## Rubrique « Vie de la recherche en formation des adultes »
**commune aux Revues *Savoirs* et *TransFormations***

# N°5 – Repérage des thèses traitant de la formation des adultes : étude des âges et des rattachements disciplinaires

***Olivier Las Vergnas, CIREL-Trigone (EA 4354) et CREF-AFA (EA 1589)***
***et Patrick Bury, Société Clevermind (Data Scientist)***

Résumé : Cet article est le cinquième de la rubrique dédiée à la recherche francophone sur la formation des adultes. Il est consacré aux caractéristiques des thèses récentes liées à ce champ. Après avoir complété la liste des 40 thèses soutenues en 2014-2015 (voir article N°1 de la présente rubrique) par 42 autres soutenues en 2016-2017, il s'intéresse à leur adossement disciplinaire ainsi qu'à l'âge des doctorants. Ce travail confirme d'abord deux constats antérieurs : d'une part le fait qu'il n'est toujours pas possible de définir un contour précis à ce champ via des requêtes bibliométriques en raison du caractère désuet l'appellation « éducation des adultes » encore employée dans l'indexation, d'autre part son caractère largement multidisciplinaire. Si l'on considère néanmoins que cette liste de thèses fournit une vision d'ensemble du noyau de la thématique, on peut tirer deux autres conclusions. La première concerne la spécificité des pyramides des âges des doctorants et jury (âgés de 18 ans de plus que la médiane de ceux de tous les autres champs de recherche confondus) ; la seconde porte sur les disciplines de rattachement (plus de la moitié en sciences de l'éducation et l'autre moitié très dispersées, avec quelques isolats dans lesquels les jurys sont spécifiques, comme les aspects langagiers). En revanche, les analyses lexicales temporelles se confirment toujours comme étant excessivement difficiles à conduire, au regard de la grande dispersion des centres d'intérêt et des vocabulaires, en lien avec l'absence catastrophique de thésaurus.

Abstract: This is the fifth article in the section dedicated to the observation of Francophone research on adult education. It is devoted to the characteristics of recent PhD theses related to this field. After completing the list of 40 theses defended in 2014-2015 (see article N°1 in this section) with 42 others defended in 2016-2017 identified through the same types of requests in the national thesis database (via sudoc.abes.fr), it focuses on their disciplinary affiliation as well as on the age of doctoral candidates. This work first confirms the two main observations presented in the first issue of this section: on the one hand, the fact that it is not possible to define a precise outline for this field via bibliometric queries due to the outdated nature of the term "adult education" still used in indexing, and on the other hand, the largely multidisciplinary nature of the field. Subject to accepting the idea that this list of 82 theses nevertheless provides an overview of the core of the theme, two other conclusions can be drawn. The first concerns the age pyramids of doctoral candidates and juries (18 years older than the median of all other research fields combined), the disciplines to which they belong (more than half in education sciences and the other half very scattered, with some isolates or juries are specialized, such as language issues). On the other hand, temporal lexical analysis is always confirmed as being excessively difficult to conduct, given the wide dispersion of interests and vocabularies linked to the catastrophic absence of thesauri.

Mots clefs : publications de recherche, bibliométrie, lexicométrie, formation des adultes, analyse lexicale.
Key words: research publications, bibliometrics, lexicometry, adult education, lexical analysis.

*Note de l'auteur : les dernières données rapportées ici correspondent à des requêtes effectuées dans la base du Sudoc via Sudoc.abes.fr en juin et juillet 2018.*

Olivier Las Vergnas, CIREL-Trigone (EA 4354) et CREF-AFA (EA 1589) olivier.las-vergnas@univ-lille.fr
Patrick Bury, PhD sciences de l'ingénieur, Société CleverMind patrick.s.bury@gmail.com

## 1. Rappel du contexte, de la finalité et des difficultés

Ce cinquième article[1] de la rubrique « vie de la recherche en formation des adultes » s'inscrit dans la volonté de proposer un état dynamique de la production francophone de connaissances dans ce champ. Dans la suite des deux premiers, parus en 2016 (VdR1 et VdR2) et des travaux antérieurs de Beillerot (1993), Beillerot et Demory (1997), Leclercq (2007, 2008) et Laot (2007) ainsi que ceux de Laot dans la précédente version de cette rubrique dans la revue *Savoirs*, il actualise et approfondit l'analyse des caractéristiques des thèses d'université liées à cette thématique. Il s'intéresse à leurs spécificités en termes d'âge et d'origine disciplinaire des jurés.

De la même façon qu'avait été repéré dans l'article VdR1 un groupe d'une quarantaine[2] de thèses soutenues en 2014 et 2015, une deuxième liste de 42 thèses soutenues en 2016 et 2017[3] a été identifiée. Elles ont été repérées dans la base du Service universitaire de documentation (Sudoc, http://sudoc.abes.fr) par une requête générique sur les termes « formation des adultes », « éducation des adultes » ou « formation tout au long de la vie » complétée dans l'esprit du travail présenté dans l'article VdR3 (voir encadré 1) par une batterie d'autres requêtes plus fines. Celles –ci combinent des recherches complémentaires sur des mots bien plus spécialisés[4] (« apprenance », « échange de savoir » « VAE » « didactique professionnelle » « formation professionnelle continue » dans leur résumé) avec des filtrages destinés à vérifier qu'il ne s'agissait ni de thèses centrées sur les « enseignants » (c'est-à-dire le monde de la formation initiale), ni de thèses « d'exercice », en médecine, pharmacie ou maïeutique.

### Encadré 1 : Requêtes et difficultés d'extraction des thèses sur la formation des adultes

Comme précédemment (voir les articles VdR1 à VdR4), l'établissement de telles listes de publications caractéristiques du champ de la « recherche en formation des adultes » se heurte à l'absence d'une définition consensuelle explicite de ce champ mais aussi à des problèmes d'indexation. Ces deux obstacles sont en résonnance l'un avec l'autre, comme dans toute relation linguistique entre signifiants et signifiés.

Même si l'on se limite, comme c'est le cas dans cet article, à l'analyse des thèses d'université francophones, on est directement confronté à trois problèmes bibliographiques : (1) les mots clefs donnés par les auteurs sont peu utilisables car non hiérarchisés au sein d'un thésaurus et donc soit très flous, soit beaucoup trop fins et spécifiques, (2) les résumés des thèses varient en longueur du simple au triple, ne traitent pas vraiment les mêmes rubriques et se révèlent totalement disparates en termes de niveau de détails et de précision de vocabulaire, (3) l'indexation qui se fait dans les bibliothèques universitaires -en appui sur le langage Rameau- se révèle distordue, comme cela a été signalé en VdR1, par l'existence d'une vedette matière « éducation des adultes » qui véhicule une connotation datée, celle de la seule promotion sociale.

