Croisement de classifications issus de différents corpus

Pierre Wavresky[1]

[1]UMR Cesaer Agrosup Inrae Ubfc – pierre.wavresky@inrae.fr

Abstract

This textometric work on the productions of a sociology and economics laboratory was initiated by a request from its direction, in order to have a statistical description of the evolution of publications, presentations at conferences, research reports, books, etc., between 2005 and 2019.
In addition to the conclusions of this work in terms of analysis of the evolution of research themes, various methodological aspects will be presented. The first is the choice of the corpus treated: it is possible to focus on the title, keywords, abstract or full production (in the case of articles but hardly in the case of audio-visual documents). Clusterings will be carried out on these different corpora, which will be cross-classified to see their overlapping.
A second methodological aspect concerns the effect of missing values: if the title is always present, it is not the same for keywords, abstracts and obviously the full production.
A third methodological aspect is the treatment of the language (or languages, when there is a translation) in which the title, abstract, etc. is written, essentially French and English, and the fact that it is possible to make a global and comparative analysis of this multilingual corpus.

Keywords: bibliometry, clustering, textometry.

Résumé

Ce travail de textométrie sur les productions d'un laboratoire de sociologie et d'économie a été initié par une demande de sa direction, afin d'avoir une description statistique de l'évolution des publications, présentations à des congrès, rapports de recherches, livres, etc., entre 2005 et 2019. Hormis les conclusions de ce travail en termes d'analyse des évolutions de thématiques de recherche, seront présentés différents aspects méthodologiques. Le premier est le choix du corpus traité : il est possible de s'intéresser au titre, aux mots clés, au résumé ou à la production intégrale (dans le cas d'articles mais guère dans celui de documents audio-visuels). Des classifications seront effectuées sur ces différents corpus, classifications qui seront croisées pour voir leur recouvrement. Un deuxième aspect méthodologique concerne l'effet des valeurs manquantes : si le titre est toujours présent, il n'en est pas de même pour les mots clés, les résumés et évidemment la production intégrale. Un troisième aspect méthodologique est le traitement de la langue (ou des langues, quand il y a une traduction) dans laquelle est écrit le titre, le résumé, etc., essentiellement le français et l'anglais, et du fait de pouvoir faire une analyse globale et comparative de ce corpus multilingue.

Mots clés : bibliométrie, classification, textométrie.

1. Introduction

Dans le cadre de son évaluation par le Haut Conseil de l'Évaluation et de la Recherche et de l'Enseignement Supérieur (HCÉRES), la direction d'un laboratoire d'économie et de

sociologie a souhaité disposer d'une analyse textométrique des travaux menés entre 2015 et 2020, avec une comparaison avec les années antérieures, en vue de mettre en évidence, de façon statistique, l'évolution des thématiques de recherche sur la période récente.

Afin de mener un premier travail exploratoire, la liste des productions[1] de ce laboratoire entre les années 2005 et mi-2019, issue du catalogue Prodinra[2], a été analysée par les logiciels Iramuteq (Ratinaud, 2009) et R (R Development Core Team, 2005).

Sera dans un premier temps présenté le laboratoire de recherche et les caractéristiques de ses membres (rattachement disciplinaire).

Sera ensuite décrit le corpus, avec ses différentes caractéristiques (type de produit, titre, résumé, langue, auteurs…). Cette description fait ressortir différents aspects des productions du laboratoire : des années plus riches en nombres de production, deux axes de recherche utilisant des supports de publication différents…

Cette première analyse met en lumière certaines caractéristiques du corpus influençant l'analyse textuelle, comme l'absence de résumé ou de mot-clé pour certaines publications, et des publications qui ne sont pas toutes dans la même langue. Nous sommes donc en présence de plusieurs corpus pouvant être étudiés : le premier critère est la langue (c'est-à-dire l'anglais ou le français), le deuxième étant le type de texte (titre, mots-clés, résumé, corps de l'article). Dans ce papier, l'analyse textuelle (Lebart et Salem, 1994) concernera quatre corpus, celui des titres[3] et celui des résumés, dans les deux langues susmentionnées. De plus chaque entrée du corpus résumé peut être considérée comme un tout (l'entrée sera alors appelée « paragraphe ») ou découpée en un certain nombre de segments de texte, afin de tenir compte de différents types de discours pouvant être présents dans un même texte.

Il paraît donc intéressant, dans une dernière partie, après avoir analysé les résultats des différentes analyses textuelles, de croiser les différentes classifications qui en sont issues pour voir leur recouvrement : les mêmes discours sont-ils présents, que le corpus soit français ou anglais ? La classification sur les paragraphes est-elle similaire à celle sur les segments de texte ?

2. Le laboratoire de recherche

Le Centre d'Economie et de Sociologie Appliquées aux Espaces Ruraux (CESAER)[4] est une unité mixte de recherche regroupant des membres de l'Inrae et d'Agrosup Dijon. Il est constitué d'un collectif de sociologues, d'économistes, de géographes… Au sein de ce laboratoire ont été définis deux axes de recherche, Dynamique et Aménagement du Territoire (DAT) et Groupes Sociaux et Mondes Ruraux (GSMR), ce deuxième axe étant essentiellement constitué de sociologues.

[1] Le terme de *production*, ou de *produit*, est celui utilisé par Prodinra (cf infra). Un produit peut en effet être un document audiovisuel, une présentation orale à un congrès, un cours… Les travaux recensés ne se limitent donc pas aux articles, aux rapports de recherche ou aux chapitres d'ouvrage, ce qui explique l'utilisation de ce terme et non celui de *publication*.

