Analyse du corpus de référence des thèses françaises concernant la recherche sur la formation des adultes soutenues depuis 2010 : Thématiques, rattachements et spécificités ; Rubrique " Vie de la recherche " , commune aux Revues Savoirs et TransFormations, Octobre 2020

Olivier Las Vergnas, Patrick Bury

Rubrique « Vie de la recherche » , commune aux Revues Savoirs et TransFormations, Octobre 2020

N°7 - Analyse du corpus de référence des thèses françaises concernant la recherche sur la formation des adultes soutenues depuis 2010 : Thématiques, rattachements et spécificités

Olivier LAS-VERGNAS[1,2] et Patrick BURY[3]

1 : équipe Trigone - CIREL, Université de Lille
2 : équipe AP'FORD - CREF, Université Paris - Nanterre
3 : Société CleverMind – Datascientist

Résumé :

Ce travail analyse le corpus de référence des thèses françaises concernant la recherche sur la formation des adultes soutenues depuis 2010 publié dans le précédent article (VdR6) de cette rubrique. Celui-ci est composé de deux groupes de thèses, le « noyau dur » et le « second cercle ». Le premier regroupe les 175 thèses explicitement centrées sur cette thématique et repérables par une simple requête bibliographique ; le second y ajoute 444 autres contenant des sous-parties éclairantes pour ce champ identifiées grâce à la composition de leurs jurys et une analyse automatique des termes employés de leurs résumés. L'analyse de ce double corpus confirme que les thèses du noyau sont aux trois quarts soutenues en sciences de l'éducation tandis que celles du second cercle ne le sont plus qu'aux deux tiers ; sur l'ensemble des deux listes on constate aussi que l'âge médian des doctorants est largement plus élevé que dans les autres disciplines (10 ans d'écart avec la sociologie). Enfin, les analyses lexicales montrent que ce corpus peut se découper en quatre classes de thématiques (analyse de l'activité des enseignants, politiques de formation, apprentissages linguistiques des adultes, approches biographiques et identitaires) qui peuvent s'affiner en six (avec apparition d'une sous classe liée à l'apprenance et aux soignants et d'une autre liée à l'autoconfrontation, l'analyse de l'activité).

Mots-clefs : formation des adultes, recherche doctorale, bibliométrie, âge des doctorants, composition des jurys de thèses

Abstract:

This work analyzes the corpus of French PhD theses concerning research on adult education defended since 2010 published in the previous article (VdR6) of this section. This corpus is composed of two groups of PhD theses, the "hard core" and the "second circle". The first group includes 175 theses explicitly centered on this theme and has be found by a simple bibliographic query; the second adds 444 other theses containing illuminating subparts for this field identified through the composition of their juries and the automatic analysis of the terms used in their abstracts. The analysis of this double corpus confirms that three quarters of the theses of the hard core are defended in educational sciences, while those of the second circle are only two-thirds; on both lists as a whole, we also note that the median age of doctoral students is much higher than in other disciplines (10 years difference with sociology). Finally, lexical analyses show that this corpus can be divided into four classes of themes (analysis of teacher activity, training policies, adult language learning, biographical and identity approaches) or into six (with the appearance of a sub-class related to "learnance" and caregivers and of another one related to self-confrontation and activity analysis).

Keywords: adult education, doctoral research, bibliometrics, age of doctoral students, composition of thesis juries.

1. Rappel de la méthode de constitution d'un double corpus de référence

Après plusieurs tentatives décrites dans les premiers articles de cette rubrique « vie de la recherche » mettant en évidence des problèmes de définition et de sélection sous-jacents à toute constitution de corpus bibliométriques non purement disciplinaires, les auteurs ont finalement établi dans leur dernière contribution (VdR6) deux listes de thèses soutenues en France depuis 2010, constituant un double corpus de référence pour l'étude de la recherche en formation des adultes (RFdA)[1].

1.1. Un noyau dur de 175 thèses issu de simples requêtes bibliographiques

La première liste a été définie comme correspondant au « noyau dur » de cette RFdA puisqu'elle énumère les 175 thèses soutenues en France depuis 2010 faisant explicitement référence à l' « éducation des adultes », la « formation des adultes » ou à quelques autres termes spécifiques. Cette première liste a été obtenue par des requêtes bibliométriques classiques sur ces termes suivis d'une épuration par lecture flottante de l'un des auteurs ; concrètement, cette procédure de nettoyage consistait à valider le fait que le résumé traitait bien de RFdA et si oui à identifier empiriquement la ou les thématiques abordées qui justifiaient -aux yeux du lecteur- cette appartenance. Cette phase a conduit à l'élimination de 16 thèses qui se sont révélées être des faux positifs[2] ; en parallèle avec la validation des 175 autres, elle a aussi conduit à l'établissement d'une catégorisation des thématiques ayant permis la justification de leur appartenance à la RFdA, présentées en encadré 1.

Encadré 1 : validation du noyau dur du corpus RFdA et catégorisation par lecture flottante

Le travail vérificatoire de lecture flottante des résumés des thèses du noyau dur a permis de repérer 7 catégories thématiques constituant une représentation non pondérée des sous-champs de la RFdA :

- Andragogie : IOX (62 thèses concernées)
- Socio économie de la Formation : EDM (11)
- Monde scolaire : SCO (33)
- Monde universitaire, jeunes adultes en formation : UJI (33)
- Didactiques disciplinaires ou pro, TICE : DFP (88)
- Questions anthropologiques ou politiques : POA (41)
- Linguistique et langages : LAN (29)

Il est d'ailleurs à noter que beaucoup de ces 175 thèses ont été jugées comme liées simultanément à plusieurs de ces thématiques.

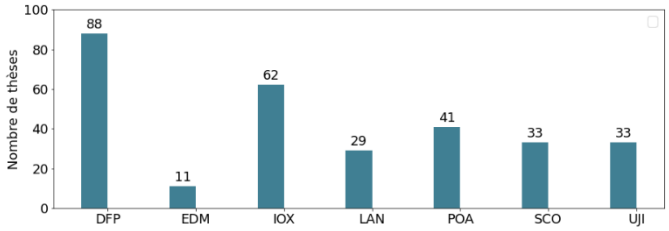

Figure 1 : Nombre de thèses pour chacune des catégories obtenues en lecture flottante pour le noyau

[1] La liste publiée dans l'article précédent (VdR6) et en ligne à https://gitlab.com/pbury/VdR6/-/tree/1.0/tables_reference est arrêtée au 1er mai 2020, mais les auteurs ont également publié à la même adresse les requêtes et algorithmes permettant de réactualiser ces listes régulièrement sur une base reproductible.

[2] Dus par exemple à la polysémie de termes utilisés dans les résumés de thèses comme « VAE » qui peut désigner une « vélo à assistance électrique » ou comme « formation » qui peut concerner les végétaux adultes

De plus, l'analyse des jurys de ces 175 thèses a permis de constituer une table des directeurs et rapporteurs de jury « impliqués dans la RFdA » destinée à servir de référence pour la constitution d'un second cercle de thèses.

1.2. Un second cercle de 444 thèses en partie intéressantes pour la RFdA

La deuxième liste de thèses, désignée sous le nom de « second cercle », a été constituée avec cette fois la volonté de recenser non plus les thèses consacrées dans leur intégralité à la RFdA, mais d'autres thèses susceptibles de contenir certains développements intéressants pour le champ.

