Analyse des 44 opérations d'expertise scientifique collective, d'étude et de prospective réalisées par l'Inra de 2000 à 2020, Volume 1 : analyse textuelle des résumés, INRAE (France)

Regina Dashkina, Marc Barbier, Guy Richard, Audrey Bethinger, Marc Antoine Caillaud, Catherine Donnars, Agnès Girard, Kim Girard, Chantal Le Mouël, Sophie Le Perchec, et al.

INRAE

20 ans d'expertise scientifique collective, de prospective et d'étude à l'Inra

Analyse des 44 opérations réalisées entre 2000 et 2020

Direction de l'Expertise scientifique collective, de la Prospective et des Études (DEPE)

Juin 2021

Rédacteurs principaux :
Regina Dashkina (DEPE), Marc Barbier (LISIS) et Guy Richard (DEPE)

Contributeurs :
Audrey Béthinger, Marc-Antoine Caillaud, Catherine Donnars, Agnès Girard (PHASE), Kim Girard, Chantal Le Mouël, Sophie Le Perchec (DipSO), Sophie Leenhardt, Virginie Lelièvre (AgroEcoSystem), Hugues Leiser (DipSO), Olivier Mora, Mégane Raulet, Olivier Réchauchère, Isabelle Savini, Anaïs Tibi

Directeur de la publication :
Guy Richard - guy.richard@inrae.fr

Pour citer ce document :
Dashkina R., Barbier M., Béthinger A., Caillaud M.-A., Donnars C., Girard A., Girard K., Le Mouël C., Le Perchec S., Leenhardt S., Lelièvre V., Leiser H., Mora O., Raulet M., Réchauchère O., Savini I., Tibi A., Richard G., 2021, Analyse des 44 opérations d'expertise scientifique collective, d'étude et de prospective réalisées par l'Inra de 2000 à 2020, Volume 1 : analyse textuelle des résumés, INRAE (France), 52 pages.

Dataset :
https://doi.org/10.15454/QRIZCR
Document imprimé en juin 2021

DOI : 10.15454/pqer-jw13

Analyse des 44 opérations d'expertise scientifique collective, d'étude et de prospective réalisées par l'Inra de 2000 à 2020

Volume 1 : analyse textuelle des résumés

Direction de l'Expertise scientifique collective, de la Prospective et des Études (DEPE)
Juin 2021

TABLE DES MATIÈRES

TABLE DES FIGURES

1. Introduction

Depuis le début des années 2000, l'Inra, devenu INRAE le 1er janvier 2020, a conduit une quarantaine d'opérations d'expertise scientifique collective, d'étude et de prospective sur des thématiques relevant de l'agriculture, de l'alimentation et de l'environnement. Ces opérations ont été coordonnées à partir de 2010 par la Délégation à l'Expertise scientifique collective, à la prospective et aux études (DEPE)[1].

Caractériser la diversité des thèmes couverts, les situer au sein du périmètre scientifique d'INRAE, mettre en évidence les liens entre les opérations, caractériser les collectifs d'experts mobilisés et analyser les corpus bibliographiques associés aux opérations apparaît pertinent pour mieux faire apparaître ce qui a été fait, comment cela a été fait, et par qui cela a été fait. C'est également une réflexion qui peut venir alimenter le plan stratégique INRAE2030. Pour cela, la DEPE a choisi de développer des compétences individuelles et collectives pour réaliser une relecture commune de l'ensemble des productions existantes par un travail d'analyse des contenus, des experts mobilisés et des corpus documentaires. Ces trois champs d'analyse ouvrent un large spectre d'exploration de la nature et du positionnement des opérations conduites, et nécessitent de nouvelles compétences venant des « data sciences » et de la bibliométrie.

Un groupe de travail et de formation a été constitué en interne à la DEPE sous la forme d'une « Master Class ». Le groupe de travail a mobilisé deux outils d'analyse textuelle : le logiciel Iramuteq[2] et la plateforme CorTexT[3]. Ce projet, d'une durée de deux ans, s'inscrit dans une perspective d'open science avec un accompagnement des collègues de la Direction pour la science ouverte (DipSO).