Dans ce contexte, l'article VdR3 avait comparé (non plus pour les seules thèses, mais sur l'ensemble des publications référencées dans HAL) deux méthodes extrêmes : la première se limitant à chercher sur les seuls termes génériques[5] « formation des adultes » et la seconde cherchant à englober tout ce qui peut être considéré comme en lien avec le champ grâce à un faisceau

[1] Dans la suite, les quatre articles précédents (Las Vergnas 2016 – 2017) parus simultanément dans les deux revues *Savoirs* et *TransFormations* seront désignés respectivement par VdR1, dR2, VdR3 et VdR4.
[2] Il y en avait en fait 40 différentes et non 41 comme indiqué dans VR1 car un doublon n'avait pas été identifié et a dû être enlevé *a posteriori*.
[3] Les requêtes à la base de ce travail ayant été effectuées à la fin du mois de juillet 2018, il n'y a pas de garantie que l'ensemble des thèses soutenues en 2017 aient été intégrées à la base.
[4] Dans l'article VdR1 un choix différent avait été fait, mais le résultat se révèle de fait très proche : nous avions cherché les thèses dont les résumés ou titres contenaient « formation continue » mais cela avait apporté de multiples thèses d'université hors champ (ne mentionnant qu'anecdotiquement une conséquence marginale en termes de FC) ce qui avait obligé à une laborieuse sélection heuristique pour ne pouvoir n'en garder qu'un tiers. Cette fois ci nous avons donc remplacé cette trop large requête complémentaire unique par cette série de mots plus spécifiques, selon une logique inspirée de l'article VdR3.
[5] Pour information, le même type de requête effectuée sur la totalité du Sudoc a conduit à identifier 275 thèses contenant strictement ces mêmes mots comme mots sujets.

d'une soixantaine de requêtes inspirées par un sommaire encyclopédique de référence (en l'occurrence celui du « Traité des sciences et des techniques de la formation » de Carré et Caspar, 2017).

Ce travail comparatif avait montré d'abord que la plupart des documents proposent des résumés ou des mots clefs beaucoup trop spécifiques pour pouvoir être repérés par une requête générique du type « formation des adultes » : il faut mélanger les deux types de recherche pour ne pas laisser passer l'essentiel de la production ; les listes que l'on établit alors ne peuvent de toute façon n'être que considérées comme un « noyau central » à faire ensuite grossir par des requêtes complémentaires.

Prétendre ainsi situer des frontières précises du champ est illusoire car deux types de problèmes se superposent. D'un côté le fait de retenir ou non une thèse identifiée via une requête pose le problème de déterminer la limite à partir de laquelle une thèse qui effleure la thématique devrait être rejetée : cela concerne en particulier l'acceptation ou non des quatre catégories signalées en VdR1 comme périphériques : la transition vers la vie adulte, le développement professionnel, les formations multimédia universitaire ou la formation des enseignants. D'un autre côté, aucune méthode ne peut garantir l'exhaustivité du recueil de « toutes » les thèses concernées : même en partant d'une liste encyclopédique de sous-thèmes, on ne peut pas s'assurer -faute de thésaurus partagé et stabilisé- que les choix de terminologies seraient pertinents au regard des pratiques rédactionnelles des rédacteurs des résumés.

## 2. Deux analyses de spécificités communes : l'âge et la dispersion disciplinaire

Les limitations présentées dans l'encadré 1 aboutissent à une conclusion simple : toute tentative d'analyse fine de variations annuelles ou biennales des thématiques abordées ne peut être séparée de la critique des requêtes aboutissant au corpus de thèses étudié, comme cela sera illustré dans la dernière partie de cet article. C'est pourquoi dans cet article ce sont des caractéristiques, qui seraient communes aux thèses sur la formation des adultes mais qui les différentieraient des autres champs, qui ont ainsi été recherchées. Deux aspects souvent évoqués dans la littérature antérieure ont ainsi été explorés : d'une part l'étendue des rattachements et origines disciplinaires de ces thèses et de leurs jurys et d'autre part la distribution des âges des doctorants et des membres des jurys.

Informatiquement, le travail a été mené grâce à l'aspiration des données de la base theses.fr du Sudoc via un court module en langage de programmation Python et à un outil informatique appelé « base de données orientée graphe », permettant de représenter simplement les réseaux relationnels entre ses composants. L'encadré 2 présente les raisons de ce choix, les modélisations mises en œuvre et les méthodes d'extractions de ces données.

### Encadré 2 : Modélisation et utilisation de base – graphe

**Source et base de données**

Lorsque l'on s'intéresse aux travaux de thèse le site « theses.fr » et le Sudoc sont précieux, ils permettent de rechercher dans une immense base, facilement et rapidement. En revanche, pour l'analyse, ils ne proposent qu'un simple moteur de recherche. Pour produire cet article, nous avons cherché un outil plus adapté à visualiser des relations par des graphes : Nous voulions en effet étudier les relations entre les thèses, leurs auteurs et membres de jury.

Dans le monde de l'informatique les bases de données relationnelles permettent facilement de stocker un grand nombre de relations et sont très connues, Toutefois, ces bases ne sont pas très performantes pour traiter de grands volumes de données, et encore moins pour explorer les données. De plus elles présentent un défaut important : le langage utilisé est éloigné du langage courant des chercheurs. À côté de celles-ci il existe une nouvelle catégorie de bases, les « bases graphes » spécifiquement adaptées à de telles études et requêtes. Ces dernières stockent les informations sous la forme d'un graphe avec des nœuds et des arcs (qui relient des nœuds entre eux, ont une nature et un sens ainsi que des propriétés). Ce type d'outil paraissant idéal pour travailler sur les relations entre les catégories que nous souhaitons étudier, nous avons retenu la base *Neo4j*, qui a l'avantage d'être fiable, bien documentée et gratuite si elle est utilisée sur une seule machine.

**Construction du modèle et types de données**

Pour construire les graphes nous intéressant à partir des données du Sudoc, nous avons défini quatre types de nœuds :
- les thèses, avec leurs titres, année de soutenance, identifiant ainsi que l'indication de son appartenance ou non à l'échantillon,

- les personnes, avec leurs noms et prénoms, date de naissance, date de décès et identifiant unique,
- les disciplines, avec leurs intitulés habituels,
- ce que nous avons appelé des méta-disciplines, avec des dénominations simplifiées communes à plusieurs disciplines.

Nous avons défini 4 relations principales, certaines directement à partir de la base Sudoc, d'autres après calcul :
1. JUGEE_PAR qui relie une thèse et un membre de son jury
2. ECRITE_PAR qui relie une personne et la thèse dont il est l'auteur
3. DU_DOMAINE qui indique de quelle discipline fait partie une thèse (utilisée pour relier un auteur et une thèse)
4. MEME_JURY qui relie les disciplines ayant jugé une même thèse ainsi que pour les Méta-disciplines

**Préparation de la base**

Nous avons donc dans un premier temps téléchargé toutes les thèses disponibles depuis le Sudoc dans une base locale, (en juillet 2018) car nous ne voulions pas ensuite surcharger ce serveur partagé par nos nombreuses requêtes simultanées, la phase d'acquisition a donc été longue (en nous limitant à environ une requête toutes les deux secondes, cela représente environ 20 jours de téléchargement en tache de fond). À l'issue de cette première étape, nous disposions en local d'environ 370 000 références de documents (thèses) ainsi que d'environ 460 000 notices, que nous avons injecté dans une base graphe locale. Nous avons nettoyé les données par exemple pour des intitulés des disciplines (« sciences du » en lieu et place de « sciences de »), des différences typographiques (« Science de », « Sciences de », « sciences de ») : ce nettoyage a été effectué à partir d'un fichier de référence, complété manuellement, en essayant de limiter au maximum le *'garbage in, garbage out'* qui s'avère toujours vrai, néanmoins il subsiste des écarts résiduels. Concernant les années prises en compte dans cette étude, nous disposons de 47 678 thèses pour les années comprises entre 2014 et 2017.

**Exemples de requêtes**

Dans le langage *cypher* qui est utilisé par la base *Neo4j*, la requête permettant de récupérer les données sur les âges des doctorants à la soutenance est la suivante : *MATCH ((these:Theses{final_2014_2016)-[r:ECRITE_PAR]-(p:Personne)) RETURN p.naissance, these.annee* que l'on peut lire comme suit : renvoyer l'année de naissance (p.naissance) et l'année de soutenance de la thèse (t.annee) de toutes les personnes ayant rédigé une thèse faisant partie de la présente étude. Pour les âges des membres de jury, la requête devient : *MATCH(these:These{final_2014_2016)-[r2:JUGEE_PAR]-(jure:Personne)) return juge.naissance, these.annee ;* pour les anciennetés dans le grade de docteur des membres de jury : *MATCH(these:These{final_2014_2016)-[r2:JUGEE_PAR]-(jure:Personne)-[:ECRITE_PAR]-(these_jure:These) return these_juge. annee , these. annee .*

Ces exemples de requêtes montrent la simplicité de la construction du modèle de données ainsi que sa facilité d'utilisation, dans un langage qui n'est pas abscons. Le problème clef est avant tout l'impérieuse nécessité de disposer de données de qualité, il faut donc particulièrement soigner les imports et les prétraitements apportés à celles-ci.

