Rubrique « Vie de la Recherche »
commune aux Revues *Savoirs* et *TransFormations*

Olivier Las Vergnas, CIREL-Trigone (EA 4354) et CREF-AFA (EA 1589)

N°1 - Méthode de repérage des thèses soutenues en 2014 et 2015 liées à la formation des adultes

Mai - 2016

Résumé :

Ce court article a pour objet d'annoncer la reprise d'une rubrique « Vie de la recherche » en formation des adultes[1]. Il est consacré à la méthode qui peut être utilisée pour tenter de donner une vision compréhensive de la production doctorale dans ce champ. En effet, le champ de la formation des adultes est l'objet de multiples thèses. Dans une rubrique telle que celle-ci, consacrée à la vie de la recherche, en donner annuellement ou biennalement une vision d'ensemble apparaît comme une nécessité. Or, justement, on peut supposer que le développement des bases de données documentaires et des big data en facilite aujourd'hui la production. On peut penser aussi que la création d'une requête bibliographique standard pour les thèses pourrait ensuite être transposée à l'extraction d'autres types de publications scientifiques (ouvrages, articles ACL). Le propos du présent article est donc de voir comment pourrait être produite systématiquement une synthèse.

Mots clefs : Thèses de doctorat, bibliométrie, formation des adultes

[1] En accord entre les rédactions des deux revues, cette rubrique sera commune à la revue « Savoirs » et à la revue « TransFormations ».

1. L'observation de la production doctorale et les bases de données bibliographiques

Le champ de la formation des adultes est l'objet de multiples thèses. Dans une rubrique telle que celle-ci, consacrée à la vie de la recherche, en donner annuellement ou biennalement une vision d'ensemble apparait comme une nécessité. Or, justement on peut supposer que le développement des bases de données documentaires et de l'approche des *big data* en facilite aujourd'hui la production. On peut penser aussi que la création d'une requête bibliographique standard pour les thèses pourrait ensuite être transposée à l'extraction d'autres types de publications scientifiques (ouvrages, articles ACL).

Le propos du présent article est donc de voir comment pourrait être produite systématiquement une synthèse analogue à celles proposées pour les sciences de l'éducation en général par Beillerot (1993), Beillerot & Demori (1997), Feyfant (2005) ou Leclerc (2008) et, spécifiquement pour la formation des adultes, par Françoise Laot à partir de 2006.

2. Sudoc et l'extraction des thèses consacrées à la formation des adultes

De fait, pour la France, on trouve des informations sur nombre de ces thèses dans plusieurs bases de données multidisciplinaires ou spécialisées[2] comme celle de l'Institut français d'éducation (Ife) à http://ife.ens-lyon.fr/vst/Recherches/AccueilTheses.php. Parmi ces bases, le catalogue du Système Universitaire de Documentation (Sudoc) est, comme cela est rappelé sur son site à l'adresse http://www.Sudoc.abes.fr/, « *le catalogue collectif français réalisé par les bibliothèques et centres de documentation de l'enseignement supérieur et de la recherche. Il comprend plus de 10 millions de notices bibliographiques qui décrivent tous les types de documents [… et] il a pour mission de recenser l'ensemble des thèses produites en France* ».

Concrètement, faute d'un système de dépôt automatisé, le Sudoc ne peut vraiment prétendre à l'exhaustivité, mais il constitue tout de même le plus complet des catalogues de thèses soutenues en France ; il peut donc logiquement être utilisé[3] pour tenter de systématiser la production d'un état périodique des thèses dans un champ thématique précis, comme cela a été le cas en sciences de l'éducation avec le travail de Macarie-Florea, Rodriguez & Serbanescu-Lestrade pour l'AREF (2010).

3. L'interrogation par une requête sur les « mots sujets », complétée par un examen heuristique

Reste donc à préciser quelle peut être la requête à utiliser pour extraire les thèses concernant la « formation des adultes ». Or, le Sudoc s'appuie sur le vocabulaire contrôlé du langage d'indexation Rameau[4] pour le remplissage du champ « mots sujet ». Celui-ci prévoyant l'usage des deux termes « éducation des adultes » et « éducation permanente » pour qualifier les documents au sein de la rubrique éducation et enseignement (370) on pourrait penser qu'il suffit de les utiliser comme requête pour extraire toutes les thèses qui nous intéressent.

