculée à l'aide de FR(P,M) et FP(P,M), appelée FF(P,M) est F-mesure de la variable taille des pieds pour les hommes.

Une application aux mondes lexicaux issus de la méthode Reinert est présentée à la figure 2. Sans rentrer dans les détails qui peuvent être trouvés dans l'article originel (Reinert, 1983), il est nécessaire de préciser que le corpus est découpé en segments de textes. Chaque terme du corpus (T), mot ou lemme selon le choix de l'utilisateur, est présent (1) ou absent (0) de chaque segment de texte formant les lignes du tableau de la figure 2. A partir de ces données, la méthode Reinert établit les mondes lexicaux. Les termes du corpus correspondent par rapport à l'exemple précédent aux variables. La F-mesure du premier terme (T1) pour le premier monde lexical (ML1) est facilement calculable en suivant l'exemple de la figure 1.

T1	T2	T3	T4	…	Monde lexical
1	0	1	1	…	ML 1
1	0	0	1	…	ML 1
0	0	1	0	…	ML 2
…	…	…	…	…	…

Fig 1 : Un exemple de calcul d'une F-mesure tiré de N Dugué, 2016

Fig 2 : Application du calcul de la F-mesure à une construct

En calculant les F-mesure pour chaque variable (ou terme) et chaque classe (ou monde lexical), il est possible d'obtenir à partir de la figure 1 le tableau présenté à la figure 3. F(x,M) est la F-mesure pour la variable x affichée en ligne concernant les hommes. F(x,F) est une mesure similaire effectuée pour les femmes. La colonne suivante (F(x,.)) est une moyenne des deux résultats précédents. Enfin, F(.,.) est la moyenne des moyenne.

	F(x,M)	F(x,F)	F(x,.)
Longueur Cheveux	0.3	0.66	0.48
Taille Pieds	0.48	0.22	0.35
Taille Nez	0,3	0,24	0,27
			F(…) 0.38

Une variable est retenue comme caractérisant une classe si sa F-mesure pour la classe est supérieure à la moyenne (F(x,.)) et à la moyenne des moyennes (F(.,.)). Ainsi, la longueur des cheveux sera retenue pour les femmes (0,66>0,53 et 0,66>0,38) ; la taille des pieds sera retenue pour les hommes ; La taille de nez ne sera pas retenue comme caractéristique de ces deux classes.

es variables
6 b

Ainsi, la F-mesure permet par conséquent de trouver les variables les plus caractéristiques de chaque classe comme le Khi 2 qui est utilisé dans IRaMuTeQ pour représenter les termes les plus représentatifs de chaque monde lexical. Il est légitime à ce stade de se demander si la F-mesure est une mesure appropriée sur un tableau constitué que de 0 et de 1 et si elle apporte une réelle plus-value par rapport au Khi 2. Ces questions seront abordées ultérieurement dans le premier temps de l'expérimentation pratique.

2.2. L'établissement d'une proximité entre deux clusters

Pour illustrer la méthode permettant de déterminer si un cluster d'une période « n+1 » a un contenu proche d'un cluster d'une période « n », nous examinons ci-dessous un exemple volontairement réduit. Le premier cluster à gauche est appelé l'ensemble source (S_s) d'une période « n ». Il est composé de variables représentatives sélectionnées par la méthode précédemment exposée. Entre parenthèse est indiquée pour chaque variable (f) sa F-mesure (FF_s). A droite, l'ensemble cible (S_t) avec ses variables représentatives sélectionnées par la même méthode et ses F-mesures (FF_t). Les termes communs aux deux clusters sont en gras.

élevage (0,18), bétail (0,16), culture (0,16), **animal (0,14)**

Un exemple de cluster source : Ss

élevage (0,15), animal (0,14), machine (0,13), pré (0,11)

Un exemple de cluster cible : St

Un premier indicateur est calculé pour le cluster cible :

$$I(t) = \sum_{f \in Ss \cap St} FFt\,(f) \Big/ \sum_{f \in St} FFt\,(f)$$

Un calcul similaire est réalisé pour le cluster source en calculant I(s). A partir de l'exemple :

I(t) = (0,15 + 0,14) / (0,15 + 0, 14 + 0,13 + 0,11) = 0,29 / 0,53 = 0,56

I(s) = (0,18 +0,14) / (0,18 +0,16 +0,16 +0,14)= 0,32 /0,64 = 0,5

Cette présentation nous amène à un écart par rapport à la formule présentée[3] dans l'article de Diachronic'Explorer (Dugué et al, 2016 a). Cet écart peut sembler à première vue mineur puisqu'il s'agit juste d'un changement d'une notation en terme de probabilité (P(t|s)) pour une notation en terme d'indicateur (I(t)). Pourtant, le changement est bien plus fondamental qu'il n'y paraît au premier abord puisqu'il conduit à remettre en question tout l'arrière fond bayésien sur lequel repose le modèle. Ce qui a motivé cette reformulation est la prise de conscience que le calcul présenté dans Diachronic'Explorer ne correspond pas à la formule bayésienne classique : P(A|B) = P(A ∩ B) / P(B). Pour être plus précis, le problème vient dans la formule présentée du dénominateur qui ne fait pas référence au cluster source contrairement à ce qui pourrait être attendu.