## 3. Première analyse : Affectation disciplinaire des thèses et mixités des origines des jurés

Dans la sélection de 82 thèses, 53 (près des deux tiers) ont été soutenues en sciences de l'éducation et certaines (issues du CNAM) sont repérées avec la mention d'une sous-spécialité « formation des adultes (22 parmi ces 53) ; les autres se répartissent entre les sciences du langage (9), le droit et la socio puis psycho, socio, lettres modernes, comme le montre la figure 1.

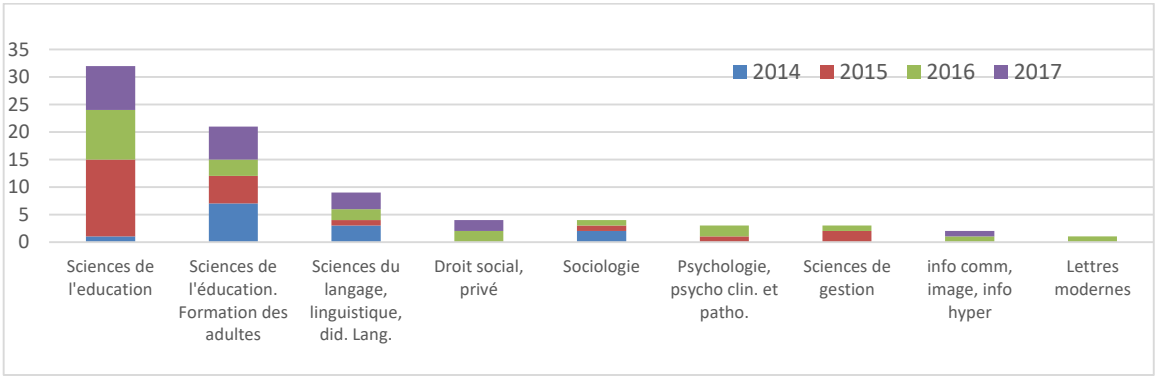



Figure 1 : répartition disciplinaire des thèses repérées comme liées à la formation des adultes dans le Sudoc (de 2014 à 2017)

On observe aussi qu'il n'y a au maximum qu'une ou deux thèses soutenues chaque année dans chaque autre discipline que les sciences de l'éducation : il en résulte que ces données sont trop peu denses pour prétendre tirer une quelconque généralisation de telles fluctuations erratiques. C'est pourquoi, afin de donner plus de visibilité à cette question de la relation aux disciplines académiques, est proposée une analyse plus fine s'intéressant cette fois à toutes celles représentées dans les jurys. La figure 2 indique ainsi les origines disciplinaires des participations aux jurys des thèses liées à la formation des adultes en 2014-2017. Elle présente le nombre de participations pour les disciplines qui en cumulent plus de 5 sur les 4 ans. On y retrouve sans surprise le poids des disciplines contributives que sont la sociologie et la psychologie.

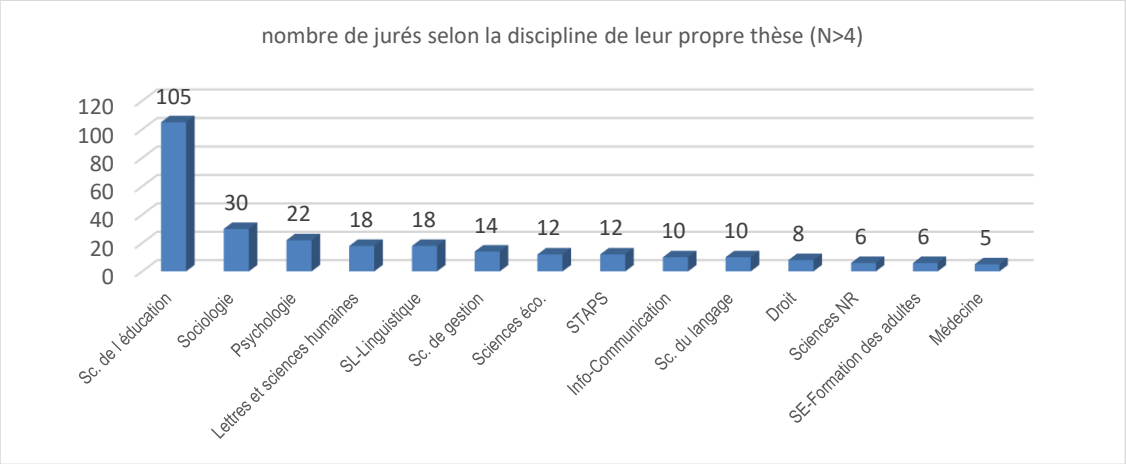



Figure 2 : origine disicplinaires des participations aux jurys des thèses liées à la formation des adultes en 2014-2017. Le graphique est limité aux disciplines qui ont fourni plus de 4 participations. « 8 » en droit, signifie qu'au total 8 participations à des jurys de soutenance ont été assurés par des titulaires d'un thèse d'université en droit.

Pour aller plus loin, les figures 3, 4 et 5 donnent des exemples de présentations plus fines des relations entre les disciplines mobilisées pour constituer les jurys de thèses et la discipline de la soutenance. Y sont représentés les origines disciplinaires des jurés.

Dans ces graphes, une flèche issue d'un disque comme « Sciences de l'éducation » et pointant vers un disque « Histoire » indique que parmi les thèses soutenues en « Sciences de l'éducation », un jury au moins a mobilisé un(e) juré (ou plusieurs) ayant lui-même soutenu sa thèse en « Histoire » ; plusieurs flèches orientées dans le même sens joignant les deux mêmes cercles « Sciences de gestion » et « Psychologie » indiquent que des thèses soutenues en Gestion ont mobilisé entre autres plusieurs jurés différents issus de la discipline « Psychologie » dans leurs jurys respectifs.

Cette modélisation originale (cf. encadré E2) a été appliquée à l'ensemble de ces 87 thèses soutenus entre 2014 et 2017. Il est néanmoins à noter que ces représentations ne sont pas complètement exhaustives : pour les 87 thèses finalement décrites par ces graphes (5 thèses liées à l'économie de la formation des adultes ont été ajoutées *a posteriori*, voir note 10) de la période 2014-2017, les données du Sudoc concernent seulement 409 jurés et ne sont renseignées que les disciplines de thèse de 251 de ceux-ci[6].

[6] La moitié des disciplines de membres de jury nous sont donc inconnues, toutefois, les tendances dégagées par cette première étude sont suffisamment fortes pour résister à ce biais.