Grâce à cette mobilisation des agents de la DEPE, un ensemble de données structurées a pu être rassemblé, traité et mis en forme pour conduire les analyses textuelles. Dans ce rapport nous restituons les résultats de ces analyses qui ont porté sur les 44 opérations[4] d'expertise scientifique collective, d'étude et de prospective réalisées par l'Inra de 2002 à 2020[5].

Ce premier volume d'une série de trois concerne la description générale des opérations et l'analyse textuelle des résumés des 44 opérations conduites. Le deuxième volume sera consacré à la connaissance des plus de 800 experts mobilisés depuis 2000. Le troisième volume sera consacré à l'analyse du corpus documentaire à la base de l'ensemble des opérations.

1 Cette délégation est devenue la Direction de l'expertise scientifique collective, de la prospective et des études (DEPE) avec la création d'INRAE.
2 http://www.iramuteq.org/ ainsi que la documentation http://www.iramuteq.org/documentation/fichiers/documentation_19_02_2014.pdf
3 https://www.cortext.net/projects/cortext-manager/
4 Page internet des opérations DEPE
5 L'analyse intégrant des opérations similaires conduites par le Cemagref puis IRSTEA n'a pas été possible. L'analyse ne fait donc référence qu'aux opérations conduites par l'Inra.

2. Description des opérations

2.1. Données utilisées

Les opérations de la DEPE sont de trois types : expertise scientifique collective (ESCo), prospective et étude (noté EPE dans le document). Chaque opération est réalisée par un comité d'une vingtaine d'experts, plusieurs livrables peuvent être produits : résumé, synthèse, rapport. Des corpus bibliographiques plus ou moins importants (quelques centaines à quelques milliers de références par opération) sont mobilisés.

Les données suivantes ont été réunies pour chacune des opérations EPE analysées :

- Le résumé (environ 10 pages).
- La synthèse dans le cas des expertises et des études (environ 100 pages).
- Le rapport des prospectives (environ 100 pages).
- Le rapport scientifique des expertises et des études (500 à 1 000 pages).
- Le corpus des références bibliographiques citées dans les rapports.
- Les experts mobilisés : origine, discipline, liste des publications scientifiques.
- Les commanditaires : statut et champ d'activité.

Les données sur les publications scientifiques produites par tous les experts scientifiques ont été extraites de la base de données internationale SCOPUS[6] qui a été préférée à la base de données WoS[7] pour mieux représenter les productions des chercheurs en sciences humaines et sociales.

Ce sont au total 44 opérations EPE qui ont été réalisées sur les 20 dernières années, par près de 850 experts scientifiques venant du monde académique public (pour les opérations de prospective, des experts n'appartenant pas au monde académique participent également).

2.2. Fiche de description des opérations

Pour chaque opération un ensemble d'informations a été recueilli sous forme d'un tableau initial de données (Annexe 1). Ces informations sont utilisables par les deux logiciels utilisés (Iramuteq et CorTexT Manager) en tant que variables pouvant être croisées avec l'analyse des contenus :

- Le type d'opération : ESCo, étude ou prospective.
- L'intitulé des opérations.
- L'année du colloque de restitution des résultats.
- Les commanditaires.
- Les domaines thématiques, caractérisés par :
  - L'objet central de l'opération (*e.g.* la forêt, les systèmes d'élevage, les territoires).
  - Le ou les domaines thématiques de l'Inra auxquels se rattache l'opération : agriculture, alimentation ou environnement.
  - La priorité du document d'orientation Inra2025[8] à laquelle se rattache l'opération : #Global, #Climat, #3Perf, #Food ou #BioRes.
  - L'Objectif de développement durable (ODD) tel que défini par l'ONU auquel se rattache l'opération.

2.3. La nature des opérations

La DEPE a conduit 44 opérations depuis 2002 (dont 35 depuis 2010) : 17 ESCo, 11 études et 16 prospectives (*Figure 1*). Les études ont été réalisées à partir de 2010, ce qui explique sans doute leur plus faible nombre. Sauf cas particuliers indiqués dans le texte, les analyses ont été réalisées sur le corpus constitué de l'ensemble de ces opérations.