Avec Anne Dourlens, documentaliste référente pour les Sciences de l'éducation à la BU de l'Université de Lille 1, nous avons donc mis en œuvre cette requête que nous appellerons R1 dans la suite ; les données 2016 étant -à la date de rédaction- beaucoup trop parcellaires dans Sudoc, nous avons choisi de nous limiter à la période biennale de soutenance de 2014 et 2015.

Ont été ainsi extraites 87 notices de documents dont 63 se révèlent être en fait non pas des thèses de doctorat d'université mais des « thèses d'exercice » propre au secteur médico-social (46 en médecine,

[22] Des mentions des soutenances sont aussi reprises dans des sites comme ceux de l'association des enseignants chercheurs en sciences de l'éducation (Aecse) ou de certains laboratoires.
[3] Signalons aussi son interface dédiée « thèses.fr » (http://theses.fr) qui donne un accès plus direct aux fiches des documents
[4] Répertoire d'autorité-matière encyclopédique et alphabétique unifié, cf http://guiderameau.bnf.fr/html/rameau_0642.html#d11e54430

10 en pharmacie, 5 en chirurgie dentaire et 2 qui sont des « mémoires de sage-femme ») dont une grande majorité se contente d'évoquer incidemment la formation continue des professionnels.

Voilà donc qui devrait nous conduire finalement à un corpus à 24 thèses d'université en formation des adultes repérées en tant que telles par la requête Rameau sur les mots sujet.

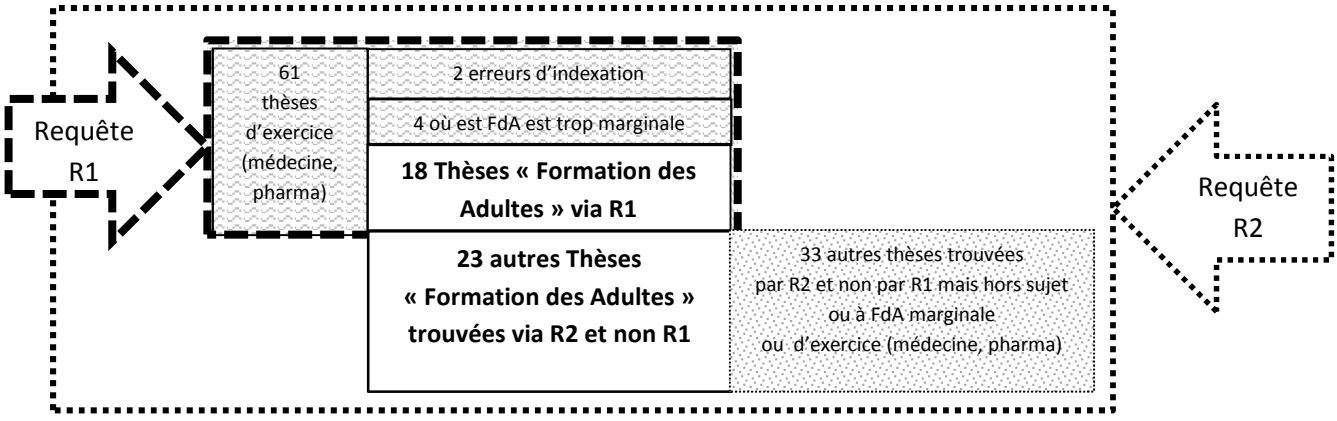
Mais nous avons dû apporter deux correctifs. D'une part il nous a fallu réduire ce corpus de 24 à 18 thèses car 6 d'entre elles ne pouvaient être conservées (2 étant des faux-positifs liés à un problème d'indexation[5] et les 4 autres n'étant -au vu du résumé- que de manière très marginale liées à la formation des adultes).