[2] https://prodinra.inra.fr/?locale=fr

[3] Celui des titres sera abordé très succinctement, car les conclusions sont similaires à celui des résumés.

[4] https://www2.dijon.inrae.fr/cesaer/

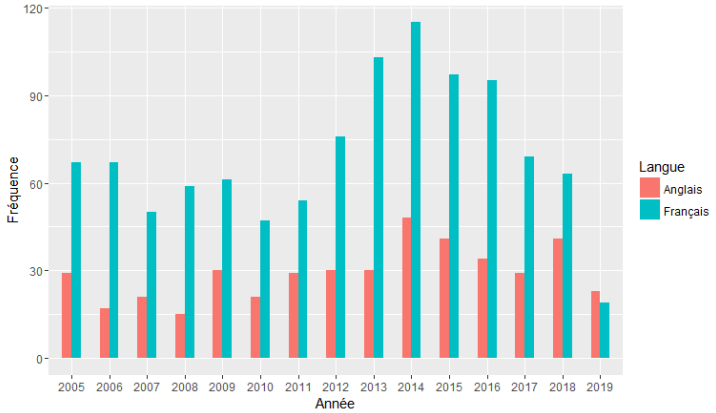
3. Les publications de 2005 à mi-2019

La période retenue pour l'analyse va de 2005 à 2019 (l'extraction finale des données à partir du site ProdInra date du 22 juillet 2019). Il y a eu 1500 produits, caractérisés par 179 critères, parmi lesquels ont été retenus les suivants : *l'identifiant* (pour pouvoir notamment croiser les classifications), le *titre*, la *langue du titre*, le *type de produit*, le *résumé*, la *langue de résumé*, les *auteurs*, le *public visé*, l'*année* de dépôt. La variable *auteurs* permettra de créer une variable *axe* à 3 modalités selon leur axe d'appartenance (DAT, GSMR, GSMR+DAT si les auteurs appartiennent à des axes différents ou si un auteur appartient aux deux axes). Il est à noter qu'en début de période, les axes DAT et GSMR n'existaient pas en tant que tels. De ce fait, des auteurs n'appartenant plus au laboratoire, parce qu'ils sont partis à la retraite par exemple, ont été classés dans un de ces deux axes selon leur champ de recherche ; toutefois certains n'ont été affectés à aucun des deux axes quand ce classement était difficile.

Au cours des 20 années étudiées, un peu plus de la moitié des productions (817) proviennent de l'axe DAT, ce qui peut s'expliquer par le nombre plus important de membres de cet axe (92 contre 36 pour l'axe GSMR, et 1 qui appartient aux deux axes[5]).

Les productions les plus courantes sont les *articles*, suivis de *papers* et des *chapitres d'ouvrage*. Mais cette répartition n'est pas homogène selon les deux axes, puisque les *livres* et *chapitres d'ouvrage* sont, comme on pouvait le penser a priori de supports privilégiés par les sociologues, surreprésentés au sein de l'axe GSMR, au contraire des *rapports de recherche* et des *thèses*.

Il y a eu une progression des publications, passant de 70 annuellement avant 2013 à100 après cette date[6]. Cette progression concerne tous les publics visés (grand public, pouvoirs publics, professionnels, scientifiques) avec une diversification entre 2012 et 2016. Ces années fastes touchent les publications dont le titre est en français ou en anglais[7], avec une progression de l'anglais, qui dépasse même le français en 2019 (mais l'année n'est pas complète).



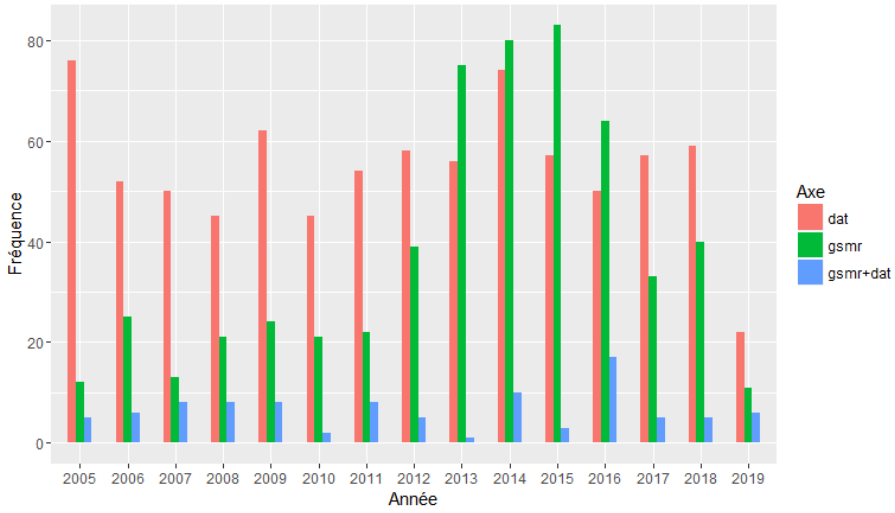
Graphique 1 : Nombre de produits par année selon la langue du titre (anglais, français)

La forte progression des publications aux alentours de 2014 semble particulièrement imputable à l'axe GSMR, comme l'indique le graphique 2.

[5] Ces comptages comprennent les membres actuels en anciens du laboratoire.

[6] L'année 2019 n'a pas été prise en compte dans le décompte car elle était incomplète.

[7] Il s'agit du titre principal, parfois il est traduit. D'autre part, certains titres, certes rares, sont en espagnol, en italien ou en portugais. Leur faible effectif explique que le graphique 1 ne comporte que le français et l'anglais.