6 https://www.scopus.com/
7 https://www.webofknowledge.com
8 https://hal.archives-ouvertes.fr/hal-01607768

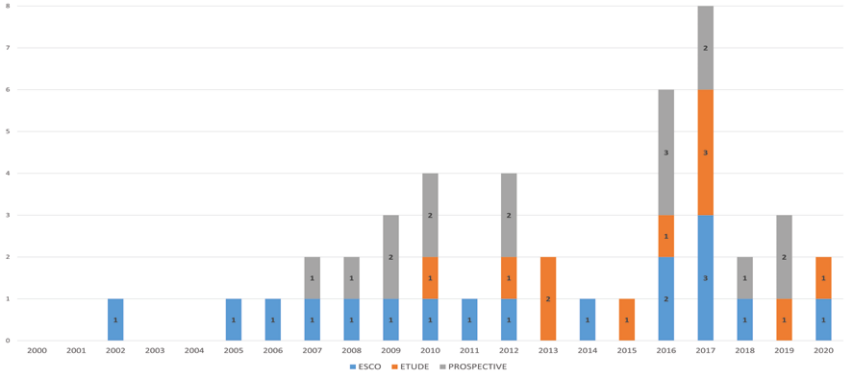


Figure 1 : Nombre annuel d'ESCo, d'études et de prospectives, réalisées par l'Inra depuis le début des années 2000

2.4. Les partenaires des opérations

Douze opérations EPE (soit environ un quart des opérations) ont été conduites en partenariat avec un autre organisme (*Figure 2*). Ces partenaires sont de natures diverses : Établissement public à caractère scientifique et technologique (EPST : CNRS, IRSTEA, IFFSTAR), Établissement public à caractère industriel et commercial (EPIC : CIRAD, IFREMER), Agence publique (Ademe), Établissement public à caractère administratif (EPA : IGN, IFCE). Notons que quatre opérations EPE ont été conduites en partenariat avec le CEMAGREF/ IRSTEA. Trois opérations EPE ont été réalisées dans le cadre d'une alliance (ALLENVI, AGREENIUM).

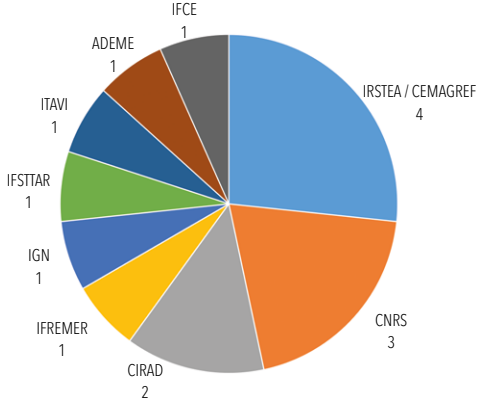


Figure 2 : Partenaires des opérations EPE (nombre d'opérations)

2.5. Les commanditaires des opérations

Les opérations EPE ont été réalisées pour 14 commanditaires externes (*Figure 3*) : des directions générales des ministères en charge de l'agriculture et de l'alimentation, ou de l'environnement ou des territoires, des agences nationales (ADEME, OFB), une agence européenne (EFSA), des instituts techniques (ACTA, IFCE, ITAB, ITAVI), des organismes ou des associations professionnelles agricoles (MSA, Crédit Agricole, Pluriagri), et un conseil régional (Aquitaine).

Les sollicitations internes (Inra, ALLENVI, AGREENIUM) ont principalement concerné des opérations de prospective. Deux éléments sont à retenir :

- Les commanditaires relèvent principalement des trois domaines de l'Inra : nous avons peu travaillé par exemple pour les Territoires et l'Énergie, voire pas du tout pour la Santé et la Ville.
- L'Inra a travaillé pour un seul commanditaire non national, l'EFSA, dans le cadre de l'ESCo sur la conscience des animaux.