[3] $P(t \mid s) = \sum_{f \in Ss \cap St} FFt\,(f) / \sum_{f \in St} FFt\,(f)$

Il résulte de cette reformulation théorique que la compréhension des calculs est beaucoup plus aisée. I(s) désigne dans l'ensemble source la proportion du poids des termes en commun par rapport au poids de tous les termes de cet ensemble source. I(t) désigne dans l'ensemble cible la proportion du poids des termes en commun par rapport au poids de tous les termes de cet ensemble cible. La force du lien est une moyenne de ces deux indicateurs basiques. Il est possible d'indiquer en plus s'il y a un déséquilibre de valeurs entre ces deux indicateurs pour détecter des cas particuliers où il est nécessaire de regarder plus attentivement à quoi correspond le résultat de la moyenne obtenue.

Pour sélectionner les liens les plus significatifs, Diachronic'Explorer propose une méthode assez proche de la sélection précédente des variables puisqu'elle consiste à comparer les valeurs obtenues de I(s) et de I(t) a des moyennes et à des moyenne de moyenne. En effet, un cluster source d'une période « n » peut avoir des termes en commun avec plusieurs clusters cible d'une période « n+1 ». Il est ainsi possible de calculer la moyenne des I(t) générées, appelée M(t). De plus, une période « n » étant constituée de plusieurs clusters sources, il est possible de calculer une moyenne des moyennes précédemment calculées, appelée MM(t) ainsi qu'un écart type appelé б(t). De manière symétrique, un cluster cible pouvant avoir des termes en commun avec plusieurs clusters sources, les mêmes calculs sont réalisables aboutissant aux valeurs de M(s), MM(s) et б(s). Pour qu'un lien soit gardé comme significatif, il faut que les deux conditions suivantes soit respectées :

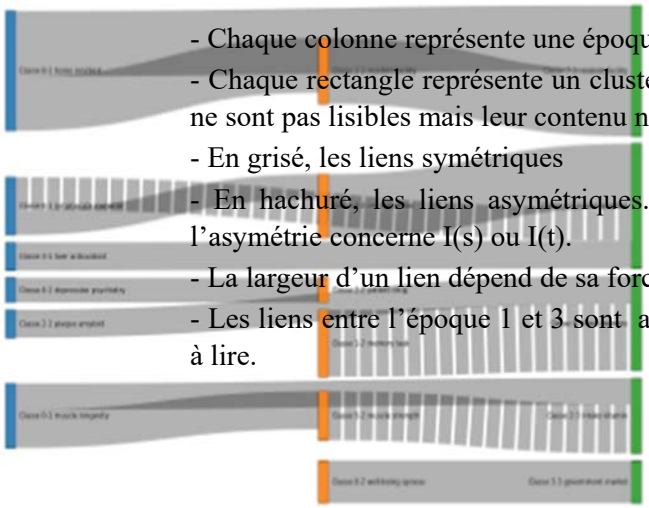
1) I(s) > M(s) et I(s) > MM(s) + б(s)

2) I(t) > M(t) et I(t) > MM(t) + б(t)

Si une seule condition est respectée le lien peut être soit rejeté, soit gardé en étant qualifié d'asymétrique suivant le choix de l'utilisateur. Cette partie peut être reprise a priori sans aucune difficulté pour être appliquée aux mondes lexicaux.

2.3. Des changements dans la représentation finale

Nous reproduisons ci contre la visualisation graphique présentée par Diachronic'Explorer pour mieux faire comprendre la raison des changements effectués sur ce point.



- Chaque colonne représente une époque
- Chaque rectangle représente un cluster. Du fait, de la petite taille de l'image, les libellés des clusters ne sont pas lisibles mais leur contenu n'a pas d'importance dans cette présentation.
- En grisé, les liens symétriques
- En hachuré, les liens asymétriques. Avec cette représentation, il n'est pas possible de savoir si l'asymétrie concerne I(s) ou I(t).
- La largeur d'un lien dépend de sa force
- Les liens entre l'époque 1 et 3 sont aussi représentés sur ce graphique ce qui rend le assez complexe à lire.

Fig 4 : Un exemple de représentation finale obtenue avec Diachronic'Explorer tiré de Dugué et al, 2016 a

En raison de notre objectif initial et de la complexité que cela engendre sur des exemples qui ne se limitent pas à trois époques, il a été décidé de ne pas représenter les liens entre époques non successives. Ensuite, plutôt que de faire dépendre la taille des liens de leur force, nous avons préféré visualiser ces différences d'intensité par des niveaux de gris. Cette modification permet de faire dépendre la taille du cluster (ou du monde lexical) non pas de ses intensités de relations mais de son poids. Par exemple, dans le cas d'un monde lexical, cette taille peut être directement reliée au nombre de segments de textes qu'il contient par rapport aux autres mondes lexicaux de la même période. Enfin, nous n'avons pas repris l'idée des hachures pour les liens asymétriques mais nous permettons à l'utilisateur d'afficher en plus cette information en bordure des liens avec deux couleurs suivant si l'asymétrie concerne I(s) ou I(t). Ce travail a été effectué en reprenant une représentation de la plateforme CORTEXT plus proche de ces directions de recherche. Suite à ces réflexions, nous abordons maintenant leurs mises en pratique avec la présentation d'une expérimentation.