Pour obtenir cette seconde liste, la méthode se construit en deux étapes. Dans un premier temps, un ensemble de 833 autres thèses a été présélectionné parmi la totalité des thèses de SHS restantes dans le Sudoc en se fondant sur la composition des jurys. Ces thèses étaient celles dont le jury associait une forte proportion de spécialistes de la RFdA : soit directeurs ou rapporteurs d'une ou plusieurs des thèses du noyau dur (repérés grâce à la table de jurés établie plus haut), soit membres des comités éditoriaux de revues scientifiques dédiées à la RFdA. Dans un second temps, un filtrage complémentaire a été effectué parmi ces 833 thèses pour identifier par *Machine Learning* (régression logistique) le sous ensemble de thèses lexicalement proches à plus de 90% d'au moins une des 7 catégories thématiques trouvées en lecture flottante dans les 174 résumés du noyau dur (cf. encadré 1). En définitive, l'opération a identifié 444 thèses qui respectaient ce critère.

Comme cela a déjà été indiqué dans l'article VdR6, ces deux listes de thèses (noyau dur et second cercle) sont accessibles en ligne (sous licence libre **CC-BY-SA)** à l'adresse https://gitlab.com/pbury/VdR6/-/tree/1.0/tables_reference. Dans cette suite de nos travaux, ces deux groupes seront considérés comme base de référence et appelés **« corpus RFdA ».**

2. Répartition disciplinaire, bibliothécaire et géographique du corpus RFdA

Une fois ce double corpus obtenu, il parait tout d'abord pertinent d'en décrire les constituants en fonction des principales caractéristiques universitaires. Sont donc proposées ici des diagrammes décrivant la répartition de ces 619 thèses (1) selon leurs rattachements par discipline au sens du conseil national des universités (CNU), (2) par la façon dont leurs contenus sont catégorisés au sein de la classification Dewey dans les bibliothèques universitaires, (3) par année de soutenance et enfin (4) par établissement universitaire de soutenance.

2.1. Répartition par disciplines (par numéro de section CNU)

Les figures 2 et 3 représentent ainsi les répartitions en pourcentages des 619 thèses dans les différentes sections CNU, respectivement pour celles du noyau dur (figure 2) puis celles du second cercle (figure 3). On retrouve ainsi le poids déjà signalé (Laot, 2006 ; Las Vergnas, 2016, 2017) de la section 70 (Sciences de l'éducation) qui fournit les trois quarts du noyau dur, mais les deux tiers seulement du second cercle (pour lequel on observe un apport de plus d'un quart de thèses soutenues en section 7, c'est-à-dire en linguistique et sciences du langage).

2.2. Répartition par type de contenus d'un point de vue bibliothécaire (par classification Dewey)

De la même façon, les figures 4 et 5 donnent les répartitions en pourcentages des 619 thèses selon les différentes cotations Dewey dans lesquelles elles ont été cataloguées dans les bibliothèques universitaires. Afin de rester lisibles, les figures et tableaux sont limités au premier chiffre des cotes Dewey qui correspond aux grandes catégories.

Section	Nombre de thèses
Sciences de l'éducation et de la formation	133
Sciences du langage	14
Sciences de l'information et de la communication	5
Psychologie et ergonomie	4
Sociologie, démographie	4
Langue et littérature françaises	3
Sciences de gestion et du management	3
Sciences économiques	2
Science politique	2
Sciences et techniques des activités physiques et sportives	1
Génie informatique, automatique et traitement du signal	1
Informatique	1
Biologie des populations et écologie	1
Philosophie	1

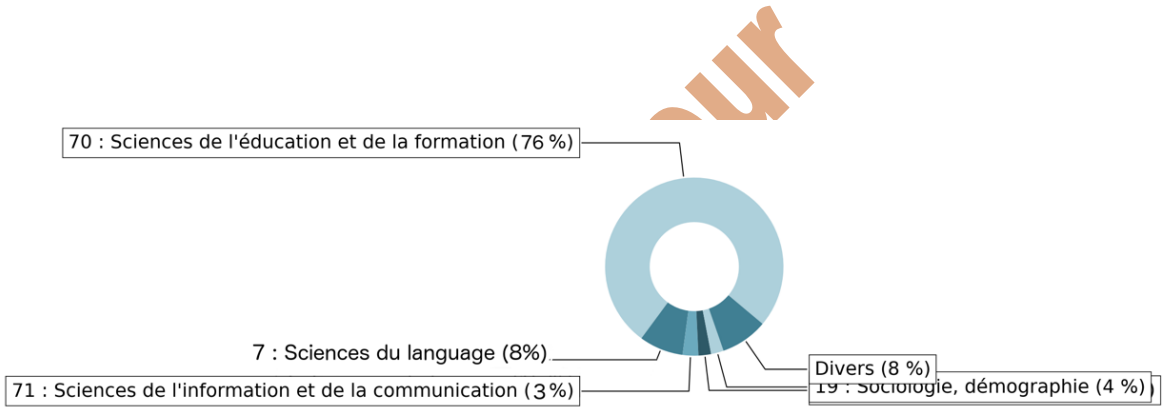

Figure 2 : Répartition des thèses du noyau par section CNU

Section	Nombre de thèses
Sciences de l'éducation et de la formation	293
Sciences du langage	69
Psychologie et ergonomie	19
Sociologie, démographie	18
Sciences de l'information et de la communication	7
Sciences et techniques des activités physiques et sportives	7
Philosophie	5
Sciences économiques	4
Langue et littérature françaises	4
Informatique	4
Génie informatique, automatique et traitement du signal	3
Sciences de gestion et du management	3
Mécanique, génie mécanique, génie civil	1
Droit privé et sciences criminelles	1
Science politique	1
Études anglophones	1
Aménagement de l'espace, urbanisme	1
Biologie des populations et écologie	1
Mathématiques	1
Langues et littératures anciennes	1

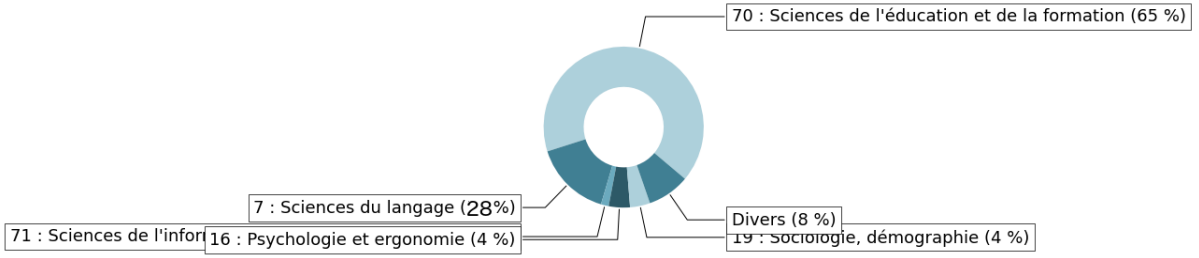

Figure 3 : Répartition des thèses du second cercle par section CNU

Dewey	Nombre de thèses
370 : Éducation	122
410 : Linguistique	10
300 : Sciences sociales	9
400 : Langues	7
150 : Psychologie	4
020 : Bibliothéconomie et sciences de l'information	4
610 : Sciences médicales Médecine	3
330 : Économie politique	3
650 : Gestion et services auxiliaires	3
320 : Science politique	2
440 : Langues romanes Français	1
790 : Loisirs et arts du spectacle	1
620 : Ingénierie et techniques connexes	1
840 : Littératures des langues romanes	1
100 : Philosophie et psychologie	1
630 : Agriculture	1
000 : Généralités	1