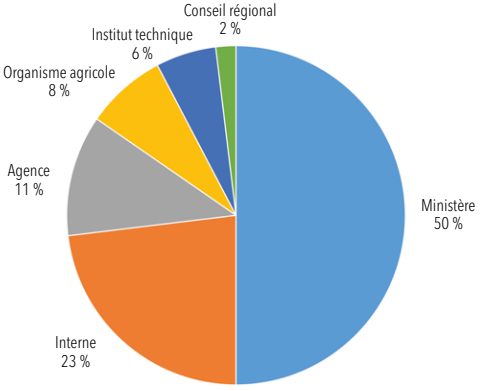


Figure 3 : Opérations EPE par commanditaire (100 = nombre total de commanditaire.opérations)

2.6. Les thématiques des opérations

Parmi les trois domaines d'INRAE, *i.e.* Agriculture, Alimentation et Environnement, l'Agriculture est le domaine principal des activités d'expertise et de prospective (64 %). L'Environnement est le deuxième domaine (41 %) et l'Alimentation est le troisième (18%) (*Figure 4*). Cette répartition reflète en partie à la fois :

- Les forces scientifiques de l'Inra dans chacun de ces trois domaines estimées à partir des effectifs des départements : elles sont approximativement de 51%, 28% et 21%.
- L'image de l'Inra auprès de commanditaires potentiels en termes de compétences en éclairage des politiques publiques.

Il faudrait compléter ces données par les opérations d'expertise relatives aux mêmes domaines et demandées à d'autres organismes (IRD, CIRAD, CNRS, INSERM…) pour mieux évaluer l'image de l'Inra auprès de nos commanditaires potentiels.
Le domaine de l'Alimentation reste sous-représenté. Près de 30 % des opérations EPE ont concerné deux domaines et sont en interface avec l'Agriculture, aucune opération n'a pour autant concerné les trois domaines d'INRAE.

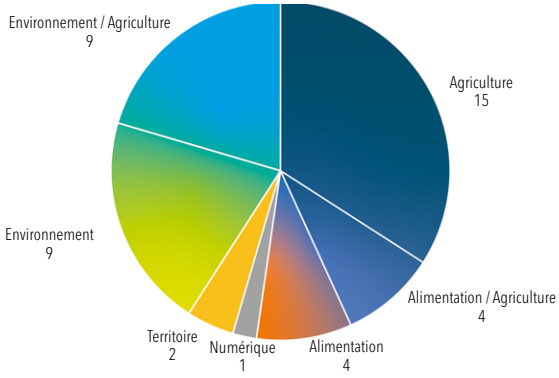


Figure 4 : Nombre d'opérations EPE par domaine de l'Inra

Parmi les cinq priorités du dernier document d'orientation de l'Inra sur la période 2015-2025, *i.e.* #Global, #Climat, #3Perf, #Food et #BioRes (*Figure 5*), les opérations EPE se rapportent principalement à #3Perf (18 opérations), puis à #Climat (14 opérations). La priorité #Global est concernée par six opérations. La priorité #Food est concernée par quatre opérations. Les priorités #BioRes et #OpenScience ne sont concernées que par une seule opération.

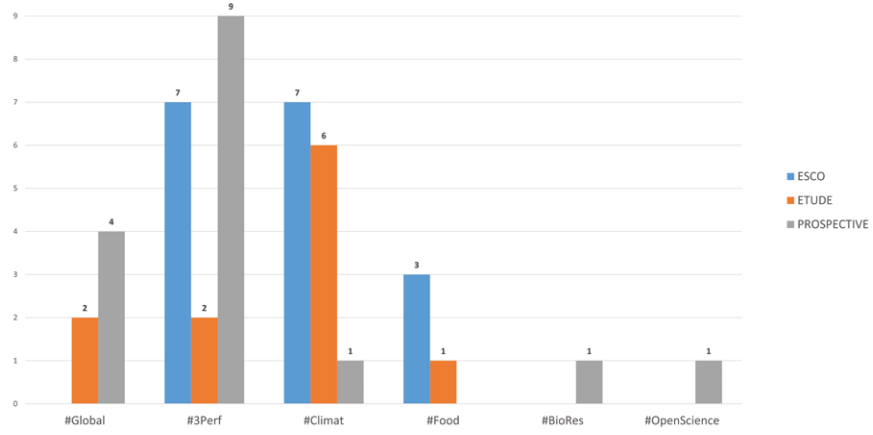


Figure 5 : Nombre d'opérations EPE par priorité scientifique du document d'orientation Inra2025