Graphique 2 : Nombre de produits par année selon l'axe d'appartenance des auteurs

Il est possible d'étudier le corpus des titres, des résumés, des mots-clés. Mais il paraît difficile de mener une analyse textuelle sur les articles entiers, vu leur disponibilité limitée au sein du corpus Prodinra. Nous nous focaliserons ici sur les corpus les plus fournis, celui des titres (1500 produits) et celui des résumés (1153 produits)[8].

Il est à noter qu'on ne peut pas considérer les éléments du corpus résumé comme un sous-ensemble de ceux du corpus titre, et ce à cause de la langue dans laquelle sont écrits le titre et le résumé. Il existe en effet, par exemple, 21 produits ayant un titre uniquement en anglais et un résumé dans les deux langues : le corpus résumé français contiendra ces produits contrairement au corpus titre français. Il y a même 11 produits ayant un titre uniquement en anglais et un résumé uniquement en français. Mais ces occurrences sont rares, comme le montre le tableau suivant :

		Langue du résumé					
		(absent)	Anglais	Anglais-Français	Français	Langue autre	TOTAL
Langue du titre	Anglais	65	285	21	11	0	382
	Anglais-Français	8	8	178	121	0	315
	Français	274	4	64	457	3	802
	Langue autre	0	0	0	0	1	1
	TOTAL	347	297	263	589	4	1500

Tableau 1 : Croisement langue du titre et langue du résumé

D'après ce tableau, on voit que si on travaille sur les résumés, a priori plus riches que les simples titres, on perdra davantage de produits sur le corpus français que sur le corpus anglais. Ce qui explique au moins en partie un biais en défaveur de l'*axe* GSMR, pour lequel 34% des produits n'ont pas de résumé, contre 16% pour l'axe DAT. Les *pouvoirs publics* et le *grand public* seront également sous-représentés dans le corpus résumé, de même que les *années* fastes (2012 à 2015). Enfin les *papers* et *chapitres d'ouvrages* seront sous-représentés au bénéfice des *articles* et *ouvrages*.

[8] Le corpus mots-clés des auteurs comportent 1078 produits. Il y a en fait deux variables *mots-clés*, les « mots-clés des auteurs », qui ont l'inconvénient d'être écrits dans plusieurs langues, donc nécessitent comme les résumés de mener deux analyses textuelles, une pour l'anglais et l'autre pour le français, et les « mots-clés inra », qui ont l'avantage d'être en français mais sont très peu renseignés dans les années récentes.

4. Le corpus *résumé*

La comparaison des corpus français et anglais permettra de voir si les discours sont similaires dans les deux langues. Pour mener à bien cette comparaison, seront mis en regard dans un premier temps les lemmes les plus fréquents dans les deux langues, puis les deux classifications qui en sont issues, afin de répondre à une question formulée de deux façons différentes : les classes de discours ont-elles la même signification quelle que soit la langue et, d'une façon plus statistique, le croisement des classes permet-il de les faire coïncider ? A contrario, y a-t-il des trous, des discours présents dans un seul des deux corpus ?

Pour croiser les 2 classifications, il est nécessaire d'avoir un identifiant commun. Ce sera celui du produit ; il en découle que la classification se fera non sur des segments de texte d'une certaine longueur, mais à partir du résumé entier, avec l'inconvénient que ces résumés ne sont pas de même longueur.

Cette comparaison des corpus dans les deux langues fera l'objet d'un premier point.

Il sera intéressant également de comparer la classification issue d'observations « paragraphe » avec celle issue d'observations « segment de texte », afin de mesurer le pourcentage de recouvrement, de voir s'il y a des classes de paragraphe relativement pures, c'est-à-dire constituées de segments de texte appartenant à la même classe ou aux même classes.

Cette comparaison, menée sur le corpus français, fera l'objet du deuxième point.

4.1. Corpus résumés anglais et français

L'analyse des lemmes les plus fréquents dans les deux corpus souligne des différences de contenu de discours, du moins des différences dans les thèmes les plus abordés. Ainsi *model* est le lemme actif le plus courant dans le corpus anglais, ce qui est loin d'être le cas dans le corpus français, ce qui pourrait laisser penser que la modélisation (économétrique en l'occurrence) est davantage citée, donc utilisée dans le corpus anglais. Les articles publiés dans les revues internationales d'économie laissent en effet une grande part à l'économétrie. Toutefois, certains lemmes, comme *effet*, plus présent dans le corpus français que ne l'est *effect* dans le corpus anglais, peuvent modérer cette conclusion[9].

Il semble d'autre part que le thème agricole soit plus présent dans le corpus français (*agricole*, *agriculture*) que dans l'anglais (*agricultural*, *agriculture*). *Ouvrier*, *classe*, *travail*, termes à connotation sociologique, semblent également plus présents dans le corpus français, ce qui pourrait être imputable à la publication moins fréquente des travaux de sociologie en anglais, comme le montre le tableau 2.

4.1.1. Spécificité des axes de recherche selon les corpus anglais ou français

Les spécificités de l'axe mixte « gsmr+dat » s'articulent notamment, pour le corpus anglais, autour de 3 domaines : le prix du climat, celui des paysages et l'alimentation (avec, entre autres thèmes, les signes de qualité[10]). Si pour les 2 premiers domaines cette spécificité est imputable à une double appartenance d'un membre du laboratoire, pour le troisième elle doit

[9] Cette affirmation est toutefois à nuancer, car le lemme *effet* se trouve dans *gaz à effet de serre*…

[10] Les lemmes « food » et « fqs », pour « food quality schemes », sont spécifiques de l'axe gsmr+dat.

plus à une collaboration de membres de GSMR et de membres de DAT[11]. Cette collaboration inter-axes autour de l'alimentation est moins visible dans le corpus français : *alimentation* n'est pas spécifique, *qualité* un peu. Ce qui laisse penser que les discours sur l'alimentation doivent différer selon que le résumé est en français ou en anglais.