3. Une expérimentation pratique

3.1. La préparation des données

Les articles de deux revues de géographie française ont été récupérés grâce aux portails Persée et Cairn : *Les Annales de Géographie* sur la période 1892-2014 et l'*Espace Géographique* sur la période 1972-2014. Seuls les articles en français et de plus de trois pages ont été gardés. Les bibliographies et les notes de bas de page ont été également retirées car elles relèvent d'un genre textuel spécifique. Quelques erreurs liées à la Reconnaissance Optique de Caractère ont été corrigées. Seuls les mots pleins (nom, adjectif, adverbe, verbe) ont été conservés et les formes ont été lemmatisées en utilisant le logiciel IRaMuTeQ.

3.2. Le choix entre la F-mesure ou le Khi 2

Pour répondre tout d'abord à la question concernant la pertinence d'utiliser la F-mesure dans notre cas, nous utilisons la méthode Reinert sur un corpus réduit à la revue *Les Annales de Géographie* sur la période 1892-1911. Cette réduction s'explique par le fait qu'à ce stade, il ne s'agit pas de chercher des continuités entre différentes époques mais seulement de comparer les termes sélectionnés en utilisant la F-mesure par rapport à ceux mis en avant par la méthode classique d'IRaMuTeQ en utilisant le Khi 2. Afin de rendre ces résultats comparables, nous avons divisé les valeurs de Khi 2 obtenues par 100 et multipliées les valeurs de F-mesure obtenues par 1000. Soulignons que les calculs dans Diachronic'Explorer reposant entièrement sur des rapports, le fait de multiplier ou de diviser les résultats pour faciliter leur présentation et leur comparabilité n'a pas d'impact sur la suite.

Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6

géographie (34)	plateau (13)	rivière (35)	population (31)	fer (25)	calcaire (32)
étude (24)	nord (13)	eau (22)	culture (24)	voie (20)	érosion (29)
géographique (16)	sud (13)	cours (21)	habitant (17)	port (19)	roche (25)
science (15)	chaîne (11)	rive (21)	agricole (9)	chemin (18)	couche (23)
observation (12)	altitude (11)	fleuve (20)	vie (9)	commerce (16)	dépôt (19)
question (12)	côte (11)	vallée (17)	nomade (9)	ville (12)	pli (18)
...	...	...	...	...	.

Fig 5 : Les 6 premiers lemmes représentatifs des mondes lexicaux obtenus en utilisant la méthode du Khi 2. Les valeurs de Khi 2 entre parenthèse ont été divisées par 100

Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
étude (14)	grand (21)	eau (33)	pays (18)	grand (18)	région (17)
géographie (14)	plateau (20)	vallée (29)	grand (17)	voie (13)	vallée (16)
voir (13)	nord (18)	rivière (22)	population (17)	fer (13)	calcaire (15)
donner (11)	sud (18)	cours (20)	culture (14)	ville (13)	eau (14)
premier (9)	haut (17)	grand (17)	terre (13)	pays (12)	érosion (13)
général (9)	région (16)	fleuve (15)	région (12)	chemin (12)	partie (12)
...	...	...	...	...	...

Fig 6 : Les 6 premiers lemmes représentatifs des même mondes lexicaux que la Fig 5 en utilisant la méthode de la F-mesure. Les valeurs de F-mesure entre parenthèse ont été multipliées par 1000

Si on compare les deux tableaux précédents, les premiers termes caractérisant les mondes lexicaux sont finalement assez proches que l'on utilise le Khi 2 ou la F-mesure. Toutefois, les F-mesures font remonter de manière importante quelques lemmes plus généraux comme « grand » ou « région ». Ce phénomène peut renforcer par la suite la créations de proximités (entre mondes lexicaux d'une époque « n » et ceux d'une époque « n+1 ») qui ne sont pas forcément pertinentes car ces termes généraux sont très partagés et ne sont pas des marqueurs thématiques convaincants.

Un avantage non négligeable de la F-mesure est de sélectionner moins de mots représentatifs d'un monde lexical. En effet, les classes, que ce soit dans le cas du Khi 2 ou de la F-mesure sont beaucoup moins cohérentes d'un point de vue thématique après 200 lemmes environ dans cet exemple. Le problème reste que même en prenant la F-mesure, les classes continuent d'avoir plus de 500 lemmes dans ce cas. Au final, il nous semble préférable de donner la possibilité à l'utilisateur de fixer lui même un seuil plutôt que de faire confiance à la méthode de la F-mesure pour déterminer une limite de cohérence thématique.

Nous avons choisi par la suite de travailler plutôt en privilégiant le Khi 2 car les avantages de la F-mesure ne sont pas évident à justifier à partir de cet exemple. Comme l'objectif était de faire une proposition à partir des mondes lexicaux très souvent produits à partir d'IRaMuTeQ dont la présentation est structurée par le Khi 2, la F-mesure nous a semblé à partir de cette expérimentation introduire une complexification qui n'était finalement pas nécessaire.