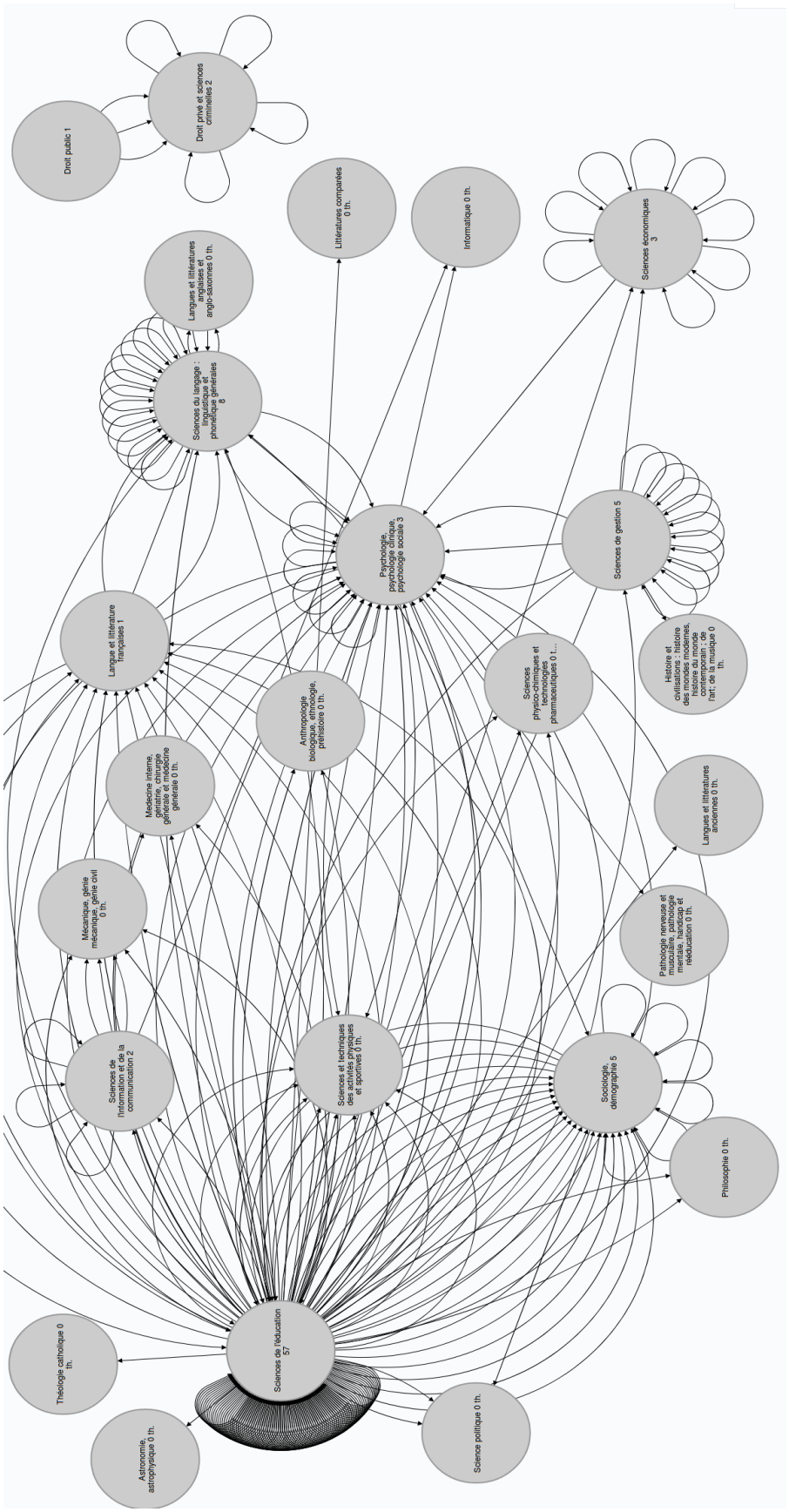
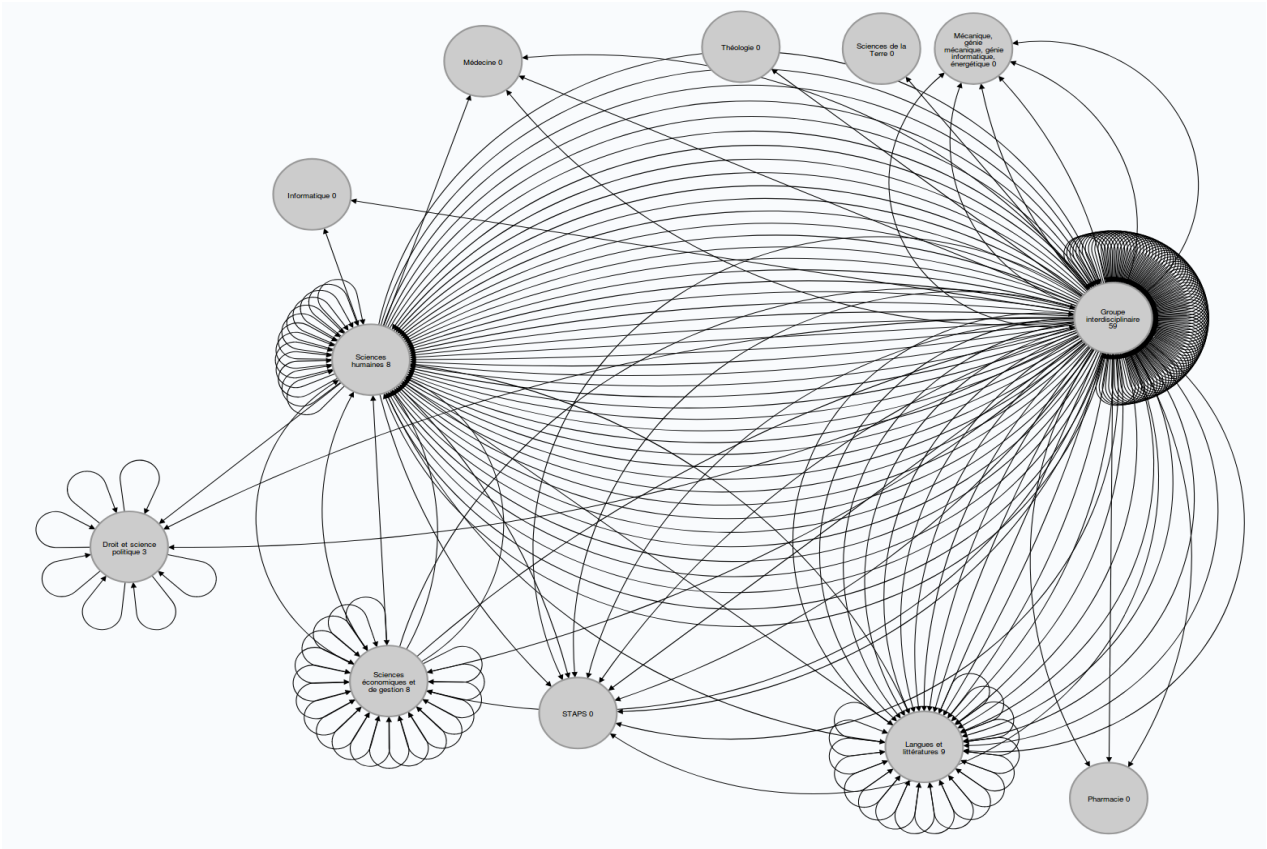
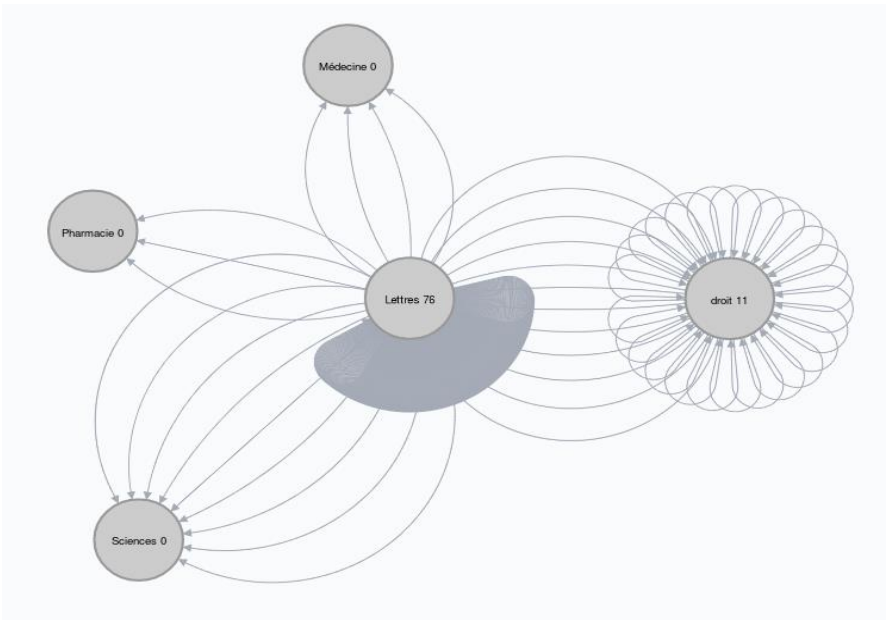