		Langue du résumé					
		(absent)	Anglais	Anglais-Français	Français	Langue autre	**TOTAL**
axe	pas d'axe	8	0	3	12	0	**23**
	dat	133	245	165	273	1	**817**
	gsmr	193	37	74	256	3	**563**
	gsmr+dat	13	15	21	48	0	**97**
	TOTAL	**347**	**297**	**263**	**589**	**4**	**1500**

Tableau 2 : Langue du corpus résumé et axe de recherche

Populaire et *ouvrier* sont fortement spécifiques de l'axe GSMR dans le corpus français, les publications où ces deux lemmes sont présents sont très souvent uniquement en Français. Evidemment certains lemmes comme *classe*, *social*, *militant*[12], sont présent dans les deux corpus. Il est notable que si *agriculteur* est spécifique, dans le corpus français, de l'axe GSMR, *agricole* est antispécifique (le lemme *agricole* est très souvent précédé de *politique*, *marché*, *secteur*, *développement*, des lemmes provenant plutôt du discours économique).

Emission et *carbon* sont des lemmes très spécifiques de l'axe DAT dans le corpus anglais, qui le sont un peu moins dans le corpus français. Pour cet axe de recherche, beaucoup de lemmes bien représentés dans un corpus le sont dans l'autre : *spatial*[13] (lemme le plus spécifique de l'axe DAT dans le corpus anglais), *coût*, *modèle*, *entreprise*, *proximité*… Remarquons que *modèle* n'est que faiblement surreprésenté, car présent aussi dans l'axe gsmr+dat.

4.1.2. Classifications des résumés anglais et français

Sur les 852 résumés français, 666 ont été classifiés dans 18 classes différentes.

Le discours des classes de la partie gauche de l'arbre (graphique 3) concerne l'aspect local (*localement*, *bassin* pour la classe 7, *pôle*, *urbain*, *périurbanisation* pour la classe 6) avec un aspect économétrique (*estimer*, *régression* pour la classe 5).

Vient ensuite un bloc de cinq classes assez hétérogènes en termes de contenu : une classe assez typique, puisque retrouvée dans toutes les classifications faites avec différents nombres de classes, concerne l'innovation (*innovation*, *innover*, *frein*), l'*organisation* et l'*entreprise*, en prenant en compte la *proximité* (classe 18). La classe 16 concerne les gaz à effet de serre (*serrer*[14], *gaz*, *ges*). La classe 13 tourne autour du conseil agricole (*conseiller*, *chambre*,

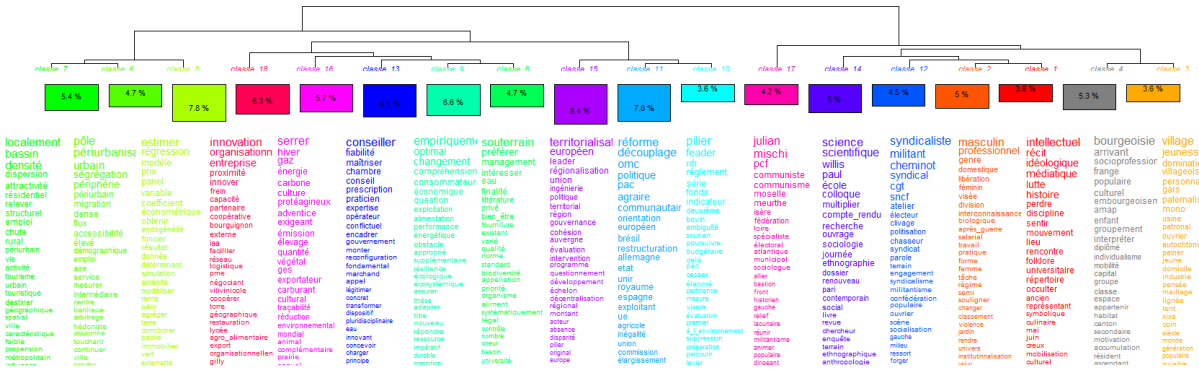
[11] De ce fait il aurait été préférable de construire une appartenance spécifique pour l'agent membre des 2 axes, afin de pouvoir mettre en évidence des collaborations entre les membres des 2 axes. Ici, il y a confusion entre les deux situations.

[12] Class, social, activist.

[13] La direction du laboratoire insiste sur l'importance du territoire comme élément central dans les thématiques de recherche, et le spatial est un des aspects du territoire.

[14] La forme *serre* a été faussement attribuée au lemme *serrer*.

conseil, *prescription*, *expertise*, *encadrer*…). La classe 9, comportant beaucoup de thèses, s'intéresse au calcul empirique (*empirique*, *empiriquement*) et aux diverses performances en agriculture (*performance*, *exploitation*, *alimentation*, *consommateur*, *énergétique*, *écosystémique*) ; tandis que la classe 8, assez proche, parle de qualité (*qualité*, *norme*, *standard*, *appellation*), notamment concernant l'eau (*eau*, *souterrain*).