3.3. Le résultat obtenu par la méthodologie de Diachronic Explorer

Le corpus a été découpé en six époques : 1891-1911, 1912-1931, 1932-1951, 1952-1971, 1972-1992, 1993-2014. Étant donné ce qui a été précédemment vu, nous avons intégré la possibilité de définir un seuil de nombre de lemmes marquant la limite de cohérence thématique des variables dans la définition de la classe. Il aurait été logique de le fixer à 200 dans notre exemple. Toutefois, étant donné l'objectif de se comparer aux résultats obtenus avec la méthodologie Labbé + Ward où il n'est pas possible de fixer pour l'instant un seuil, cette option

n'a pas été utilisée dans les résultats ici présentés. Nous avons gardé donc tous les lemmes mis en avant par IRaMuTeQ pour définir le profil de chacune des classes.

Ce choix a aggravé un effet inattendu. En effet, dans les conditions pour déterminer si un lien entre deux mondes lexicaux est sélectionné comme important, il y a le calcul de moyennes et de moyenne de moyennes. Or, dans l'implémentation initiale de Diachronic'Explorer, si un cluster source n'a aucun terme en commun avec un cluster cible, il ne participe pas au calcul de ces moyennes. En prenant en compte autant de lemmes, il y a toujours au moins un terme en commun, ce qui diminue de beaucoup les moyennes et les moyennes de moyenne. En conséquence, beaucoup de liens, même faibles, se retrouvent au dessus de ces moyennes et sont donc sélectionnés comme significatif par la méthode. Le graphique devient alors difficilement lisible. Nous tenons à souligner qu'en prenant 200 lemmes, le problème reste présent car il suffit d'un lemme en commun pour que ce phénomène se produise. En pratique, cela donne beaucoup de poids à des phénomènes très réduits puisqu'un lemme en commun peut avoir un impact important sur le calcul de ces moyennes.

Nous avons donc opté pour la solution de mettre un seuil manuel permettant d'éliminer ces liens les moins importants. De plus, ce seuil est réglable dynamiquement en laissant la possibilité à l'utilisateur de voir quels effets directs ses changements de seuil provoquent sur le graphique. Dans notre cas, ce seuil a été fixé à 0,6 dans le graphique présenté à la figure 8 pour conserver un maximum d'informations tout en gardant un ensemble lisible. Cela signifie que tous les liens entre deux mondes lexicaux qui avaient une valeur inférieure à 0,6 ont été enlevés. Cette proposition modifie la philosophie de fond de Diachronic'Explorer qui essaye de proposer une méthode de sélection automatique. Nous assumons totalement ce changement au vue de l'expérimentation menée.

Sur la figure 8 qui présente le résultat obtenu, nous avons laissé volontairement un élément dynamique (l'encadré noir). En effet, une partie en javascript permet à l'utilisateur de passer sur un rectangle représentant un monde lexical ou sur un lien pour bénéficier de l'affichage d'informations sur leurs compositions. Ici, c'est le lien en violet qui est détaillé en affichant les mondes lexicaux qu'il relie, sa force, son asymétrie (ce qui n'est pas le cas ici), les 5 premiers lemmes en communs, les 5 premiers lemmes uniquement dans la source et enfin les 5 premiers lemmes qui sont uniquement dans la cible. Cette fonctionnalité a été ici réduite à 5 lemmes pour ne pas surcharger la figure mais elle permet au lecteur de comprendre tout l'intérêt de cette représentation pour suivre l'évolution de la composition d'un monde lexical.

3.4. Le résultat obtenu par la méthode Labbé + Ward

Le résultat produit à l'aide d'IRaMuTeQ en utilisant la méthode de Labbé + Ward est présenté à la figure 7. L'arbre obtenu est à l'origine sans couleur. Ce qui a motivé sa coloration est un travail mené pour obtenir un résultat comparable au précédent.