En analysant plus en détail les thèmes des 44 opérations EPE (*Figure 6*), il apparaît que les aspects biotechniques, au sens large, des productions végétales et animales, ont été abordés avec la même importance que les aspects Filière/ Territoire/Politique. Ces derniers correspondent tous à des prospectives. Les productions végétales ont essentiellement été abordées au travers des

questions de protection des cultures.
Les différents compartiments de l'environnement ont été abordés : Climat, Eau, Sol, Biodiversité et Forêt.
Quatre des six opérations réalisées à l'échelle mondiale s'inscrivent dans la lignée de la prospective Inra-CIRAD AgriMonde, les deux autres traitent de scénarios environnementaux à la demande d'ALLENVI. Aucune ESCo n'a été réalisée dans un cadre global.

Dans le domaine de l'alimentation, les notions de type de produits (fruits et légumes, produits animaux), de pertes et gaspillages, et de comportements alimentaires ont été abordées.
Enfin, seul l'usage non alimentaire des produits végétaux a été abordé dans le domaine de la bioéconomie.
Plusieurs de ces opérations EPE s'inscrivent directement dans six des 17 ODD (*Figure 7*).

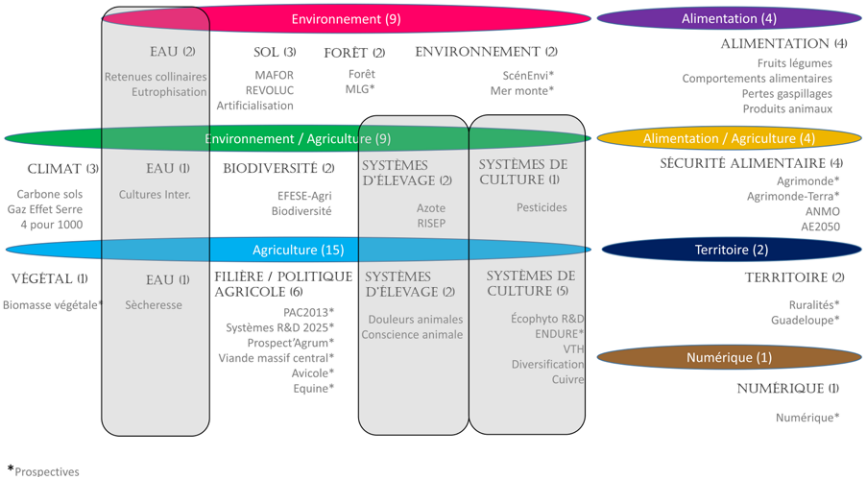


Figure 6 : Nombre d'opérations EPE par objet central de recherche (un seul objet a été attribué à chaque opération)

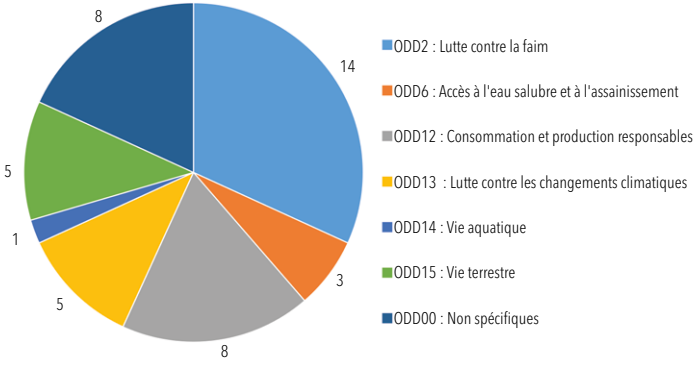


Figure 7 : Nombre d'opérations EPE par objectif de développement durable de l'ONU

2.7. Conclusion

Les opérations d'expertise et de prospective ont largement couvert le périmètre Inra/INRAE depuis une vingtaine d'années, avec cependant plusieurs thématiques peu abordées : la forêt et les prairies (contrairement aux grandes cultures), la santé des animaux et des élevages (contrairement à la santé des cultures), la santé des consommateurs, la bioéconomie. Par ailleurs, aucune opération n'a concerné simultanément les trois domaines du tripode : agriculture, alimentation et environnement.