D'autre part, il était aussi logique de chercher à confirmer la pertinence de cette extraction en la comparant à celles que donneraient d'autres requêtes qui devraient être convergentes, de façon à chercher à mettre en évidence cette fois des « faux négatifs ». Pour cela nous avons décidé de tester une seconde requête[6] (R2) dans le Sudoc portant sur « tous les mots » (i.e. dans le texte de la notice ou dans le texte intégral du résumé) et non pas seulement les « mots sujet ». Nous avons ainsi cherché cette fois les documents contenant dans « tous les mots » l'un des trois termes « éducation des adultes », « éducation permanente » ou « formation continue ».

On obtient alors une autre liste de 153 thèses qui contient, en plus de l'intégralité des 87 thèses correspondant à la requête R1, 66 autres thèses dont à nouveau, bon nombre se révèlent être concernées seulement marginalement par la formation continue et sont des thèses d'exercice. Néanmoins, après un filtrage heuristique effectué par un examen du contenu du résumé, nous avons extrait parmi ces 66 une liste complémentaire de 23 autres thèses d'université non précédemment identifiées par la requête R1 alors que pourtant leurs questions de recherche intéressent significativement la formation des adultes.

4. Un corpus Sudoc de 41 thèses d'université indexées en formation des adultes (2014-2015).

En les fusionnant avec les 18 déjà sélectionnées grâce à la requête R1, il ressort au total un corpus de 18+23 = 41 thèses à prendre en compte qui ont été soutenues en France en 2014 ou 2015 et dont on peut dire qu'elles sont indexées dans Sudoc comme traitant de manière significative de formation des adultes. La figure 1 rend compte de ce travail.



[5] Il s'agit en l'occurrence de 2 thèses consacrées l'une à « la thématique de l'amour dans les blogs et journaux collégiens et lycéens » et l'autre à « l'usage de la littérature de jeunesse dans l'éducation des filles au XIXe siècle ».

[6] On pourrait aussi envisager deux autres méthodes complémentaires, non mobilisées ici : l'une (M3) serait de procéder par itération en recherchant (par un logiciel d'analyse lexicale, comme Iramuteq, cf plus loin) les mots à utiliser pour définir parmi ceux qui justement caractérisent un premier noyau de thèses sélectionnées : M4 inversement procéderait par exclusion de termes, en cherchant celles qui contiennent par exemple « formation » et « apprentissage » mais pas « élèves » et « scolaires ».

Figure 1 : résultat des requêtes de recherche des thèses R1 et R2 sur la formation des adultes en 2014-2015

Le fait que ces 23 thèses supplémentaires aient été des « faux négatifs », c'est à dire qu'elles n'aient pas été identifiées par R1 peut s'expliquer par le fait qu'elles ont été indexées trop finement, i.e. avec des mots clefs plus précis que les vedettes matières que sont « éducation des adultes » et « éducation permanente ». Et comme le langage Rameau ne fonctionne pas comme un thésaurus hiérarchique, une telle indexation trop fine n'est pas rattachée dans la notice à des grandes familles plus générales : il en résulte que si l'on ne cherche que sur ces vedettes matières (comme c'est le cas avec R1) on ne trouve pas trace de ces thèses indexées sur des mots plus fins.

Cependant une fois ce constat posé, la question des faux positifs n'est pas complètement explorée. Rien ne prouve en effet que nous ayons récupéré par R2 la totalité des thèses qui peuvent être considérées comme concernent la formation des adultes. Pour examiner cette question, il nous faut procéder à l'envers en partant d'une liste de thèses soutenues dans un laboratoire de formation des adultes et voir dans quelle mesure nous les avons retrouvées ou non par nos requetes R1 ou R2. Ce ce que nous avons fait en prenant comme point de départ la liste exhaustive des 14 thèses soutenues en 2014 et 2015 au sein des équipes CIREL-Trigone (6 thèses) à Lille et CREF-Apprenance et Formation des Adultes à Paris-Ouest-Nanterre (8 thèses).