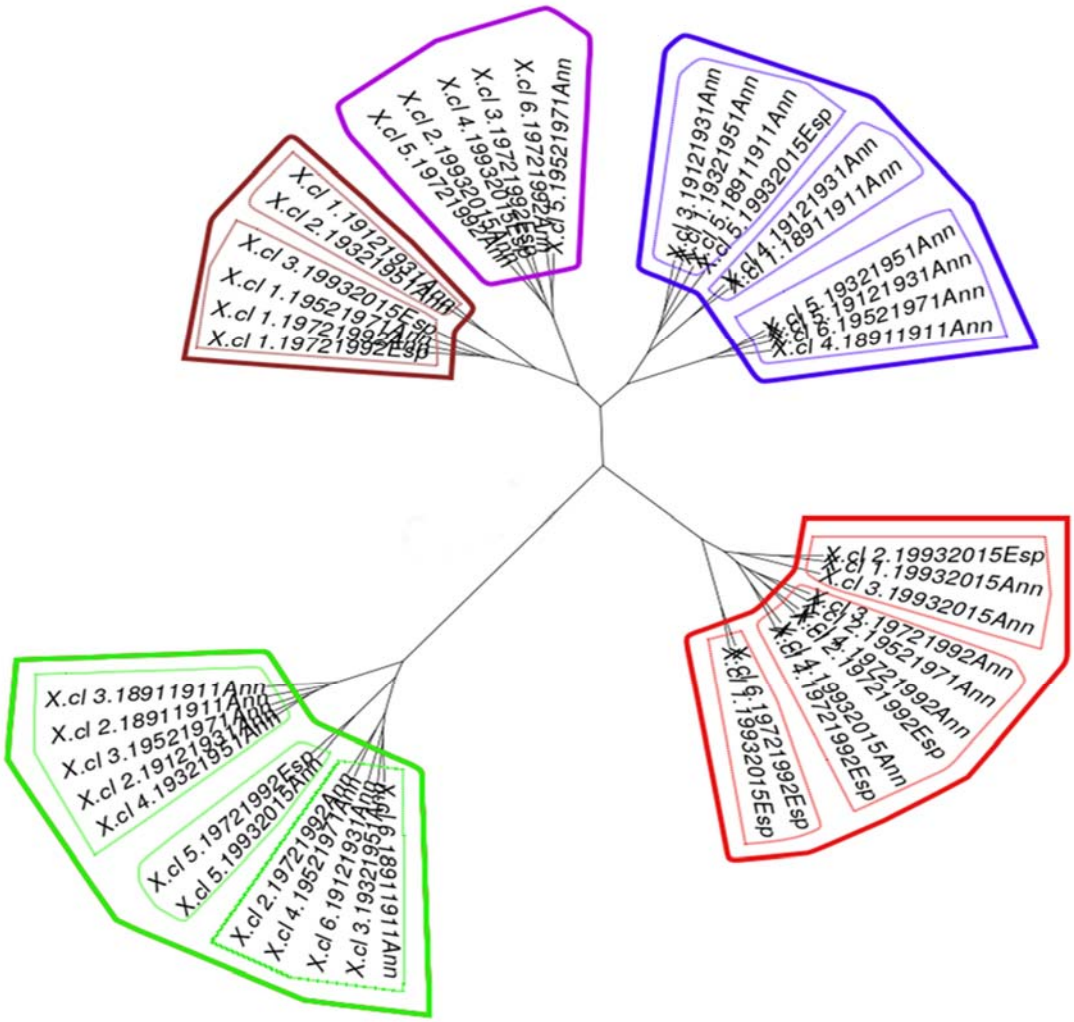

X.cl 6.19721992Ann
X.cl 3.19721992Ann
X.cl 4.19932015Esp
X.cl 2.19932015Ann
X.cl 5.19521971Ann
X.cl 1.19121931Ann
X.cl 2.19321951Ann
X.cl 3.19932015Esp
X.cl 1.19521971Ann
X.cl 1.19721992Esp
X.cl 3.19121931Ann
X.cl 1.19321951Ann
X.cl 5.18911911Ann
X.cl 5.19932015Esp
X.cl 4.19121931Ann
X.cl 1.18911911Ann
X.cl 5.19321951Ann
X.cl 6.19121931Ann
X.cl 4.18911911Ann
X.cl 2.19932015Esp
X.cl 1.19932015Ann
X.cl 3.19932015Ann
X.cl 3.19721992Ann
X.cl 2.19521971Ann
X.cl 4.19721992Ann
X.cl 4.19932015Ann
X.cl 4.19721992Esp
X.cl 1.19932015Esp
X.cl 3.18911911Ann
X.cl 2.18911911Ann
X.cl 3.19521971Ann
X.cl 2.19121931Ann
X.cl 4.19321951Ann
X.cl 5.19721992Esp
X.cl 5.19932015Ann
X.cl 2.19721992Ann
X.cl 4.19521971Ann
X.cl 6.19121931Ann
X.cl 3.19321951Ann
X.cl 6.18911911Ann

Fig 7 : Rapprochement entre les mondes lexicaux effectué par la méthode Ward + Labbé. Sortie brute par le logiciel IRaMuTeQ. Colorisation réalisée par les auteurs.