Figure 3 : graphe des recours aux diverses disciplines pour constituer les jurys des thèses francophones (2014 et 2017) concernant la formation des adultes. Une flèche issue de sciences de l'éducation et pointant vers astronomie indique que parmi les thèses soutenues en Sciences de l'éducation, a été mobilisé un juré ayant lui-même soutenu sa thèse en astronomie. Le nombre indiqué sur les cercles est le nombre de thèses soutenues dans la discipline. À noter à droite 3 thèses n'ayant mobilisé que des jurés titulaires d'un doctorat en droit.

Sur la figure 3, les disciplines mentionnées sont directement celles qui sont indiquées dans la base des thèses ; sur les figures 4 et 5, les graphes ont été densifiés en utilisant les groupements de disciplines en sous-groupes et en grandes disciplines à partir de l'arborescence CNU donnée dans le tableau placé en annexe. On retrouve sans surprise le poids des jurés issus des sciences de l'éducation (donc en figure 4 du « groupe interdisciplinaire » puis en figure 5 des « Lettres », auxquelles le CNU les a rattachées) mais aussi d'autres disciplines contributives, pour lesquelles on peut observer plus ou moins de consanguinité (cf. langues, sciences économiques ou droit).




Figures 4 et 5 : graphes des recours aux divers sous-groupes et disciplines du CNU (voir table en annexe) pour constituer les jurys des thèses francophones soutenues entre 2014 et 2017 concernant la formation des adultes. Le « groupe interdisciplinaire » CNU est constitué des sciences de l'éducation, de l'information/communication, de l'épistémologie des STAPS et des langues et cultures régionales.

Ces graphes visualisent les interconnexions entre positionnement disciplinaire de la soutenance et origine des expertises mobilisées pour le jury : sur la figure 4 plus simple à lire, on retrouve logiquement les rôles des méta-disciplines contributives (comme psychologie, sociologie, sciences économiques, politiques) et les disciplines majeures de soutenance situées au sein d'un réseau de sources d'expertises communes, plus ou moins dépendantes (avec un noyau central autour de sciences de l'éducation et didactique) ou indépendantes (en périphérie, comme linguistique d'un côté ou gestion de l'autre).

Certes, de telles modélisations sont -à ce stade- loin d'être exemptes de problèmes. Avant tout, il faut rappeler les importants problèmes concernant la liste des disciplines dans la base Sudoc : se cumulent des orthographes disparates et des confusions entre niveaux et sous-niveaux avec de réels glissements des nomenclatures académiques d'une période à l'autre. De plus, cet indicateur (la discipline) devant être « corrigé » manuellement, la méthode utilisée est difficilement généralisable à l'ensemble des disciplines. Au-delà de ces approximations propres au renseignement de la base elle-même, un autre problème, lui de méthode, résulte de l'assimilation entre la discipline de la thèse d'un juré et l'expertise que l'on a voulu mobiliser dans le jury. Il serait ainsi peut être plus pertinent dans l'avenir d'y substituer (ou d'y ajouter) la discipline des HDR ou thèses d'État des jurés dès lors qu'ils en auraient soutenu une dans une discipline différente.

## 4. Deuxième analyse : Age des doctorants et des jurées, délai de participation à un jury

Comme annoncé plus haut, un des éléments les plus significatifs est la pyramide des âges des impétrants au moment des doctorats. A ce propos, la figure 6 montre la répartition en pourcentage des âges des doctorants soutenant (sur la période 2014-2017) à propos de la formation des adultes et ce par différence avec plusieurs disciplines,

Est d'abord présentée une courbe correspondant à la totalité des doctorants (toutes disciplines du Sudoc regroupées) dont l'âge médian de soutenance s'est situé à 29 ans ; sont aussi figurées des courbes spécialisées pour les doctorants en mathématiques (âge médian légèrement plus jeune, Me=28 ans), ou la sociologie (Me=33 ans). Ce qui est frappant par contraste c'est la répartition des âges de soutenance des thèses de notre extraction correspondant aux recherches sur la formation des adultes : ceux-ci se distribuent de 25 à 66 ans, avec une médiane à 43 ans et une répartition décennale quasiment uniforme (entre 6 à 10 thèses sur chaque tranche de dix ans).

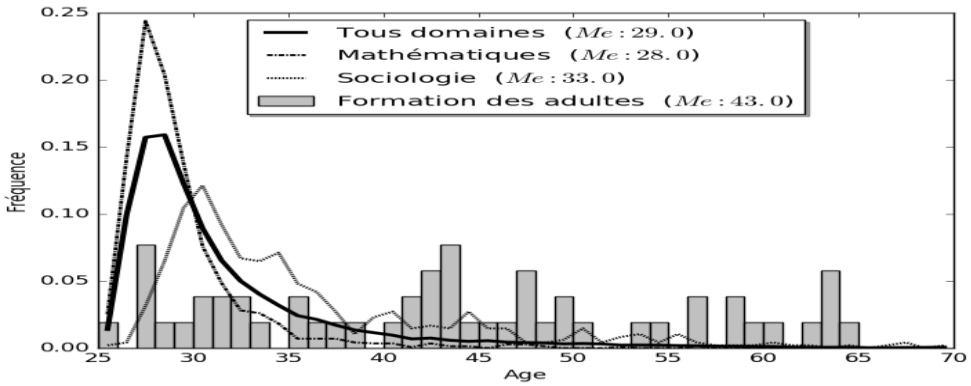



Figure 6 : Répartition en ratio des âges des doctorants en 2014 et 2017 selon les disciplines (0.12 sur l'axe vertical pour 31 ans signifie que 12% de ceux de sociologie avaient 31 ans lors de leur soutenance). La courbe pleine correspondant à la totalité des doctorants (toutes disciplines du Sudoc regroupées) celle en pointillés doubles aux doctorants en mathématiques ; celle en léger pointillé à la sociologie et l'histogramme gris à celle concernant la formation des adultes.

Le décalage de la médiane des âges des doctorats liés à la formation des adultes (+ 14 ans pour la médiane) par rapport aux autres domaines est donc de 14 ans. Face à un tel résultat on est conduit à se demander quelles en sont les conséquences sur les carrières ultérieures, par exemple en matière d'âge des premières participations à des jurys en tant que juré. Afin de donner une visibilité sur cette question, la figure 7 fournit, cette fois justement pour les membres de jurys, le délai écoulé entre leur première participation en tant que juré à une soutenance et leur propre soutenance : on observe que s'ajoute à nouveau un allongement de ces délais (+ 2 ans) pour la formation des adultes, qui est de 18 ans contre 16 ans sur l'ensemble des disciplines.

Une des conséquences est que les périodes d'activités de tels jurés sont forcément beaucoup plus courtes que dans les autres domaines et que les départs à la retraite des titulaires doivent y avoir beaucoup plus de répercussions.