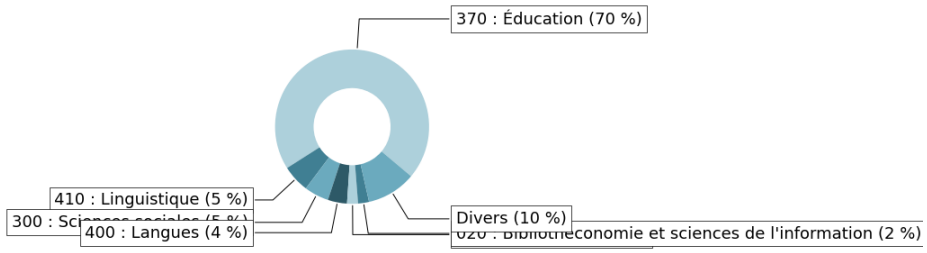

Figure 4 : répartition des thèses du noyau par leurs catégories de dépôt selon la classification Dewey

Dewey	Nombre de thèses
370 : Éducation	277
400 : Langues	36
300 : Sciences sociales	34
410 : Linguistique	19
440 : Langues romanes Français	16
150 : Psychologie	15
790 : Loisirs et arts du spectacle	6
620 : Ingénierie et techniques connexes	6
100 : Philosophie et psychologie	5
020 : Bibliothéconomie et sciences de l'information	5
330 : Économie politique	4
650 : Gestion et services auxiliaires	3
610 : Sciences médicales Médecine	2
630 : Agriculture	2
190 : Philosophie occidentale moderne	2
530 : Physique	1
460 : Espagnol et portugais	1
360 : Problèmes et services sociaux; associations	1
420 : Anglais et vieil anglais	1
320 : Science politique	1
340 : Droit	1
510 : Mathématiques	1

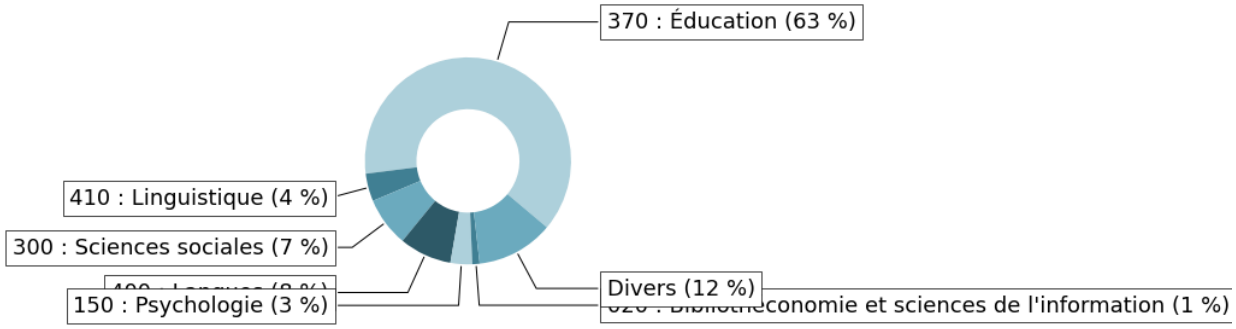

Figure 5 : répartition des thèses du second cercle par leurs catégories de dépôt selon la classification Dewey

2.3. Répartition par année de soutenance

La figure 6 présente les variations des nombres de soutenances par année, respectivement pour le noyau dur puis le second cercle. On observe que le nombre de soutenances de thèses du noyau dur se situe un peu en dessous d'une vingtaine par an (entre 11 en 2013 et 22 en 2018) alors qu'il est proche de la cinquantaine pour le second cercle (entre 36 en 2010, 2011 et 2019 et 63 en 2014). On voit aussi que la proportion entre second cercle et noyau augmente jusqu'à 2014 puis se réduit ensuite.

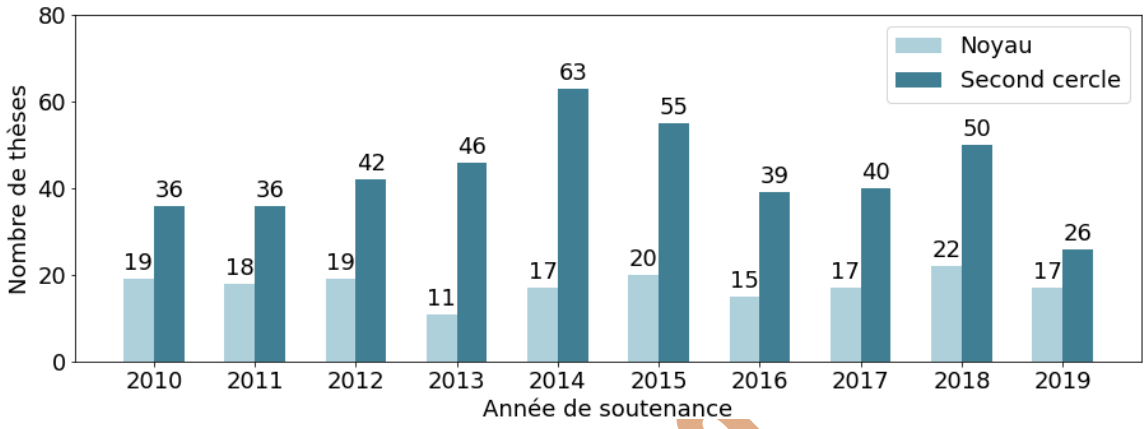

Figure 6 : répartition annuelle des thèses du noyau et du second cercle

2.4. Répartition par établissement de soutenance (code d'établissement issu du N° de thèse)

La figure 7 quant à elle présente la répartition des 619 thèses par établissement de soutenance. On y note en particulier la forte contribution du pôle Aix-Marseille au second cercle.

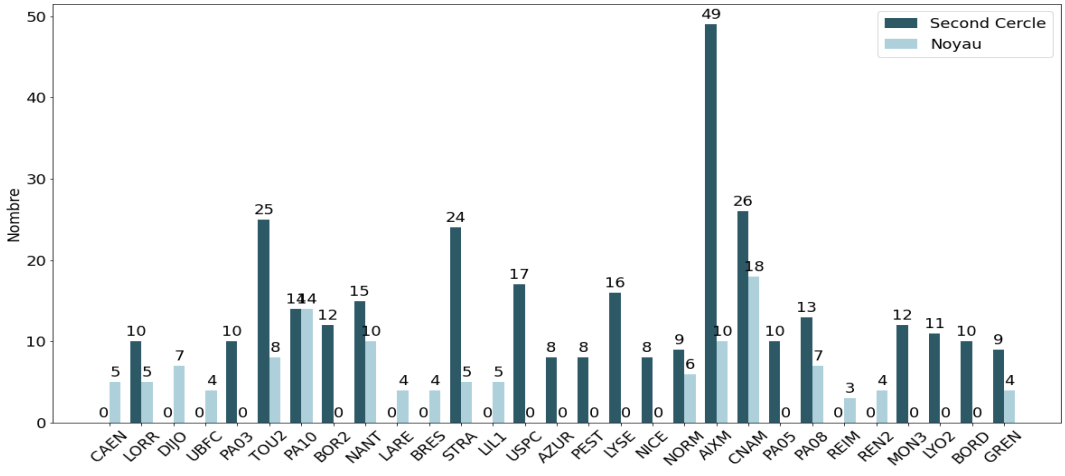

Figure 7 : répartition des thèses du noyau et du second cercle par établissement de soutenance

3. Spécificités du corpus RFdA par rapport aux autres thématiques : âge des doctorants

Dès lors que l'on cherche à mettre en évidence des particularités de ce champ de la RFdA, il est logique de s'intéresser à l'âge des doctorants, puisque l'on peut penser que ce sont des personnes déjà engagées dans une vie adulte, qui se consacrent à cette thématique. Un premier résultat limité aux thèses de la période 2014-2017 avait déjà été présenté dans l'article VdR5.