Apres avoir constaté que seules 3 d'entre elles avaient bien été trouvées par la requête R1 ou R2 nous avons regardé si nous devions considérer les 11 autres comme des faux négatifs. Le tableau T1 résume ce travail. On observe d'abord que 2 ne sont pas présentes ou pas indiquées comme « soutenues » dans Sudoc et que 2 autres ont visiblement été indexées trop finement (« infirmières – formation » ou « ingénieurs -- formation ») confirmant le problème indiqué plus haut. Restent enfin les 7 autres, qui renvoient à la question du périmètre que nous voulons prendre compte : en effet, 2 s'intéressent à la formation des enseignants ou formateurs (donc de fait des adultes), 2 autres à la transition de jeunes à adultes et enfin 3 l'ingénierie multimédia (que l'on peut considérer comme FTLV). Le tableau T1 indique également les mots clefs qui auraient permis d'inclure (ou non) ces types de thèses dans notre requête.

Au-delà des difficultés liées à l'indexation par le langage Rameau, cette situation ne doit pas nous étonner. Le champ de la « formation des adultes » est loin d'être défini de manière univoque et ne constitue pas non plus une catégorie universitaire reconnue : nous ne pouvons donc pas éluder notre responsabilité d'en choisir une définition pour cette rubrique. Le corpus obtenu par R1+R1 correspond à une définition « bibliométrique » limitée à la présence de mots clefs ; son élargissement par une recherche par laboratoire oblige à interroger notre vision du périmètre que nous jugeons intéressant.

N°	Nom du doctorant	Equipe d'accueil	Sudoc	Theses.fr	trouvé par R1	trouvé par R2(153)	retenu dans R2final	retenus R1ouR2 finaux	Est-ce un FAUX NEGATIF?	Vedette Matière Rameau	autres mots clefs discri-minants (1)	autre mot clef (2)
1	**Alshami**	CIREL-Trigone	Oui	Oui	Non	Non	Non	**NON**	OUI si l'on veut inclure la formation des enseignants ou formateurs	Enseignants--Formation		
2	**Amblard**	CIREL-Trigone	Oui	Oui	Oui	Oui	Oui	**OUI**	NON, trouvé par R1 et R2	***Formateurs (éducation des adultes)***		
3	**Bellegarde**	CIREL-Trigone	Oui	Oui	Oui	Oui	Oui	**OUI**	NON, trouvé par R1 et R2	***Ecriture (éducation des adultes)***		
4	**Danquigny**	CIREL-Trigone	Oui	Oui	Non	Non	Non	**NON**	si l'on veut inclure les dispositifs multimedia	Universités virtuelles	Dispositifs ouverts	
5	**Haidar**	CREF-Apprenance	Oui	Oui	Non	Oui	Non	**NON**	si l'on veut inclure la transition vers adultes	Transition jeunes / adultes		
6	**Lawinski**	CIREL-Trigone	Oui	Oui	Non	Non	Non	**NON**	OUI si l'on veut inclure la formation des enseignants ou formateurs	Apprentissage professionnel	Formation de formateurs	CFA
7	**Lemaire**	CREF-Apprenance	Non	Oui	Non	Non	Non	**NON**	**OUI mais pas noté comme "soutenue" dans SUDOC**	Formation professionnele continue	Apprenance	
8	**Leroux**	CREF-Apprenance	Oui	Oui	Non	Non	Non	**NON**	si l'on veut inclure la transition vers adultes	Apprentissage	Parcours professionnel	VIE
9	**Muller**	CREF-Apprenance	Oui	Oui	Non	Non	Non	**NON**	**OUI (indexé trop fin)**	Infirmières--Formation		
10	**Nucci**	CREF-Apprenance	Oui	Oui	Non	Non	Non	**NON**	si l'on veut inclure les dispositifs multimedia	Enseignants--Formation	Cours en ligne ouverts à tous	
11	**Raujol**	CREF-Apprenance	Non	Non	Non	Non	Non	**NON**	**OUI mais pas présent dans SUDOC**	Apprentissage professionnel	Formation de formateurs	Agricole
12	**Striff**	CREF-Apprenance	Oui	Oui	Non	Non	Non	**NON**	**OUI (indexé trop fin)**	Ingérieurs--Formation		
13	**Tarrit**	CIREL-Trigone	Oui	Oui	Non	Non	Non	**NON**	si l'on veut inclure les dispositifs multimedia	Universités virtuelles	Dispositifs ouverts	
14	**Yennek**	CREF-Apprenance	Oui	Oui	Oui	Oui	Oui	**OUI**	NON, trouvé par R1 et R2	***Formateurs (éducation des adultes)***		