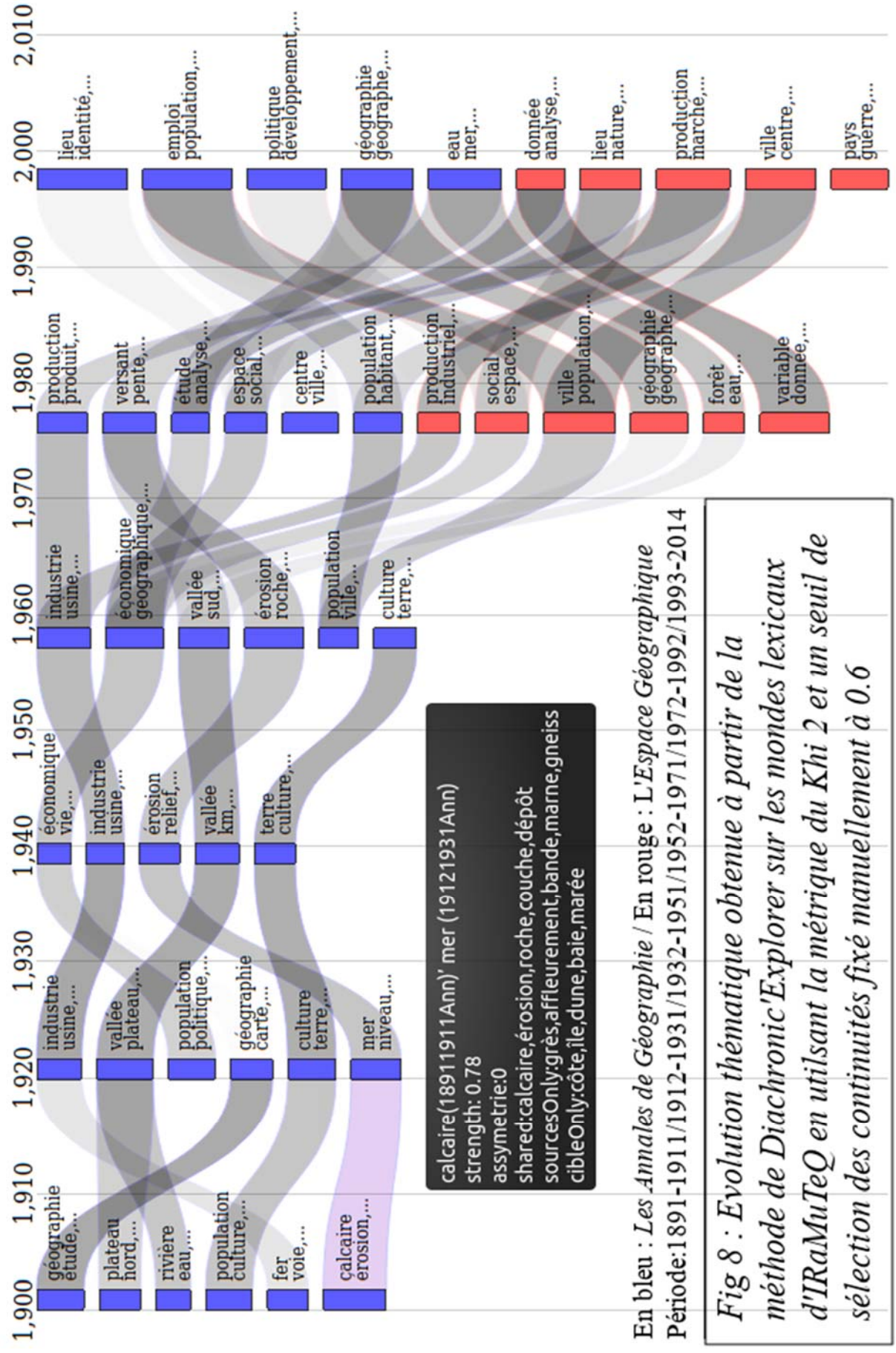


Fig 8 : Evolution thématique obtenue à partir de la méthode de Diachronic'Explorer sur les mondes lexicaux d'IRaMuTeQ en utilsant la métrique du Khi 2 et un seuil de sélection des continuités fixé manuellement à 0.6

Le découpage réalisé n'est pas si évident qu'il n'y paraît. Par exemple, faut il fusionner les ensembles violet et marron pour former un ensemble commun ? Il n'y pas sur ce sujet de réponse simple et les découpages proposés peuvent tout à fait être discutés.

La réflexion nous a amené à construire des sous-ensembles (en pointillé) dans l'établissement des relations entre mondes lexicaux. Par exemple, dans le cluster vert, il semble opportun de a relier Cl2.1891-1911Ann[4] et Cl3.1891-1911Ann avec Cl2.1912-1931Ann mais il ne semble pas judicieux de les relier avec Cl6.1912-1931Ann qui a plutôt une proximité avec Cl6.1891-1911Ann du fait de leur appartenance à des sous-ensembles distincts. Quand il n'existait pas de proximité dans le sous-ensemble mais qu'il était possible d'en trouver dans l'ensemble correspondant, il nous a semblé intéressant de le signifier en marquant un lien en pointillé comme entre Cl2.1932-1951Ann et Cl1.1952-1971Ann. De plus, sur les liens entre l'avant dernière et la dernière période, il a été décidé de ne pas ajouter de proximité provenant de l'autre revue quand il existait déjà une proximité dans la revue même. Ainsi, Cl5.1972-1992Ann n'a par exemple pas été relié à Cl4.1993-2014Esp car il existait déjà une proximité avec Cl3.1972-1992Ann. Cette règle crée une différence entre les deux graphiques sur cette période qui se justifie seulement par le fait que sans elle, la représentation des continuités entre cette avant-dernière et dernière période devenait très difficilement lisible.

En suivant ces règles, le graphique de la figure 10 a été construit à la main. Dans l'ensemble, une forte ressemblance est notable avec les résultats précédemment obtenus (figure 8). Un avantage de la méthode Labbé + Ward est de créer des méta-ensembles qui ont du sens. Par exemple, l'ensemble vert qui se retrouve sur la figure 10 matérialisé par les flèches vertes correspond à la géographie physique. Cela étant dit, il est possible d'appliquer la méthode de Ward sur des distances issues de la méthode de Diachronic'Explorer pour trouver des résultats proches. Notons qu'il existe également quelques différences entre les deux représentations réalisées (figures 8 et 10). Par exemple, la méthode de Diachronic Explorer rapproche Cl5.1891-1911Ann de Cl1.1912-1931Ann alors que la méthode Labbé + Ward rapproche ce monde lexical plutôt de Cl3.1912-1931Ann. Le détail des 25 premiers lemmes de Cl5.1891-1911Ann en détaillant ceux en commun avec les 200 premiers lemmes de Cl1.1912-1931Ann et de Cl3.1912-1931Ann permet de détailler le contenu de cette différence par rapport aux termes les plus représentatifs de ces mondes lexicaux.