Ce résultat est confirmé quand on considère tout le corpus RFdA, comme on le constate sur la figure 8. Concernant les 175 thèses du noyau, les âges de soutenance se répartissent de manière quasi-uniforme jusqu'à 65 ans (médiane à 43 ans, entre 6 à 10 thèses sur chaque tranche de dix ans) ; le second cercle confirme aussi des particularités d'âge, avec une distribution légèrement plus en cloche et une médiane à 40 ans. Cela s'explique sans doute par la plus forte présence de thèses soutenues dans des disciplines autres que les sciences de l'éducation correspondant à une sociologie différente des impétrants.

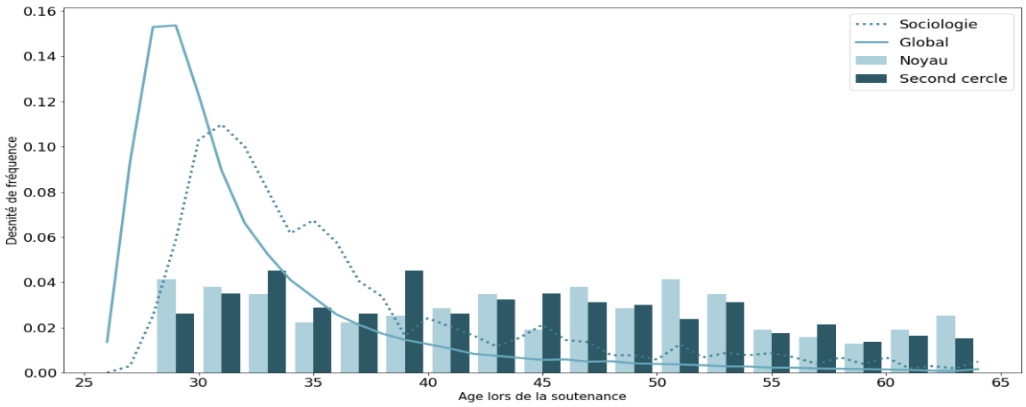

Figure 8 : répartition des thèses par âges des doctorants au moment de leur soutenance

Pour faciliter la comparaison avec d'autres champs et disciplines, la figure 8 indique aussi la répartition correspondant aux thèses de sociologie ainsi qu'à la totalité des soutenances, toutes disciplines du Sudoc regroupées. On y observe (dans le prolongement de ce qui avait été proposé dans l'article VdR5) que l'âge médian de soutenance toutes disciplines confondues est situé à 29 ans ; en sociologie il est de 33 ans.

4. Analyse lexicale des thématiques abordées dans les thèses du noyau dur et du second cercle

Les logiciels d'analyse lexicale comme le logiciel libre Iramuteq permettent d'effectuer des repérages et des classifications automatiques des thématiques traitées dans ces thèses de manière complémentaire de ce qui peut être fait par des lectures humaines flottantes. Avec Iramuteq, il est ainsi possible de repérer des classes de mots, regroupant chacune ceux qui sont les plus fréquemment utilisés ensemble dans les mêmes résumés de thèses en utilisant un algorithme développé par Max Reinert (1987)[3]. Ce repérage de telles classes de mots permet ainsi de disposer symétriquement d'une classification des résumés[4] (et par conséquent des thèses) en groupes correspondant à ceux qui utilisent le plus telle ou telle classe de mot. Cette méthode a ainsi été utilisée sur ce double corpus.

Une première analyse brute des 619 titres et résumés montre qu'ils utilisent des termes qui peuvent d'abord se regrouper en quatre grandes classes, (analyse de l'activité des enseignants, politiques de

[3] Antérieurement uniquement diffusé commercialement dans le cadre du logiciel Alceste®

[4] En fait Iramuteq permet de choisir de chercher les classes de mots les plus utilisés ensemble soit au sein des mêmes textes (ici les résumés de thèses) soit plus finement au sein juste des mêmes phrases. Ici c'est la première option qui est retenue, ce qui fait que les classes de mots correspondent exactement à des classes de thèses.

formation, apprentissages linguistiques des adultes, approches biographiques et identitaires) comme on le voit sur le dendrogramme de la figure 9.

En utilisant les possibilités offertes par cet algorithme de chercher à affiner la classification[5], on peut observer la fracturation de ces classes, ce qui aboutit à la figure 10. On observe que les deux classes « politique » et « langue » restent homogènes, mais que les deux autres sont fragmentables chacune en deux : la classe « identitaire » s'est scindée en deux sous classes (l'une qui reste centrée sur « vie, biographie, récit, identitaire » et une autre qui est focalisée sur « apprenance, collectif, soignant ») tandis que la classe « activité, enseignant » quant à elle s'est séparée en un groupe « élève, utilisation, outil, enseignant » et en un autre « auto confrontation, activité » clairement relié à l'analyse de l'activité et la didactique professionnelle.

Figure 9 : Classification en quatre classes des mots utilisés les plus fréquemment ensemble dans les titres et résumés des thèses du noyau dur et second cercle.