Tableau 1 : analyse des 14 thèses soutenue en 14-15 au sein de CIREL Trigone ou CREF-apprenance et formation des adultes

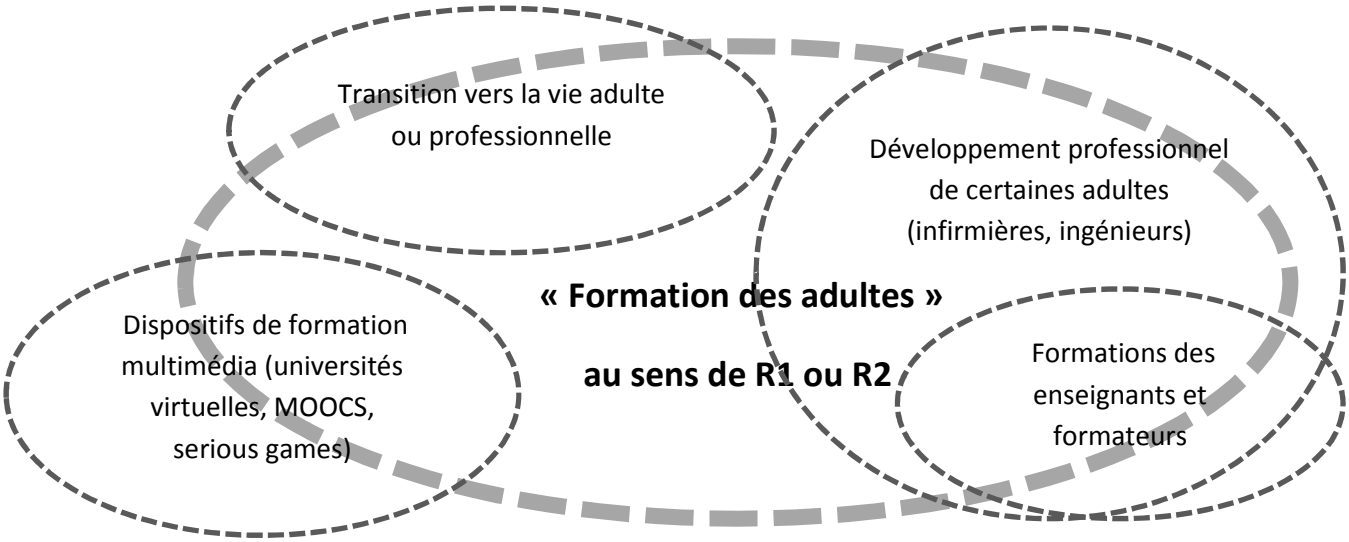


Figure 2: visualisation des champs pouvant être pris en compte à la suite de l'analyse du tableau de faux négatifs.

Concrètement, si l'on s'intéresse -pour l'avenir de cette rubrique- à automatiser une requête que l'on pourra utiliser régulièrement dans cette rubrique pour interroger les documents, on pourra reproduire sans difficulté dans les requêtes R1 et R2. On pourra aussi les filtrer en leur ajoutant la condition : et non note de thèse = « exercice » et non note de thèse = « sage femme ». Cela éliminera les thèses d'exercice de médecine et de pharmacie ainsi que les rapports de sage-femme. En revanche, le filtrage heuristique des faux positifs devra se faire à la main. Et quant au traitement des faux négatifs, il devra être décidé du périmètre à retenir. S'agira-t-il de rendre compte, selon ce qui est illustré en figure 2 :

- Strictement des thèses indexées via Rameau comme « formation des adultes » (R1)
- Plus largement de celles qui utilisent explicitement les vocables de R2
- Ou d'y ajouter explicitement la formation des enseignants, la transition vers la vie adulte et l'ingénierie multimédia ?