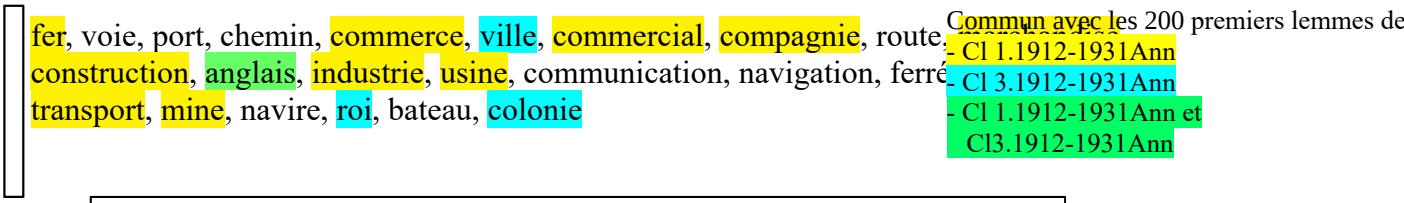
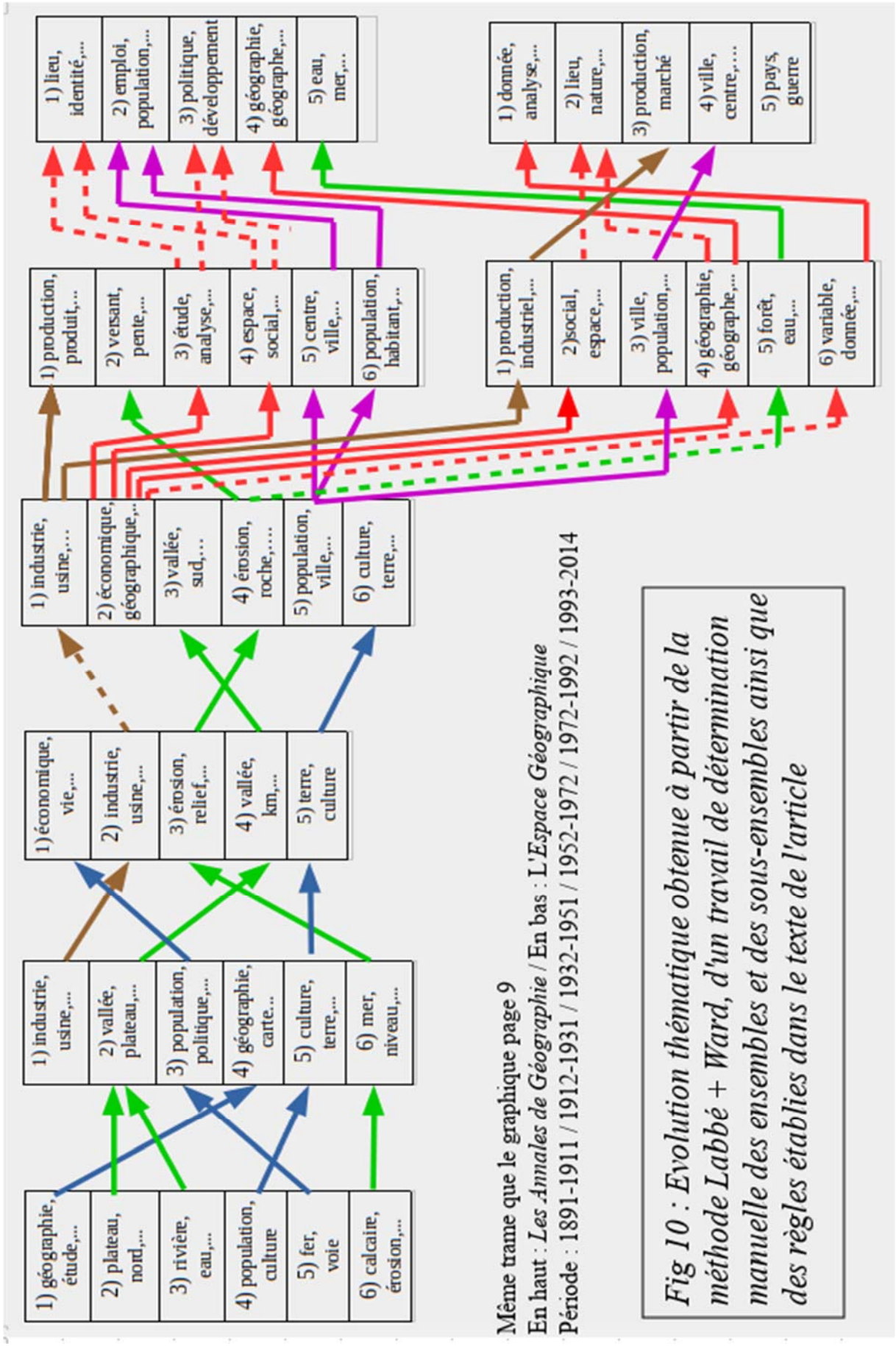