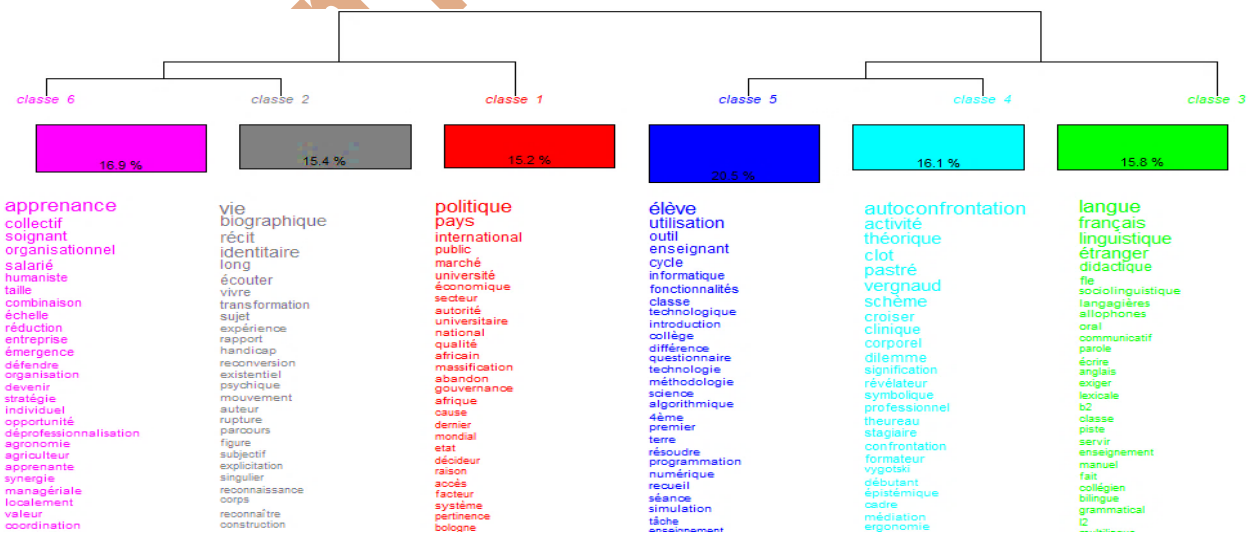

Figure 10 : Classification en six classes des mots utilisés les plus fréquemment ensemble dans les titres et résumés des thèses du noyau dur et second cercle.

[5] Il est en effet possible de paramétrer des analyses plus ou moins fines dans le logiciel Iramuteq en jouant sur le paramètre NCT (nombre de classes terminales souhaitées).

Cette classification plus poussée fait donc apparaitre finalement six grandes classes thématiques, détaillées figure 10 :

- Classe 1 (regroupant 15% des thèses) : « politique, pays international » correspondant aux thèses centrées sur les politiques de formation en particulier économiques (donc en rapport avec la thématique appelée précédemment EDM/POA)
- Classe 2 (15%) : « vie, biographie, récit », correspondant ainsi aux thèses centrées sur les histoires de vie et la biographisation, (donc en rapport avec la thématique appelée précédemment IOX)
- Classe 3 (16%) : « langue, français, linguistique » en rapport avec la thématique appelée précédemment LAN
- Classe 4 (16%) : « autoconfrontation, activité, théorique » correspondant ainsi aux thèses centrées sur l'analyse de l'activité et la didactique professionnelle (donc en rapport avec la thématique appelée précédemment DFP)
- Classe 5 (21%) : « élève, utilisation, outil, enseignants », correspondant ainsi aux thèses centrées sur le système scolaire et la formation des maitres, (donc en rapport avec la thématique appelée précédemment SCO)
- Classe 6 (17%) : « apprenance, collectif, soignant », correspondant à un sous ensemble de la catégorie IOX concernant la motivation ainsi qu'aux travaux liés aux professionnels de santé.

On ne peut néanmoins qu'être un peu étonné du faible poids des questions liées à l'ingénierie socio-économique et juridique de formation et à la relation formation et emploi dans ces travaux du corpus. Est-ce réellement dû à une forme d'isolationnisme de certaines disciplines (qui se marquerait par la faible interpénétration des jurys, déjà soupçonnée dans l'article VdR5), à un problème de sémantique pour désigner ces travaux ou alors est-ce le témoin d'un vrai déficit de production. Cette question renvoie à l'analyse qualitative mais aussi quantitative faite voici quelques années par Frétigné et de Lescure (2007) sur l'articulation entre sociologie et formation.

4.2. Différences entre classification automatique lexicale et liste des catégories vérificatoires

On observe que cette classification Iramuteq en six classes est proche de la liste des catégories repérées précédemment lors de l'analyse en lecture flottante des 175 thèses du noyau dur (cf. encadré 1), mais ne la recoupe qu'en partie. Ainsi, c'est une seule et même classe « élève, utilisation, outil » qui correspond aux deux catégories SCO et UJI et la seule et même classe « politique, pays international » qui correspond aux catégories EDM et POA ; inversement les deux classes distinguées en poussant l'analyse à 6 classes (« vie, biographique » d'un côté et « apprenance, collectif, soignants ») correspondent à la seule catégorie IOX.

Pour bien interpréter cette différence entre ces deux façons de classer les thèses, il faut avoir en tête les spécificités des deux méthodes employées pour construire ces classifications : pour ce qui est de la première (encadré 1), elle a été construite de manière heuristique par un lecteur humain qui a ajouté des catégories au fur et à mesure de la lecture des sujets des 175 thèses du noyau dur dès lors qu'il lui apparaissait une nouvelle thématique correspondant à sa représentation de la RFdA et donc une seule occurrence dans le noyau dur suffisait pour qu'une classe soit créée ; a contrario, dans le cas de l'algorithme Iramuteq®, la création d'une classe nécessite qu'après l'analyse de tous les mots dans tous les résumés de thèse, il ait été repéré un groupe suffisamment important de résumés utilisant simultanément les mêmes mots. De plus la lecture flottante a eu lieu sur le seul noyau dur, alors que

la classification Iramuteq® s'est effectuée sur l'ensemble noyau + second cercle, ce qui signifie que la classification automatique intègre les thèses dites « partiellement intéressantes pour la RFdA ».

4.3. Relations entre les classes de thèmatiques et des variables descriptives des thèses

Au delà de cette simple identification des 4 ou 6 classes thématiques, le logiciel Iramuteq® permet aussi de regarder leur répartiton selon les modalités prises par différentes variables descriptives des thèses.

La figure 13 montre ainsi comment certaines de ces classes de mots et de phrases sont liées plutôt au noyau dur ou au second cercle : on y observe que les classes 2 et 5 (« scolaire » et « langue » sont plutôt caractéristiques du second cercle tandis que la classe 3 (« politique et économie » de la formation) l'est plus du noyau, tout comme à un degré moindre la classe 1 (« histoire de vie, biographisation ») ; cela signifie fort logiquement que ces deux thématiques sont celles qui sont les plus présentes dans les résumés des thèses du noyau de RFdA de la dernière décennie. L'échelle verticale de ces figures produites par le logiciel Iramuteq® est fondée sur le coefficient chi² calculé entre la modalité de la variable (par exemple « Noyau » pour la variable « corpus » notée « *Co_ ») et les mots de la classe concernée : une valeur positive élevée indique une forte corrélation, une valeur quasi nulle, une absence de relation et une valeur négative une anti-corrélation.

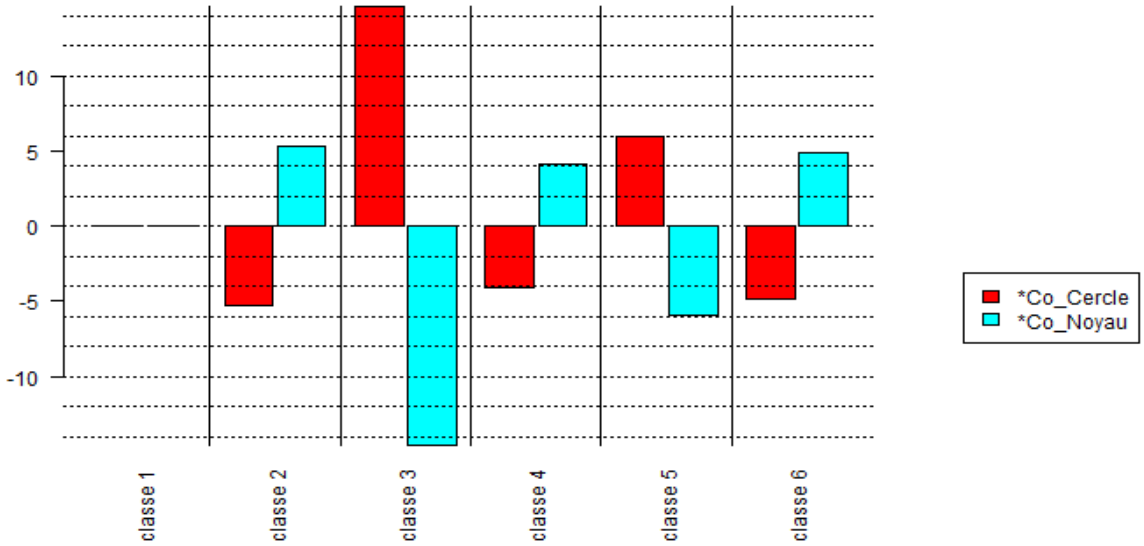

Figure 11 : relation entre les six classes de mots et l'appartenance ou non aux noyaux durs ou second cercle