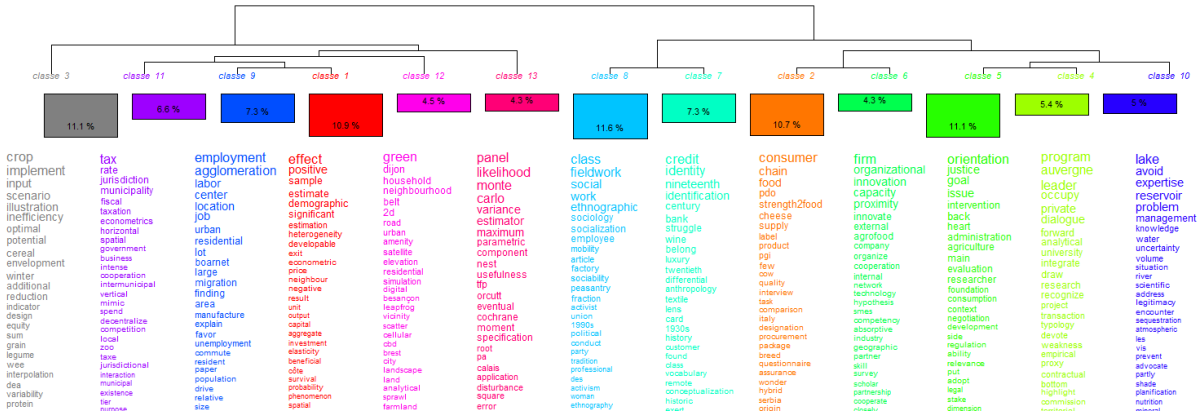


Graphique 3 : classification des résumés français : texte entier

Le bloc de 3 classes suivant est focalisé autour de la PAC (*territorialisation*, *européen*, *union*, *leader*, *régionalisation* pour la classe 15, *réforme*, *découplage*, *omc*, *politique*, *pac* pour la classe11, *deuxième pilier*, *feader*, *rdr*, *règlement*, *fonds*, *indicateur* pour la classe 10).

Enfin le bloc de droite rassemble les discours sociologiques, avec des thèmes très marqués, comme le communisme[16] (*pcf*, *communiste*, *communisme*, *fédération*…) pour la classe 17. La classe 14 parle de disciplines (*science*, *scientifique*, *école*, *sociologie*, *ethnographie*). La classe 12 apparaît comme le pendant syndical de la classe 17 (*syndicaliste*, *militant*, *cheminot*…). La classe 2 concerne le genre (*masculin*, *genre*, *libération*, *féminin*) dans les différents lieux (*domestique*, *professionnel*), tandis que la 1 concerne notamment la sociologie des mobilisations (*lutte*, *mouvement*). Enfin la classe 4 se rapporte à l'embourgeoisement (bourgeoisie, socioprofessionnel, culturel, embourgeoisement) et la 3 au paternalisme et à la jeunesse (*village*, *jeunesse*, *gars*, *domination*, *paternalisme*).

Sur les 560 résumés anglais, 441 ont été classifiés.



Graphique 4 : classification des résumés anglais : texte entier

[16] Cette classe est surtout à destination du grand-public, donc les auteurs concernés ont une volonté de vulgarisation du savoir.

Les classes anglaises semblent bien différentes des françaises. Il n'y a guère que la classe 6 (*firm, organizational, innovation, proximity*) qui semble fort similaire à la classe 18 française. On pourra remarquer la classe orange 2 qui concerne un programme de recherche européen, donc destiné principalement à des publications dans des rapports ou des revues en anglais. Ce programme (*strength2food*) concerne les consommateurs (*consumer*) les circuits de commercialisation alimentaires (*food supply chain*), les signes de qualité (*pdo*)…

Sur la gauche on remarque des classes très économétriques (*effect, positive, sample, estimate* pour la classe 1, et *panel, likelihood, monte carlo, variance, estimator* pour la classe 13). L'économétrie spatiale (*spatial, econometrics*[17]) utilisée pour étudier la taxation (*tax, rate, fiscal, taxation*…) à différents niveaux géographiques (*jurisdiction, municipality, government, intermunicipal, decentralize*…) est le principal thème abordé dans la classe 11. La classe 9 est centrée sur l'économie de l'agglomération, avec des lemmes traitant du foncier (*agglomeration, center, urban, residential, lot, area*…) et des déplacements (*migration, commute, Boarnet*), d'autres du marché du travail (*employment, labor, job, unemployement*…). La classe 8 concerne le prix du paysage (*green, road, amenity, landscape, neighbourhood, belt*…) calculé à partir de celui des transactions immobilières (*dijon, household, residential*…) et d'un modèle numérique de terrain (*satellite, elevation*…), ainsi que de modèles de croissance urbaine (*2d, road, urban*…). La classe 3 est plus difficile à interpréter, il y est question d'inefficacité (*inefficiency, scenario*), de cultures (*crop, cereal, winter, weed*).

Au centre du graphique 4, deux classes sociologiques (*class, fieldwork, social, work, ethnographic, sociology*… pour la classe 8 et *credit, identity* pour la classe 7, dont le discours ne semble pas être dans le corpus français).

Enfin au sein du bloc situé à droite du graphique, les classes 4 et 5 sont relativement proches et partagent de ce fait des lemmes caractéristiques (*territorial, policy*). La classe 5 traite de *l'agriculture,* du territoire (*territorial*), des politiques publiques (*policy, public, intervention, orientation, intervention, administration, regulation*) et de leur finalité (*justice, goal*). Dans la classe 4 il est aussi question de politique (*policy, framework, project, program, Leader*…) et de social (*socially, societal, society, socio*) avec un aspect recherche (*research, study*) prononcé. La classe 10 traite un peu différemment de politique (*management*, lié à *knowledge*) dans le domaine de l'eau (*lake, water, river*…) et aussi des gaz à effet de serre (*sequestration, carbon, atmospheric*…).