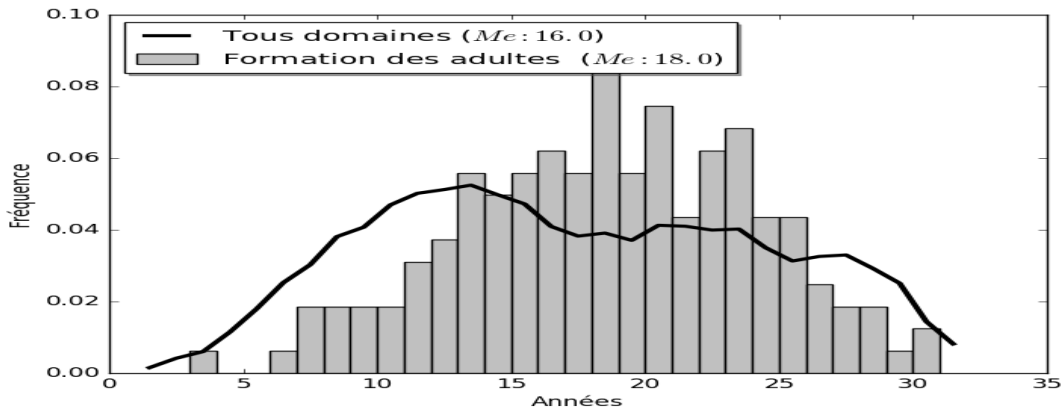



Figure 7 : Répartition pour les membres de jurys, des délais écoulés entre leur première participation en tant que juré à une soutenance et leur propre soutenance.

## 5. Regard comparatif sur les lexiques utilisés et critiques des extractions

Compte tenu des difficultés signalées plus haut concernant les études thématiques différentielles, une exploration comparative des deux corpus 2014-2015 et 2016-2017 a néanmoins été tentée pour mieux visualiser les effets pervers de ces problèmes d'extraction. Ainsi, en comparaison avec les analyses lexicales des 40 thèses pour la période 2014-2015 présentée dans l'article VdR1, une nouvelle analyse a été menée sur la sélection 2016-2017 toujours avec le logiciel *Iramuteq* (Reinert, 1987 ; Ratinaud et Déjean, 2009). Ces analyses s'appuient sur un comptage des mots signifiants employés dans les titres et résumés du corpus des thèses.

Pour rappel, la figure 8 donne un premier exemple du type de résultats que l'on peut espérer obtenir concernant l'infléchissement de l'usage des mots[7] entre les résumés de 2014-15 et 2016-17. Elle présente -pour les mots rencontrés au total sur les 4 ans plus de 25 fois dans les résumes- ceux qui ont le plus régressé ou au contraire progressé. On peut ainsi observer à gauche que les mots « action », « pédagogique » et « activité » (très présents en particulier en 2015) sont beaucoup moins utilisés en 2016-2017. Symétriquement, on observe que les termes « discours », « réseau » et « niveau » sont ceux qui ont le plus progressé en 2016-17 (surtout l'occurrence en 2016).

[7] Ou plus exactement de leur forme lemmatisée, c'est-à-dire regroupée sur le masculin singulier ou l'infinitif

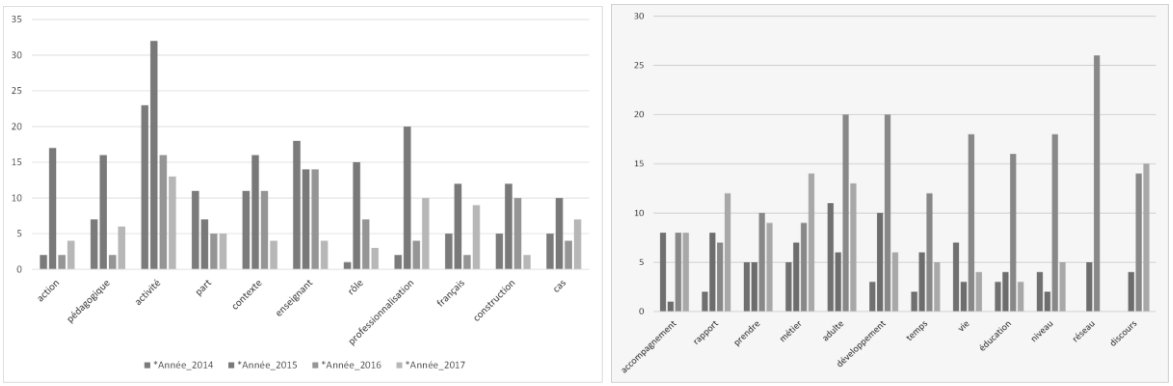



Figure 7 : exemple de variations d'emploi des mots entre 2014-15 et 2016-17. A gauche le nombre d'occurrence par année des 11 mots cités plus de 25 fois dont l'usage a le plus baissé ; à droite celles des 12 pour lequel il a le plus augmenté.

En affinant cette analyse au niveau de chaque résumé de thèse, on peut regarder les fréquences de cooccurrences des mots pour en proposer des regroupements en utilisant des méthodes de classification hiérarchique (CH). On peut ainsi regrouper ceux qui se retrouvent le plus souvent ensemble dans certains résumés et peu dans les autres.

Un noyau dur centré sur la pédagogie et les personnes

Les arborescences présentées en figures 8 et 9 permettent de comparer la formation de tels groupes de mots. Ainsi, la figure 8 -extraite de l'article VdR1- rappelle le résultat d'une CH des mots employés dans les titres et résumés de thèses de 2014-2015. Quatre groupes avaient été repérés associant chacun des mots fréquemment employés ensemble et peu mélangés avec ceux des trois autres : ces quatre univers lexicaux globalement disjoints les uns des autres, correspondaient à autant de familles de préoccupations ou de centres d'intérêt. On y trouvait un lexique (classe 3) lié à la gestion, aux technologies, à l'entreprise et aux transferts de compétences, un autre (classe 2) lié aux apprentissages, aux questions pédagogiques et particulièrement linguistiques, un troisième (classe 4) lié à la formation continue des enseignants (car nous n'avions alors pas encore décidé de filtrer spécifiquement cette thématique) et enfin un dernier (classe 1) lié à la construction des personnes, aux dynamiques identitaires et à la professionnalisation.

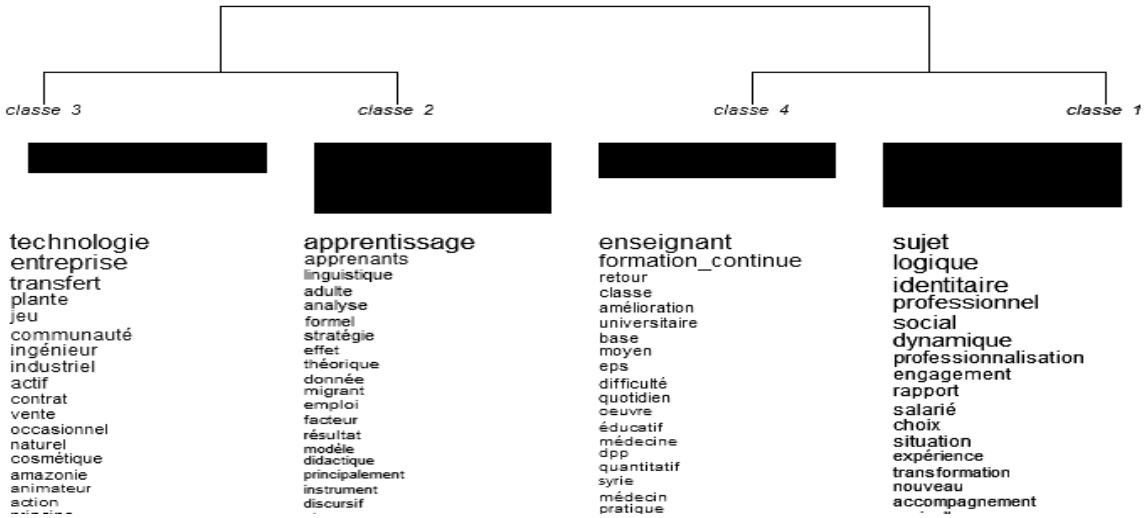



Figure 8 : classification des mots fréquemment employés ensembles
dans les thèses soutenues en 2014 ou 2015 (titre et résumés) en 4 univers lexicaux