On observe ainsi sans surprise que les classes 2 (vie, biographie), 4 (autoconfrontation, activité) et 6 (apprenance, collectif, soignant) sont bien effectivement fortement corrélées au noyau dur, et que les

classes 3 (langue, linguistique) et 5 (élève, utilisation, outil), moins spécifiquement RFdA, le sont bien au second cercle.

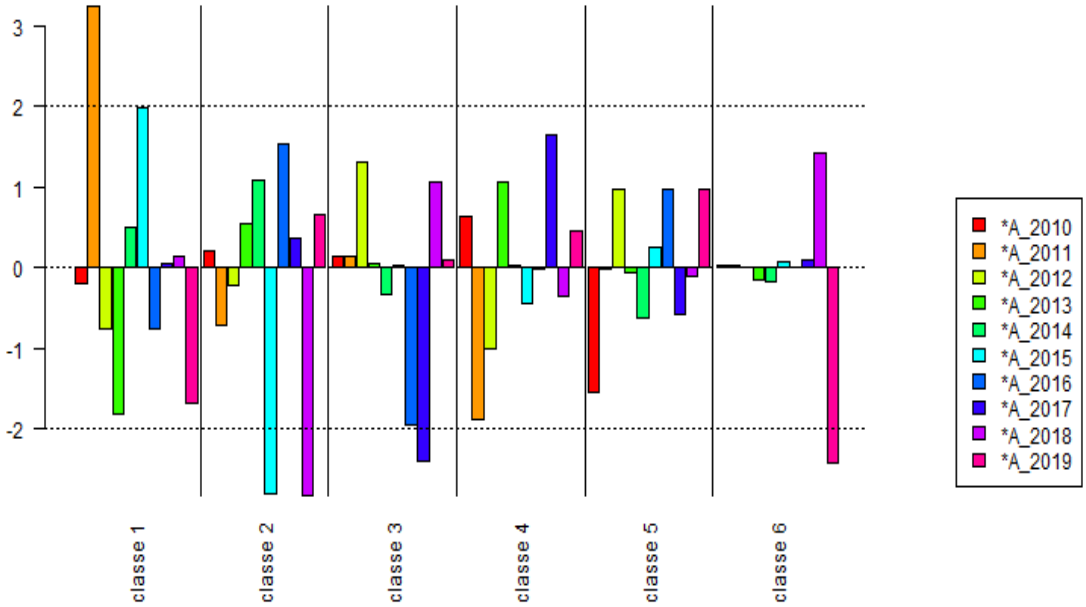

Figure 1 : relations entre les six classes de mots et les années de soutenance

En ce qui concerne les variations temporelles, la Figure 1 montre que les corrélations des thématiques avec les années sont très erratiques, ce qui s'explique par le fait qu'il n'y a en moyenne qu'une dizaine de thèses soutenues chaque année dans chaque classe. Dans de telles conditions, la conclusion que l'on peut tirer de ce graphique est qu'il n'y a pas d'évolutions pérennes majeures de la répartition entre les 6 classes thématiques sur cette décennie.

Pour ce qui est des sections CNU, la figure équivalente (Figure 13) donne surtout à voir le poids tout à fait logique des thèses soutenues en section 7 (sciences du langage) qui se retrouvent sans surprise dominantes en classe 3 (classe de mots centrés sur « langue, français, linguistique ») ; ensuite on observe la corrélation logique entre la classe 1 (« politique, pays ») et la section 19 (sociologie, démographie) ainsi que les sections 4 (sciences économiques) et 5 (sciences politiques) et on note que c'est la classe 4 (« autoconfrontation, activité ») fortement marquée par la didactique professionnelle qui se trouve la plus corrélée à la 70e section (science de l'éducation). On observe enfin aussi une relation somme toute logique entre la classe 5 (« élève, utilisation, outil ») et la section 27 (informatique) ainsi qu'entre la classe 6 (« apprenance, soignant ») et la section 6 (sciences de gestion).

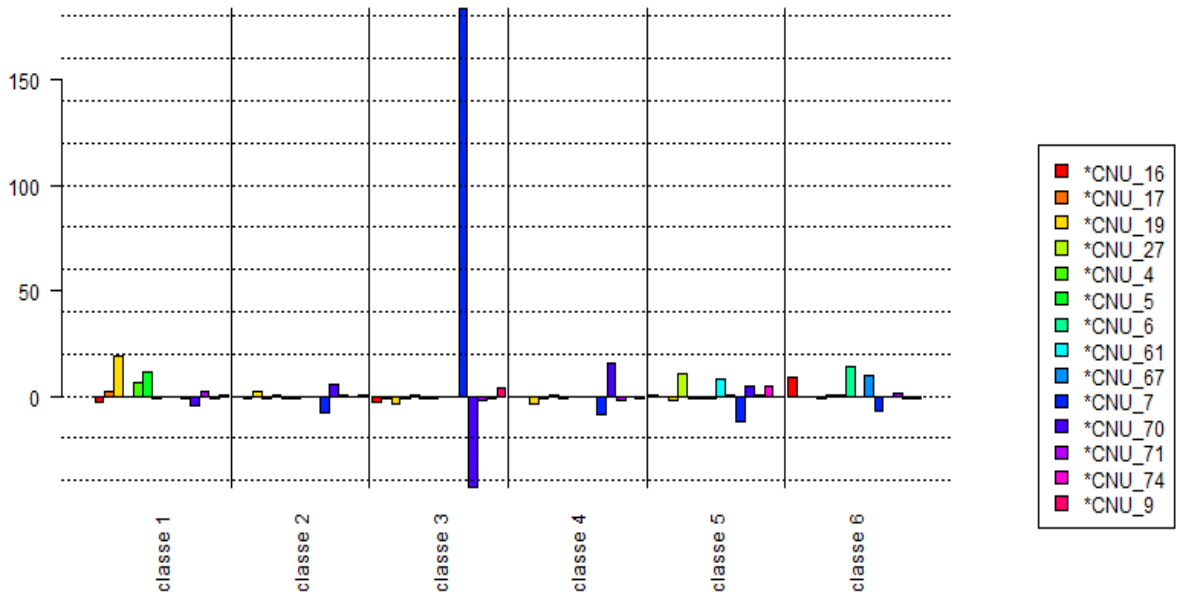

Figure 13 : appartenance aux sections CNU de soutenance par classe

4.4. Cartographie générale des relations entre classes et variables

Plus globalement, les deux figures (Figure 2 et Figure 3, faites pour se regarder en ayant été superposées) permettent de visualiser les principales relations entre les classes de mots (visibles sur la figure 16) et les modalités prises par l'ensemble des variables descriptives des thèses (visibles sur la figure 17 à savoir l'année, la section CNU, le classement bibliographique Dewey et le code de l'établissement de soutenance). En effet, en plus de fournir l'arborescence des classes de mots et les figures donnant les chi² par modalités de variables et classes de mots, le logiciel Iramuteq® permet également de projeter graphiquement, selon une analyse factorielle des correspondances (AFC), ces différentes classes.