5. De premiers éléments de synthèse sur les 41 thèses repérées

Pour l'instant, en en restant à R1+R2, plusieurs comptages peuvent être effectués à partir du premier corpus 2014-2015 de 41 thèses. On peut par exemple s'intéresser aux disciplines de dépôt : 31 d'entre elles ont été déposées en sciences de l'éducation (dont 7 -sur les 11 issues du CNAM- étrangement notées directement en « formation des adultes » dans theses.fr alors que cette expression ne correspond pas à une discipline CNU), 3 en sociologie, 2 en gestion, 1 en psychologie et 4 déposées dans des disciplines liées aux langues (sciences du langage, didactique des langues ou linguistique). Le tableau complet des thèses tel qui est en ligne à http://enviedesavoir.org/data/VdR/T1.pdf et l'espace associé dans le réseau Zotero permettront à chacun d'explorer cette production doctorale en fonction de différentes variables.

6. Un premier exemple de méthode cartographie lexicale

Bien sûr, l'analyse de ces thèses, comme celle d'un autre corpus de documents significatifs de la vie de la recherche, peut aller plus loin que de tels comptages.

Pour donner une idée du travail qui pourra être fait dans les parutions suivantes de cette rubrique nous donnons ici à titre indicatif des extraits d'une cartographie lexicale de ces thèses utilisant le logiciel Iramuteq (ex Alceste, cf http://iramuteq.org Reinert, 1987 ; Ratinaud et Déjean 2009).

Le point de départ est un comptage des mots signifiants employés dans les titres et résumés du corpus de ces 41 thèses[7]. A partir de la comparaison des fréquences de ces mots issus des différentes thèses peut être définie une distance lexicale (et *a contrario* une proximité lexicale entre les thèses) : c'est la proximité ou l'éloignement de leurs profils de co-occurrentes des mots qui est utilisée comme mesure de leur distance. Cette distance pourra servir à étudier des groupes de mots signifiants, et à en proposer des regroupements en utilisant des méthodes de classification hiérarchique (CH) ou d'analyse factorielle des correspondances (AFC).

L'arborescence présentée en figure 3 donne ainsi le résultat d'une CH des mots employés dans les titres et résumés de nos 41 thèses. Dans cet exemple, quatre groupes de mots ont été repérés comme fonctionnant ensemble dans des segments de texte : ils regroupent chacun des mots fréquemment employés ensemble et peu mélangés avec ceux des trois autres groupes : ils peuvent aussi être plus ou moins caractéristiques de certaines thèses et non des autres. On peut donc lire cette arborescence comme montrant quatre univers lexicaux globalement disjoints les uns des autres, correspondant à des familles de préoccupations ou de centre d'intérêt.

Même si le corpus actuel des 41 thèses n'est pas encore -dans cette première contribution- parfaitement affiné, on peut, en toute première analyse, en regardant dans les résumés le contexte d'emploi de ces mots considérer que l'on a là quatre univers correspondent à :

- un lexique (classe 3) lié à la gestion, aux technologies, à l'entreprise et aux transferts de compétences,
- un lexique (classe 2) lié aux apprentissages, aux questions pédagogiques et particulièrement linguistiques,
- un lexique (classe 4) lié à la formation continue des enseignants (et donc la limite de notre champ, même si l'on peut considérer les enseignants comme des adultes comme les autres)
- un lexique (classe 1) lié à la construction des personnes, aux dynamiques identitaires et à la professionnalisation,