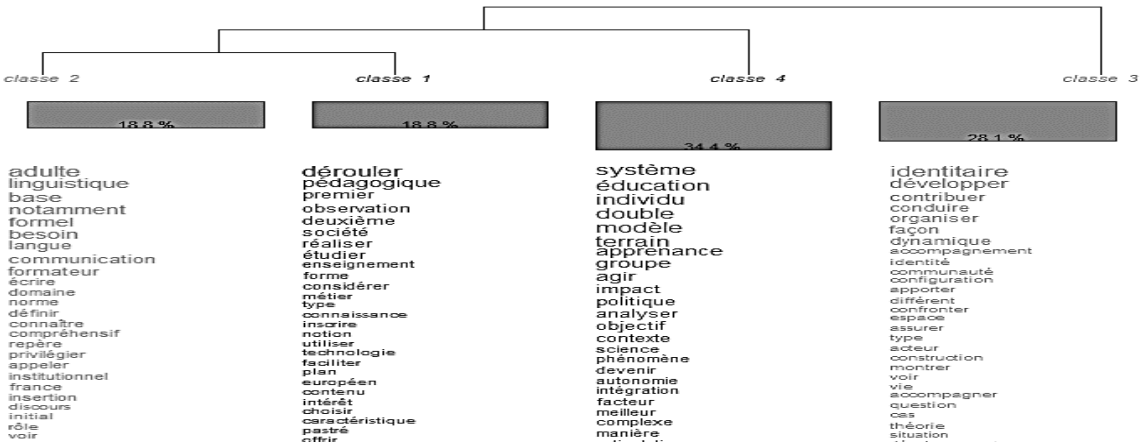



Figure 9 : classification des mots fréquemment employés ensembles
dans les thèses soutenues en 2016 ou 2017 (titre et résumés) en 4 univers lexicaux

La figure 9 présente le même type de classification pour les thèses retenues pour 2016-2017. On y retrouve comme en 2014-2015 deux classes consacrées respectivement aux aspects linguistiques des formations de base des adultes (classe 2) puis aux dynamiques identitaires (classe 3). On observe aussi fort logiquement que la classe « enseignant et formation continue[8] » (classe 4 en 2014-15) a -en tant que telle- disparu à la suite de nos filtrages spécifiques (cf. plus haut) ; pour autant ont émergé deux classes s'intéressant pour l'une aux déroulés pédagogiques (classe 1) et l'autre à l'individu dans le système de l'éducation (classe 4) : on peut penser que les mots qui les composent avaient été pour 2014-2015 amalgamés tous ensemble au lexique de la formation continue des enseignants mais que le filtrage les en a émancipés, tout en les fragmentant en deux sous-classes.

Une confirmation de la sous-représentation de la perspective macroscopique

Une autre différence entre les deux périodes est le fait que la classe 3 de 2014-2015 (« technologie », « entreprise », « transfert », « plante »…) n'a pas d'équivalent en 2016-2017, ce qui implique que la classification ne donne plus à voir de classe liée aux entreprises.

Un retour à l'analyse du corpus de 2014-2015 fait comprendre que c'est une thèse très particulière (Gagneaux, 2015) qui faisait -à elle seule- émerger cette classe 3 spécifique au « transfert » de « technologie » : explicitement notée comme soutenue dans la sous-discipline « formation des adultes » dans le Sudoc, cette thèse porte sur l' « analyse de l'activité et culture d'action des professionnels de la bioproduction de substances actives pharmacologiques, cosmétiques en système Plantes à Traire® pour la conception de formations sous l'hypothèse de l'enaction ». En l'occurrence la présence de cette thèse dans le « noyau dur » n'a rien de surprenant[9] puisqu'elle se situe en fait clairement au croisement de l'anthropologie des savoirs / communication interculturelle, de l'analyse de l'activité et de la gestion de connaissances.

De fait, *a contrario,* c'est plutôt l'absence d'autres thèses consacrées au monde économique dans ce deux corpus 2014-2015 et 2016-2017 qui peut attirer l'attention : la perspective macroscopique sur la formation des adultes (économie, sociohistoire, réforme) n'apparait que de manière très marginale, à

[8] À noter que, dans l'article VdR1, avait été regroupée -dans le corpus même- la locution « formation_continue » sous la forme d'un seul mot composé afin de renforcer le filtrage des faux positifs portant en fait sur la formation initiale.
[9] Néanmoins, elle illustre la difficulté du périmétrage de ce qui peut être considéré comme recherche en formation des adultes en matière de travaux de didactique pro ou d'analyse de l'activité dont une grande partie est thématisée par l'activité analysée (ici la culture des plantes à traire en Amazonie) : A partir de quel volume de lexique « formation des adultes » dans le résumé peut-on considérer que la « teinture » est suffisante ?

l'intérieur de la classe 4 de 2016-2017 c'est à dire vis-à-vis du « système » de « l'éducation ». Plus généralement, l'économie de la formation des adultes (ou la relation avec le monde économique) semble presque totalement absente des extractions des deux périodes. Cela avait déjà été observé dans l'article précédent (VdR4) portant sur la comparaison entre les thèses vues par Sudoc/Rameau et les archives ouvertes HAL : il avait été remarqué que c'est « la classe « didactique », « professionnalité » [qui] est surtout alimentée par les thèses du Sudoc ».

De fait, une simple requête supplémentaire recherchant les thèses contenant « économie de la formation » et « formation professionnelle » fait apparaitre cinq thèses non identifiées précédemment (1 en 2017 et 4 en 2014-2015), parlant bien de l'économie de la formation des adultes et qui ont été réintégrées[10] *a posteriori* dans les figures 3, 4 et 5. Cela confirme bien que de telles thèses existent, mais qu'elles ne sont pas spontanément référées dans le Sudoc à ce qui devrait être leur catégorie naturelle à savoir « formation des adultes », « formation professionnelle continue » ou « éducation des adultes ».

À nouveau, on retrouve le fait que dans Sudoc -ou plutôt dans son langage d'indexation Rameau- ces termes sont plus pensés comme décrivant la « promotion sociale » de certains individus, allant donc de pair avec problèmes linguistiques et dynamiques identitaires que comme des questions socio-économiques ou politiques, relevant d'une perspective macroscopique. On n'est donc pas loin ici d'un constat tautologique : l'analyse lexicale ne fait que montrer que ce que l'on injecte dans le corpus. Il ne peut pas y avoir d'« économie de la formation » si l'on n'introduit pas explicitement ce terme en requête : en effet cette thématique n'est pas spontanément perçue comme rentrant dans ce qui devrait être indexé par les bibliothécaires des thèses comme « éducation des adultes » dans Rameau.

## 6. Discussion et conclusion, pistes pour l'avenir

Comme les quatre précédents, cet article se lit à deux niveaux.