Cette visualisation fournit alors des paires d'images superposables représentant pour l'une (Figure 2) les proximités relatives des mots et pour l'autre (Figure 3) les modalités des types de segments de textes concernés (via des tags ou variables prédéfinis) et ce autour des centres de ces lexiques. Les mots centraux sont les plus communs et la distance au centre indique la spécificité de tel ou tel mot. Les axes maximisent mathématiquement la visibilité des spécificités, mais leur orientation sur la page (haut / bas et droite / gauche) sont arbitraires.

Même si une telle paire de figures semble au premier abord un peu complexe à déchiffrer, elle se révèle particulièrement intéressante pour visualiser les contributions des différents établissements (*ETAB_ suivi du code à 4 lettres du nom de l'établissement utilisé dans les numéros nationaux de thèse) ou sections de soutenance (*CNU_) à la RFdA.

Pour ce qui concerne les dépôts dans les bibliothèques, ces figures permettent de voir dans quelle section des catalogues Dewey ces thèses correspondantes ont été affectées (*DEW_, suivi de la cote Dewey à trois chiffres correspondante pour situer les cotations fines, ainsi que *DEW2_ suivi de la code Dewey arrondie aux centaines pour situer les grandes sections bibliothécaires). Les lecteurs pourront aisément faire usage de ces figures afin de repérer par exemple les spécificités des différents établissements supports des écoles doctorales et laboratoires.

5. Conclusion

Avec l'obtention dans l'article précédent d'un tel double corpus de référence et les analyses descriptives qui en ont été fournies dans cet article, il est donc possible de se faire une idée d'ensemble de la contribution que les travaux doctoraux soutenus en France apportent ces dernières années à une meilleure compréhension de la formation des adultes.

Comme cela avait déjà été observé plus généralement dans l'article VdR4, on voit bien que l'on a affaire à des travaux conduits à partir de perspectives regardant à différentes échelles.

L'échelle macro-sociale est présente dans le noyau avec les travaux de la classe 1 « politique, pays… » (15%) de même que l'échelle micro-individuelle au travers de la classe 2 « vie, biographie » (également 15%). Les quatre autres classes ont plus ou moins à voir avec l'échelle méso, qu'il s'agisse de la regarder dans le noyau sous l'angle de l'analyse de l'activité et didactique professionnelle (classe 4, 16%) ou de la mobilisation de l'apprenance par les soignants ou les autres collectifs (17%), ou dans le second cercle sous les angles des dispositifs linguistiques (classe 3, 16%) ou de formation des maitres (classe 5 : 21%),

Il est aussi très facile de constater que cet état de fait résulte de contributions multiples venant de différentes écoles doctorales, alimentant plus ou moins le noyau ou le second cercle, agissant un peu comme la construction d'un tableau impressionniste, en apportant des touches de couleurs souvent complémentaires (voir figure 15)[6].

Dans cet esprit, il est frappant que ce patchwork reste temporellement relativement en équilibre : les répartitions annuelles montrent que ces proportions entre ces différentes échelles sont à peu près stables sur la décennie. On n'observe pas de montée en puissance d'une perspective plus qu'une autre. Cela signifie en particulier que les thématiques qui émergent (comme par exemple celles liées aux vagues de technologies ou aux spécificités des différentes réformes) restent contenues dans les grandes familles de préoccupations pérennes correspondant aux quatre ou six classes décrites.

Notes des auteurs :

1) Toutes les figures (en version couleur) et tous les tableaux créés pour ce chapitre ou simplement cités sont disponibles (au format tableur) en ligne en archives ouvertes via l'adresse https://gitlab.com/pbury/vdr7 et le Q/R code ci-contre.

2) Le double corpus étudié ici est le même que celui publié dans le numéro précédent. Les données sont donc arrêtées au 1er mai 2020, ce qui explique une certaine sous-représentation des données très récentes, liée en particulier au délai des saisies post soutenance dans le catalogue Sudoc. Il y a bien 175 thèses dans le noyau dur et non 174 comme indiqué dans l'article VdR6.

[6] Ce phénomène est particulièrement visible sur une carte de France animée disponible dans les documents fournis en ligne à https://gitlab.com/pbury/vdr7