Cet apparent faible recouvrement est-il confirmé par un croisement des deux classifications ?

4.1.3. Croisement des classifications dans les deux langues

Comme on peut le déduire du tableau 1, 69% des résumés français ne sont pas traduits, ce qui peut expliquer la faible concordance thématique. Mais toutes les classes ne sont pas concernées identiquement : la classe 17 du corpus français, sur le communisme et à destination du grand-public, n'est jamais traduite, les classes 14 (sur la discipline sociologique), 6 (*pôle, urbain, périurbanisation*) et 3 (paternalisme et jeunesse) sont particulièrement peu traduites (entre 15 et 20% des textes). À l'opposé, les classes 5 (*estimer, régression*) et 18 (*innovation, organisation, entreprise*) sont traduites une fois sur deux.

[17] Le lemme *econometrics* est spécifique de cette classe, alors que *econometric* l'est de la classe 1.

Quant aux résumés anglais, 53% ne sont pas traduits en Français. Et c'est surtout le cas des classes 9 du travail (*employment, agglomeration, labor*) et 13 des méthodes économétriques (*panel, likelihood, monte, carlo, variance, estimator*…). On peut subodorer que ces classes sont typiquement celles qui visent les revues internationales.

Pour avoir une vision relativement complète des publications, il est donc nécessaire d'analyser les deux corpus, car ils ne se recouvrent pas.

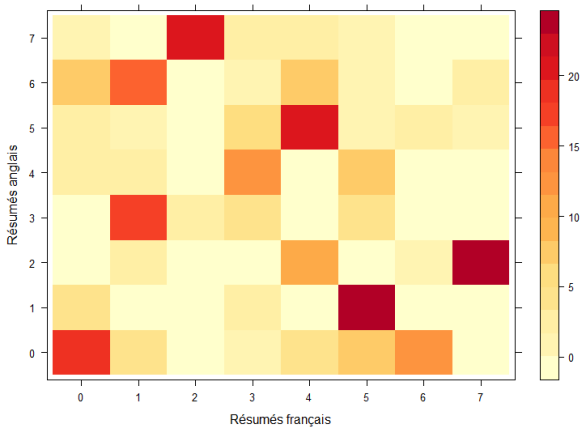
Le croisement des deux corpus ne donne pas une concordance très nette, comme attendu vu les lemmes caractéristiques des classes. Toutefois la classe française 18 (*innovation, organisation, entreprise*) correspond bien à la classe 6 anglaise, la classe 5 (*estimer, régression*) à la classe 1 anglaise (*effect, positive, sample, estimate*).

4.1.4. Corpus résumés bilingues : résultat du croisement des classifications

L'absence de concordance des classes est-il uniquement imputable aux « trous » présents dans chaque corpus ou y-a-t-il un effet langue (le lemme anglais *land*, par exemple, est traduit par *sol, terre, terrain, foncier* dans le corpus français) qui perturberait cette concordance ?

Pour répondre à cette question, des classifications ont été effectuées sur les corpus où les résumés français et anglais sont présents simultanément, c'est à-dire sur 263 publications (voir le tableau 1). Sept classes ont été définies dans chacun des deux corpus, plus la classe 0 des non classifiés.

Voici le résultat de ce croisement, sous la forme d'un « heatmap » :



Graphique 5 : heatmap du croisement des classes anglaises et française, corpus bilingue

Le recouvrement est plutôt bon, sans être parfait bien sûr (ainsi la classe française 3 n'a pas de classe anglaise bien définie). On peut en conclure que la non correspondance des classes constituées avec l'ensemble des résumés est en grande partie imputable à des publications présentes dans un seul des 2 corpus (anglais ou français).

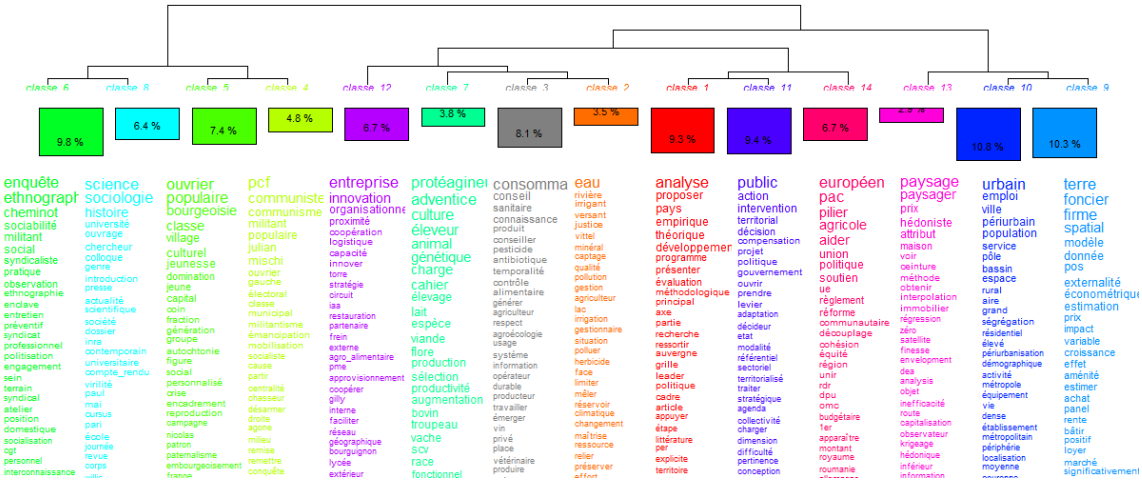
4.2. Corpus résumés français : classification des textes entiers et des segments de texte

Les résumés sont de longueur variable (le nombre médian de mots est de 153, le premier quartile de 110 et le troisième de 206), certains sont très longs et peuvent contenir plusieurs types de discours. Il paraît opportun, pour retenir chaque partie du résumé, de faire une classification sur des segments de texte, chaque résumé étant découpé en segments d'une certaine longueur. Il sera de ce fait difficile de croiser les corpus anglais et français, les segments ne coïncidant pas dans les deux langues. Par contre il sera possible de croiser une

telle classification dans une langue avec la classification faite sur les paragraphes dans la même langue. Ce croisement sera effectué sur le corpus français.