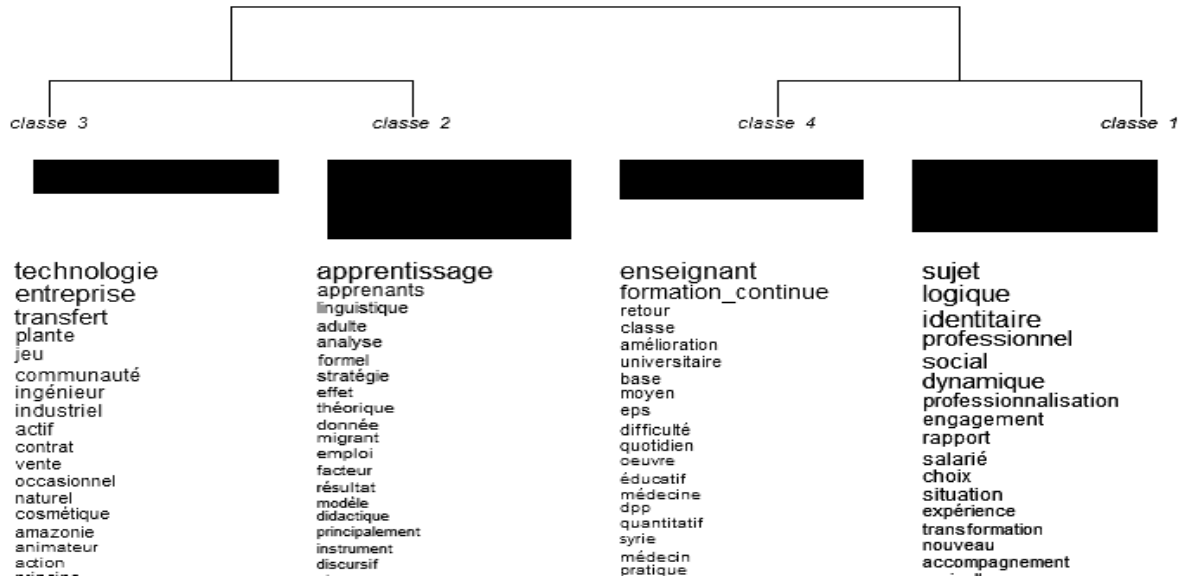


Figure 3 : classification des mots fréquemment employés ensembles dans les 41 thèses (titre et résumés) en 4 univers lexicaux

[7] Nous n'avons pas retenu leurs listes de mots clefs pour éviter les inférences entre contenu et indexation.

On peut aussi projeter ces classes de mots selon des axes définis par des analyses des correspondances (AFC). Ainsi, la figure 4 donne les positions des projections de ces mots sur le plan (1,2) d'une AFC effectuée sur ces 4 classes. Concrètement, sur le plan représenté dans la figure 2, les mots centraux sont ceux qui sont les plus communs à tous les contributeurs. A contrario, la distance au centre indique la spécificité de tel ou tel mot. Les axes 1 et 2 utilisés sont ceux qui maximisent la visibilité des spécificités des lexiques. Ont aussi été localisées des disciplines (autres que les sciences de l'éducation) dans lesquelles ces mots sont plus fréquents. Sur la page à http://enviedesavoir.org/data/VdR/T1.pdf cette figure est présentée en couleurs et commentée de manière plus détaillée. Y est ajoutée une figure 2bis permettant de voir à quels organismes et directeurs de thèse sont corrélées ces spécificités de vocabulaire : ces deux figures sont superposables, les axes utilisés et les échelles étant les mêmes.