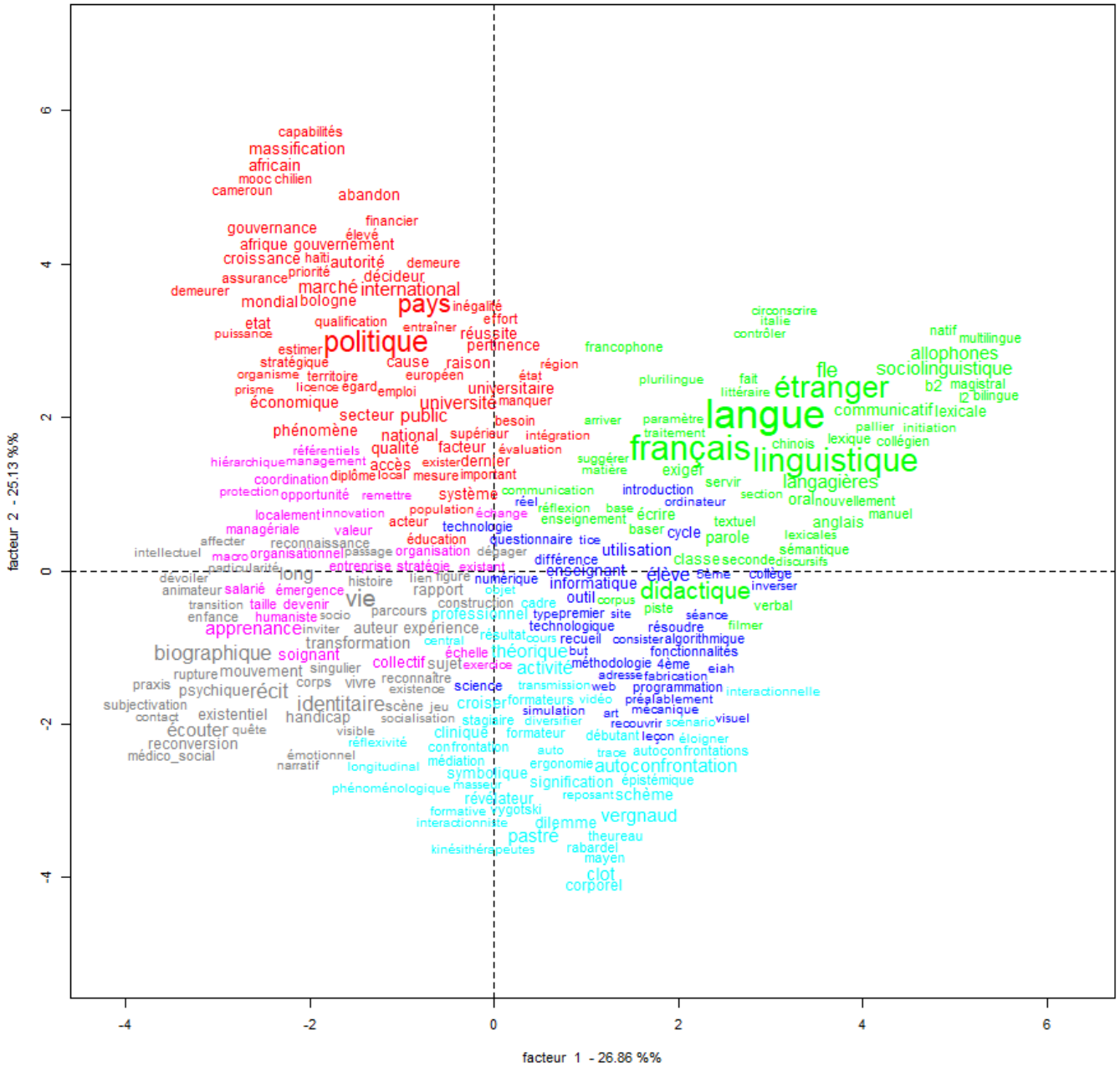

Figure 2 : proximités relatives des mots dans le plan 1,2 d'une analyse factorielle des centres des six classes

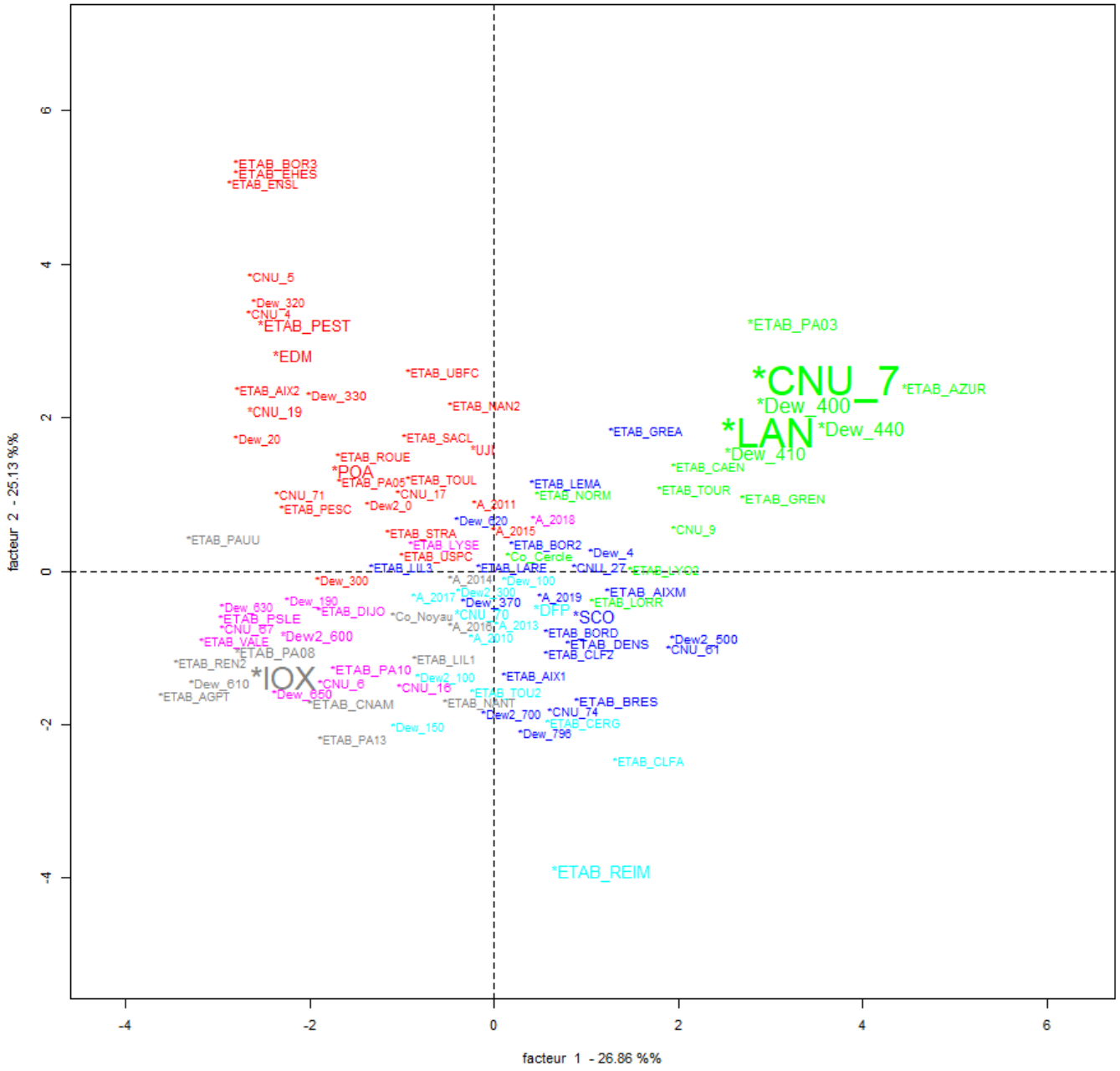

Figure 3 : modalités des textes par section CNU, établissements et codes Dewey (à superposer à la figure 16)

Figure 16 : localisation des soutenances des thèses du double corpus selon les années. Une figure animée en couleur est aussi disponible en ligne à https://gitlab.com/pbury/vdr7/-/blob/master/cartes/theses_vs_time.gif

Références bibliographiques

Fretigné C., de Lescure E. (2007). Sociologie et Formation en France, Note de synthèse. revue *Savoirs* N°15. Paris : L'Harmattan.

Laot F. (2006). « Les thèses en formation d'adultes. », *Savoirs* 1/2006 (n° 10), p. 129-132 : http://www.cairn.info/revue-savoirs-2006-1-page-129.htm .

Las Vergnas O. (2016, 2017). Rubrique « vie de la recherche », articles N°1, N°2, N°3 et N°4. Revue *Savoirs.* (N°40, 41, 42 et 43) Paris : L'Harmattan. Parus simultanément dans la Revue *TransFormations*, (N° 15-16 et 17) Lille : Université de Lille –CIREL

Las Vergnas O. et Bury P. (2018, 2019). Rubrique « vie de la recherche », articles N°5 et N°6. Revue *Savoirs.* (N°48 et 53) Paris : L'Harmattan. Parus simultanément dans la Revue *TransFormations*, (N°18) Lille : Université de Lille –CIREL

Ratinaud P. et Déjean S. (2009). IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009). Toulouse - Le Mirail.

Reinert M. (1987). Un logiciel d'analyse lexicale. *Cahiers analyse des données*, 11-4, 471-484. En ligne à http://www.numdam.org/item/CAD_1986__11_4_471_0

Walliser B. (2009) La cumulativité du savoir en SHS, Paris, Editions Éditions de l'École des hautes études en sciences sociales