4.2.1. Corpus français : classification des segments de texte

Les 852 résumés français ont été découpés en 4063 segments de texte de longueur approximative de 40 mots. 3383 de ces segments ont été classifiés en 14 classes.

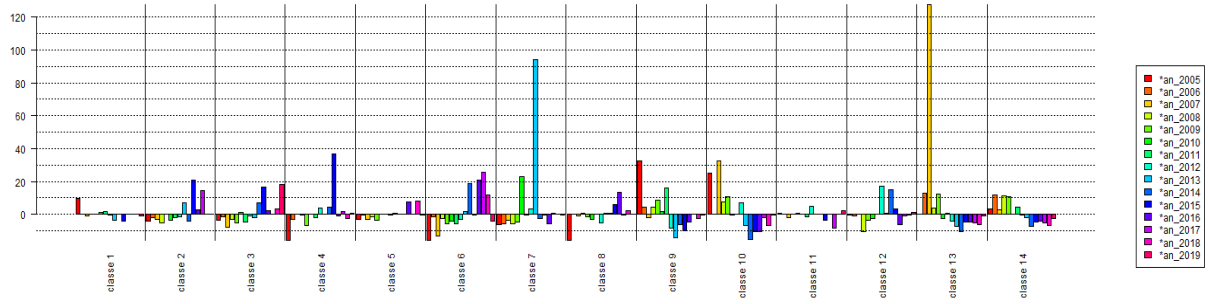


Graphique 6 : classification des résumés français : segments

Les quatre classes de gauche sont des classes sociologiques, dont beaucoup de lemmes caractéristiques ont été mis en évidence dans la classification sur les résumés entiers.

Juste à côté, la classe 12 (*entreprise*, *innovation*, *organisationnel*, *proximité*…) semble définie par le même ensemble de lemmes que la classe 18 du corpus résumé entier. La classe 14 partage beaucoup de lemmes spécifiques avec la classe 11 des paragraphes, la 9 avec la 5 des paragraphes. La classe 7 a quelques lemmes communs avec la classe 16 des paragraphes, tandis que la 11 partage ses lemmes avec les classes 6 et 7 des paragraphes. Les autres classes semblent a priori moins correspondre à des classes de résumés entiers.

Un des objectifs de ce travail textométrique est d'analyser l'évolution des thématiques du laboratoire entre 2005 et 2019. Ce qui sera fait ici, à titre d'exemple, sur le corpus français, à partir de la classification des segments de texte.



Graphique 7 : classification des résumés français : segments

Les classes 2 (*eau*), 3 (*consommateur*) et 6 (*enquête ethnographique*) sont plus spécifiques des années récentes. Contrairement aux classes 9 (*terre*, *foncier*, *spatial*), 10 (*urbain*, *emploi*, *périurbain*) 13 (*paysage*) et 14 (*européen*, *pac*).

La classe 7 est surreprésentée durant l'année 2013 et un peu moins en 2010. Elle est plutôt sous-représentée dans les années 2005 à 2009.

Les classes 1 (*analyse, empirique, théorique*) et 11 (*action publique*) ne semblent pas varier d'une année à l'autre.

4.2.2. Croisement des classifications "segment de texte" et "paragraphe" (corpus français)

Le croisement de la classification des segments de texte (14 classes) et des résumés entiers (18 classes) montre que les segments d'un même résumé ne sont pas toujours classifiés ensemble, avec l'exception notable de la classe 18 (*innovation, innover, organisation*) qui est constituée à près de 80% de segments de la classe de segments 12, c'est donc une classe homogène avec des discours qui lui sont spécifiques. Les classes 3, 5 et 8 du corpus paragraphes sont assez homogènes, mais dans une moindre mesure. La classe de paragraphes 3 (paternalisme et jeunesse) est constituée en grande partie de segments de la classe 5 (*ouvrier, populaire, bourgeoisie*…), la classe de paragraphes 5 (*estimer, régression*) de segments de la classe 9 (*terre, foncier, firme, spatial, modèle, économétrique*…), et les paragraphes de la classe 6 (*pôle, urbain, périurbanisation*) de segments de la classe 10 (*emploi, ville, périurbain*). Les autres classes de paragraphes sont au minimum constituées principalement par deux classes de segments.

4.2.3. Classification des paragraphes à partir des segments (corpus français)

Le regroupement de différentes classes des deux classifications n'homogénéise que peu les classes de paragraphes, ce qui semble normal. En effet, un discours d'une certaine longueur abordera généralement différents aspects du problème (la méthode, le sujet, les données…) ce qui se traduira par différentes classes de discours.

Pour rendre compte de cette multiplicité des discours au sein d'un même résumé, une autre classification des résumés entiers a été effectuée à partir de la classification des segments de texte : chaque classe de paragraphe est caractérisée par les classes de segments la constituant. Ces nouvelles classes seront appelées clusters pour les différencier des autres.