D'une part, en ce qui concerne le noyau dur des thèses récentes, on observe peu de différences entre 2014-2015 et 2016-2017 qui traitent globalement des mêmes thèmes. En revanche on visualise bien leurs spécificités communes par rapport aux thèses de tous les autres champs : elles sont plus tardives d'une quinzaine d'années et sont certes aux deux tiers rattachées aux sciences de l'éducation, mais associent dans leurs jurys de multiples expertises issues d'autres disciplines. D'autre part, en matière d'extraction des thèses et de constitution du corpus, cet article est l'occasion de rappeler l'intérêt qu'il y a à dépasser le simple noyau dur Rameau/Sudoc déjà étiqueté comme éducation ou formation des adultes et de fait orienté « promotion sociale ».

Les articles VdR3 et VdR4 ont montré que cela pouvait être entrepris par agrégation successive de requêtes, sous réserve de partir *in extenso* d'un sommaire encyclopédique de référence, comme cela a été fait avec celui du traité de Carré et Caspar (voir encadré 1 et article VdR3), mais cette méthode se révèle très fastidieuse. En fait on peut imaginer deux autres méthodes : l'une est de travailler à partir de l'intitulé de l'équipe de rattachement du doctorant et du directeur. En parallèle une nouvelle voie peut être proposée dans une logique dite de *machine learning* en cherchant non plus par des requêtes classiques focalisées sur quelques mots ou quelques auteurs mais en cherchant globalement une similitude des classes de mots présentes dans l'ensemble du résumé par rapport à celles d'un échantillon de résumés servant de référence. C'est ce qui sera développé dans la suite de cette rubrique.

[10] En fait les graphes concernant les disciplines ont été produits avec un corpus de 87 thèses incluant ces 4 nouvelles thèses ce qui explique l'importance des jurés issus de droit et sciences économiques.

**Références bibliographiques**


Beillerot J. (1993). *Les thèses en Sciences de l'Education : bilan de 20 années d'une discipline,* Université Paris X, 93 pages.

Beillerot J. et Demori F. (1997). *Les thèses en Sciences de l'Education de 1990 à 1994,* Université Paris X, 49 pages.

Carré P. et Caspar P. (2011). *Traité des sciences et techniques de la formation.* (Troisième édition). Paris : Dunod.

Gagneaux M.-H. (2015). *Analyse de l'activité et culture d'action des professionnels de la bioproduction de substances actives pharmacologiques, cosmétiques en système Plantes à Traire® pour la conception de formations sous l'hypothèse de l'enaction.* Thèse d'université. Paris : CNAM. http://www.theses.fr/2015CNAM0986

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

Leclercq, V. (2007). *Docteurs et doctorants en Sciences de l'Education : les trajectoires professionnelles et préoccupations scientifiques, étude d'une population de 2001-2006,* AECSE, Commission Cursus Publics - Sept. 2007, 31 pages.

Leclercq, V. (2008). Docteurs et doctorants en Sciences de l'Education : entre trajectoires professionnelles et préoccupations scientifiques, *Recherches & Educations* n°1/2008, p. 27-45. En ligne à https://rechercheseducations.revues.org/437

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. Accessible via https://docplayer.fr/10422759-Iramuteq-implementation-de-la-methode-alceste-d-analyse-de-texte-dans-un-logiciel-libre.html

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

| Grande Discipline | Groupe | Libellé sous-groupe | Section | Titre de la section du CNU |
|---|---|---|---|---|
| DROIT | 01 | Droit et science politique | 01 | Droit privé et sciences criminelles |
| | | | 02 | Droit public |
| | | | 03 | Histoire du droit et des institutions |
| | | | 04 | Science politique |
| | 02 | Sciences économiques et de gestion | 05 | Sciences économiques |
| | | | 06 | Sciences de gestion |
| LETTRES | 03 | Langues et littératures | 07 | Sciences du langage : linguistique et phonétique générales |
| | | | 08 | Langues et littératures anciennes |
| | | | 09 | Langue et littérature françaises |
| | | | 10 | Littératures comparées |
| | | | 11 | Langues et littératures anglaises et anglo-saxonnes |
| | | | 12 | Langues et littératures germaniques et scandinaves |
| | | | 13 | Langues et littératures slaves |
| | | | 14 | Langues et littératures romanes : espagnol, italien, portugais, autres langues romanes |
| | | | 15 | Langues et littératures arabes, chinoises, japonaises, hébraïques, d'autres domaines linguistiques |
| | 04 | Sciences humaines | 16 | Psychologie, psychologie clinique, psychologie sociale |
| | | | 17 | Philosophie |
| | | | 18 | Architecture (ses théories et ses pratiques) arts appliqués, arts plastiques, arts du spectacle, épistémologie des enseignements artistiques, esthétique, musicologie, musique, sciences de l'art |
| | | | 19 | Sociologie, démographie |
| | | | 20 | Anthropologie biologique, ethnologie, préhistoire |
| | | | 21 | Histoire et civilisations : histoire et archéologie des mondes anciens et des mondes médiévaux; de l'art |
| | | | 22 | Histoire et civilisations : histoire des mondes modernes, histoire du monde contemporain ; de l'art; de la musique |
| | | | 23 | Géographie physique, humaine, économique et régionale |
| | | | 24 | Aménagement de l'espace, urbanisme |
| | 12 | Groupe interdisciplinaire | 70 | Sciences de l'éducation |
| | | | 71 | Sciences de l'information et de la communication |
| | | | 72 | Epistémologie, histoire des sciences et des techniques |
| | | | 73 | Cultures et langues régionales |
| | | | 74 | Sciences et techniques des activités physiques et sportives |
| | Théologie | | 76 | Théologie catholique |
| | | | 77 | Théologie protestante |
| SCIENCES | 05 | Mathématiques et informatique | 25 | Mathématiques |
| | | | 26 | Mathématiques appliquées et applications des mathématiques |
| | | | 27 | Informatique |
| | 06 | Physique | 28 | Milieux denses et matériaux |
| | | | 29 | Constituants élémentaires |
| | | | 30 | Milieux dilués et optique |
| | 07 | Chimie | 31 | Chimie théorique, physique, analytique |
| | | | 32 | Chimie organique, minérale, industrielle |
| | | | 33 | Chimie des matériaux |
| | 08 | Sciences de la terre | 34 | Astronomie, astrophysique |
| | | | 35 | Structure et évolution de la Terre et des autres planètes |
| | | | 36 | Terre solide : géodynamique des enveloppes supérieures, paléo-biosphère |
| | | | 37 | Météorologie, océanographie physique et physique de l'environnement |
| | 09 | Mécanique, génie mécanique, génie informatique, énergétique | 60 | Mécanique, génie mécanique, génie civil |
| | | | 61 | Génie informatique, automatique et traitement du signal |
| | | | 62 | Energétique, génie des procédés |
| | | | 63 | Génie électrique, électronique, photonique et systèmes |
| | 10 | Biologie et biochimie | 64 | Biochimie et biologie moléculaire |
| | | | 65 | Biologie cellulaire |
| | | | 66 | Physiologie |
| | | | 67 | Biologie des populations et écologie |
| | | | 68 | Biologie des organismes |
| | | | 69 | Neurosciences |
| PHARMACIE | 11 | Pharmacie | 80 | Sciences physico-chimiques et ingéniérie appliquée à la santé |
| | | | 81 | Sciences du médicament et des autres produits de santé |
| | | | 82 | Sciences biologiques, fondamentales et clinique |
| | | | 85 | Sciences physico-chimiques et ingéniérie appliquée à la santé |
| | | | 86 | Sciences du médicament et des autres produits de santé |
| | | | 87 | Sciences biologiques, fondamentales et clinique |

Tableau Annexe 1 : regroupement CNU des disciplines en grandes disciplines et en groupes
Source CNU : http://cache.media.enseignementsup-recherche.gouv.fr/file/statistiques/09/1/70_938091.pdf