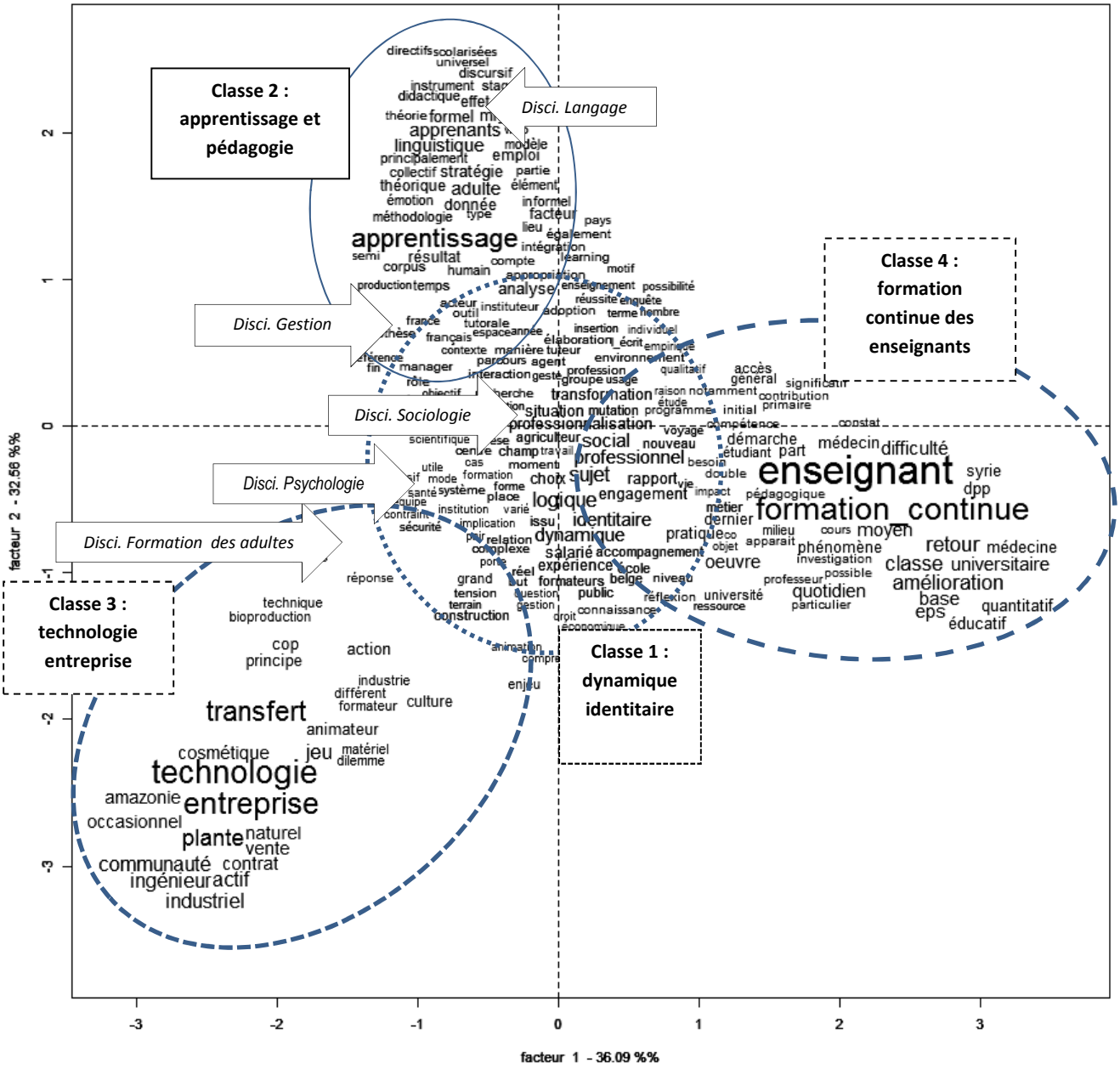


Figure 4 : projection des quatre univers lexicaux sur le 1er plan factoriel d'une AFC des résultats de la CH.

7. Un premier état des difficultés pour lancer une rubrique régulière

Comme annoncé, la présente contribution à cette rubrique « vie de la recherche » a surtout pour objectif de montrer les méthodes qui seront déployées de manière plus approfondie dans les prochains numéros. Dans la mesure où nous avons choisi une approche de bibliothéconomique et d'analyse lexicale, nous sommes à la fois tributaire des bases de données et des déposants, mais aussi de nos capacités à en proposer une valorisation compréhensive.

Nous espérons que le lecteur aura pu se faire une idée tant de la batterie de difficultés (non exhaustivité, absence de thésaurus, modélisation hasardeuse de la cartographie) que des solutions que nous allons progressivement en train de mettre en place. Nous lui donnons donc rendez-vous pour la prochaine contribution à cette rubrique pour en observer l'avancement.

Olivier Las Vergnas, CIREL-Trigone (EA 4354) et CREF-AFA (EA 1589)

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

Fayfant, Annie (2005) *Bilan des thèses concernant l'éducation Observatoire des thèses en éducation,* INRP/IFE http://ife.ens-lyon.fr/vst/DocDivers/Obs_theses_PDE_62.pdf

Laot Françoise F., « 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 .

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 CursusPublics-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

Macarie Florea, Roxana, Rodriguez, Daniela, Serbanescu-Lestrade, Karin. *Les caractéristiques des objets des theses en sciences de l'education.* Université de Geneve. Congres international AREF 2010, Sep 2010, Geneve, France. 10 p.halshs-00693539 https://halshs.archives-ouvertes.fr/halshs-00693539/document

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. Voir à http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation_mashs2009.pdf

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