Pour cela une ACM est effectuée sur la présence absence d'une classe de segments dans les paragraphes. Il y a 14 axes factoriels, construits à partir des 14 classes de segments, sur lesquels une classification ascendante hiérarchique est effectuée, donnant 17 clusters.

Les exemples de clusters suivants illustreront la façon dont les discours peuvent s'articuler au sein d'un même résumé. Les six premiers clusters regroupent de différentes manières des classes de segments sociologiques (6, 8, 5 et 4) ; ici ne seront présentés que trois clusters.

Le premier cluster est purement et « entièrement » sociologique : les 4 types de discours sociologiques ont tendance à être présents. Il n'y a toutefois que 7 paragraphes sur 43 qui ont les 4 classes. Il y a toujours la classe 4 (*pcf, communiste, mishi*), souvent (respectivement 72 et 69%) les classes 5 (*ouvrier, populaire, bourgeoisie*) et 6 (*enquête, ethnographique*…) et moins souvent la classe 8 (39%) (*science, sociologie, histoire*…).

Le quatrième cluster est centré sur la classe 8 (*science, sociologie, histoire*…), une fois sur deux sur la classe 6 et jamais sur les classes 4 et 5. Il doit y être question de méthode.

Le deuxième cluster est centré sur les classes 8 et 5 et une fois sur deux sur la classe 6. Les neuf parangons de ce cluster sont les paragraphes qui ont la plus petite distance au centre de leur classe (0.335767), ce qui correspond à la présence exclusive des classes 5, 6 et 8 (avec la classe 0 des segments non classifiés). Voici deux parangons de ce cluster, constitués de segments de la classe 5 (en vert), 6 (en bleu), 8 (en rouge) et 0 (en noir) :

obs	ident	Segment de texte	Classe de segment
651	170124	les sciences sociales ne constituent pas des sciences habituelles de l enfance par comparaison avec les sciences cliniques et médicales en particulier	8
652	170124	pourtant elles peuvent apporter beaucoup à la compréhension du jeune âge autour des notions d historicisation de l enfance de socialisation primaire ou encore de différenciation sociale des enfants	8
653	170124	cette dernière entrée est ici privilégiée sachant qu elle a l avantage de ne pas exclure les deux autres on parlera ainsi d historicisation et de socialisation différenciées nous insistons sur trois manières distinctes d_aborder la différenciation des enfants	5
654	170124	par la différenciation concrète des enfances par l identification différentielle des enfants par les perceptions enfantines des différences sociales nous évoquons par ailleurs les enjeux méthodologiques inhérents à l étude de ces enquêtés particuliers que sont les enfants	6
1610	276929	la société urbaine semble avoir envahie tous les espaces réduisant à néant l objet de la sociologie rurale et de l ethnologie de la france ces disciplines se sont constituées sur une coupure urbain rurale franche réservant à l urbain le vocabulaire canonique de la sociologie comme l analyse des classes sociales	8
1611	276929	une conceptualisation ad hoc société paysanne communauté collectivité villageoise interconnaissance villageoise notable empruntée à l anthropologie a fondé les études rurales l évolution radicale des mondes ruraux contemporains a balayé cette conceptualisation	8
1612	276929	mais les mondes ruraux contemporains sont ils aujourd_hui les stricts équivalents des mondes urbains nous défendons ici l idée que la morphologie sociale des mondes ruraux contemporains ne correspond ni à une france moyenne en réduit	0
1613	276929	ni à des particularités locales de manière récurrente on observe dans les mondes ruraux contemporains une surreprésentation des classes populaires notamment ouvrières et une sous représentation des franges culturelles des mondes supérieurs	5
1614	276929	de même des phénomènes de double ou multi résidences participent d une appartenance à divers degrés à l espace social observé nous proposons ici de reconstruire une sociologie des mondes ruraux comprise comme une sociologie de la localisation des groupes sociaux à l échelle macro	6
1615	276929	sociale et une sociologie des espaces sociaux localisés produits de la localisation différenciée des groupes sociaux sur le territoire	5

Tableau 3 : Deux parangons du cluster 2

4. Le corpus *titre*, conclusions et perspectives

Des classifications sur le corpus titre (français et anglais, respectivement en 15 et 12 classes), ne montre pas une concordance nette, ce qui était déjà le cas pour les résumés. Les résultats ne seront pas analysés ici.

Face à un tel corpus bibliographique, plusieurs façons de l'analyser de façon textométrique sont possibles : analyser les titres, les résumés, les mots-clés. Et si on retient les résumés, soit on considère le résumé comme une observation du tableau lexical, soit on le découpe en segments de texte.

Une des conclusions de ce premier travail est qu'il n'y a pas coïncidence nette entre une classification des segments et une classification des résumés entiers, ce qui reflète la possible multiplicité des discours au sein d'un même résumé. Une classification des paragraphes fondée sur celle des segments de texte permet de créer des clusters de paragraphes, caractérisés par la combinaison de classes de segments y cohabitant.

Une autre conclusion est qu'il faut analyser les corpus anglais et français (les autres langues étant rares, elles ne sont pas analysables de façon textométrique) pour avoir un panorama complet des publications du laboratoire.

Une façon d'enrichir les conclusions de l'analyse serait de croiser une classification basée sur la textométrie avec celle issue d'un réseau d'acteurs, la variable auteurs étant présente dans le corpus global.

References

Lebart L. and Salem A. (1994). *Statistique textuelle*, Dunod

R Development Core Team (2005). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0*, URL: http://www.R-project.org.

Ratinaud P. (2009). Iramuteq: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires.URL: http://www.iramuteq.org.

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