Fig 9 : Appartenance privilégiée des 25 premiers lemmes de Cl5.1891-1911Ann

Cette figure donne un avantage au rapprochement effectué par la méthode de Diachronic'Explorer. Cette constatation se retrouve sur d'autres exemples car le calcul réalisé par cette méhode est plus proche de la présentation des mondes lexicaux dans IRaMuTeQ basée

[4] Cette notation correspond au monde lexical 2 du sous corpus des Annales de Géographie sur la période 1891-1911. Le lecteur peut se référer dans ce cas à la figure 5 et ainsi comprendre le phénomène d'aller-retour qu'il est nécessaire d'effectuer pour comprendre les évolutions thématiques à partir de cet arbre.

sur le Khi 2 alors que les calculs pour la distance de Labbé repassent par une donnée liée mais moins directement structurante : la fréquence des termes (Labbé, 2003).



Même trame que le graphique page 9
En haut : *Les Annales de Géographie* / En bas : *L'Espace Géographique*
Période : 1891-1911 / 1912-1931 / 1932-1951 / 1952-1972 / 1972-1992 / 1993-2014

Fig 10 : Evolution thématique obtenue à partir de la méthode Labbé + Ward, d'un travail de détermination manuelle des ensembles et des sous-ensembles ainsi que des règles établies dans le texte de l'article

4. Conclusion

Cette expérimentation montre la pertinence d'avoir adapté Diachronic'Explorer pour étudier l'évolution des mondes lexicaux. Le premier avantage est de ne pas demander un travail long et minutieux de reconstruction diachronique. La visualisation est plus intuitive et permet de mieux remplir l'objectif principal de connaître l'évolution du contenu des mondes lexicaux. En amont, plusieurs modifications ont été réalisées : sur le plan de la théorie sous-jacente, sur la métrique utilisée, sur la sélection des continuités et sur la visualisation. Le travail à venir consiste donc à rendre disponible cet outil. Il est malheureusement difficilement intégrable directement à IRaMuTeQ car la partie en javascript nécessite un navigateur web. Par contre, le projet s'oriente vers la forme d'une interface Web. Il faut tout de même noter que pour l'instant les temps de traitement de la méthode Reinert rendent l'algorithme plutôt long. Le travail de Pierre Ratinaud sur une exécution parallélisée de cette méthode (Ratinaud, 2018) devrait régler ce problème dès qu'il sera disponible publiquement.

Au niveau du calcul de la proximité entre deux thématiques, les résultats basés sur le principe de Diachronic'Explorer en utilisant la métrique du Khi 2 permettent de prendre comme base de calcul la structuration mise en avant par IRaMuTeQ. Pour cette raison, cette proposition constitue dans ce cas à notre avis une alternative intéressante à l'utilisation des distances de Labbé. Au niveau de la méthode de Ward, l'expérimentation menée nous a montré qu'un avantage de celle-ci réside dans l'identification de méta-catégories rassemblant plusieurs mondes lexicaux mais elle est plus problématique pour établir des continuités fines. Un autre intérêt de la méthode de Ward à ne pas sous-estimer est qu'elle permet de saisir certaines proximités entre des mondes lexicaux d'époques non successives, ce qui n'est pas pris en compte dans la démarche ici proposée mais peut dans certains cas produire des informations pas inintéressantes à prendre en compte. Cette perte si elle est conscientisée et acceptée par l'utilisateur nous semble amplement compensée par le gain obtenu en visualisation. Les diagrammes alluviaux permettent à la fois d'avoir une vision de l'ensemble des dynamiques et de rentrer dans le détail de chaque évolution de manière plus aisée. Par rapport aux limites de cette représentation qui viennent d'être explicitées, la méthode de Ward peut toujours être utilisée en complément pour explorer ces points spécifiques.

References

Brunet, E. (2014). Où l'on mesure la distance entre les distances. *Texto !* Textes et Cultures, Institut Ferdinand de Saussure, Dits et inédits, publication électronique

Dugué N., Lamirel J.-C. and Cuxac P. (2016 a). Diachronic'Explorer: Keep track of your clusters. In *International Conference on Research Challenges in Information Science (RCIS),* pp. 1-2

Dugué, N., Lamirel, J-C. et Cuxac, P. (2016 b). Visualisation pour la détection d'évolutions dans des corpus de publications scientifiques. *Les Cahiers du numérique*, *12*(4), pp. 157-184

Labbé C. et Labbé D. (2003), La distance intertextuelle *Corpus.*, (2)

Ratinaud, P. et Marchand, P. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). *Mots. Les langages du politique*, (2), pp.57-77.

Ratinaud, P. (2018). Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ. In *JADT' 2018,* pp 616-625

Reinert M. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*. VIII, (2), 187-198.

Rosvall M. et Bergstrom C. T. (2010). Mapping change in large networks. PLoS one 5.1