3. Les méthodologies mobilisées

3.1. Quelques précisions sur le vocabulaire

L'analyse textuelle est un domaine fondé sur la linguistique et la statistique des composantes du corpus écrit. Domaine ancien, transformé par l'arrivée de l'informatique, des bases de données relationnelles puis de l'intelligence artificielle, il est à nouveau l'objet de transformation avec l'arrivée des « big data », de l'extension de la production d'écrits dans les réseaux sociaux et de la présence de nombreuses traces numériques dans les organisations. De nouvelles capacités d'analyse textuelle accompagnent cette mutation.
Quelques précisions sont indiquées ci-après sur le vocabulaire de base de l'analyse textuelle afin de faciliter la lecture de ce rapport :

- Une forme = un mot réduit à sa racine de déclinaison dans un texte selon les règles suivantes : les verbes sont ramenés à l'infinitif, les noms sont ramenés au singulier, et les adjectifs sont ramenés au masculin et au singulier.
- Un hapax = une forme qui apparaît une seule fois dans un texte.
- Une expression ou un N-gram ou un segment de texte = un ensemble d'un certain nombre de mots.
- Un corpus = un ensemble de textes.

3.2. Les outils utilisés

Deux logiciels d'analyse textuelle ont été mobilisés : Iramuteq et CorTexT Manager. Les textes issus des résumés ont été préparés selon la procédure présentée dans l'Annexe 2.

Iramuteq est une application en ligne, basée sur le logiciel ALCESTE[9], qui est très utilisée dans le domaine des sciences sociales. L'analyse de base est la classification hiérarchique descendante de segments de texte contenant des mots détectés à partir d'un dictionnaire. Les méthodes de classification hiérarchique descendante débutent avec une seule classe contenant la totalité des formes considérées dans l'analyse, puis la divisent à chaque étape selon un critère d'opposition entre les formes du point de vue de leurs co-occurrences jusqu'à l'obtention d'un ensemble de classes différentes.

CorTexT Manager (ci-après CorTexT) est une application en ligne développée par la plateforme CorTexT de l'UMR LISIS[10] et du LABEX SITES[11]. Il s'agit d'une plateforme proposant des scripts d'analyse textuelle variés pour les sciences sociales. L'application offre à l'usager la possibilité d'utiliser différents scripts de gestion et de transformation des données, des scripts d'analyse de contenu, des scripts de visualisation des

9 https://hal.archives-ouvertes.fr/hal-00924168/document
10 UMR LISIS = Laboratoire Interdisciplinaire Sciences Innovations Sociétés http://umr-lisis.fr/
11 LABEX SITES = Laboratoire d'Excellence SITES (Sciences, Innovation et Techniques en Société) https://ifris.org/labex/

résultats sous forme de graphes. Dans le présent projet nous avons utilisé l'analyse de réseaux d'expressions associées dénommée clusterisation, les réseaux étant visualisés avec l'algorithme de partitionnement de Louvain. Cet algorithme permet de révéler des réseaux d'expressions appelés clusters à partir d'un calcul d'indice de similitude entre les expressions en fonction de différentes métriques choisies par l'usager (Chi², information mutuelle, distribution, etc.).

3.3. Analyses menées avec Iramuteq[12]

Nous avons utilisé plusieurs méthodes d'analyse textuelle sous Iramuteq :

- Nuage de mots
- Méthode Reinert :
  - Classification simple sur segments de texte
  - Classification simple sur texte
- Analyse de similitudes

3.3.1. Nuage de mots

La construction du nuage de mots se base sur la fréquence des formes présentes dans le corpus. Dans le nuage construit, la taille des formes qui sont représentées est proportionnelle à leur fréquence d'apparition dans le corpus.

3.3.2. Méthode Reinert

La méthode Reinert est une méthode de classification hiérarchique descendante, dont le résultat se présente sous la forme d'un dendrogramme. L'objectif est de faire apparaître des « mondes lexicaux » pour mettre en évidence les thématiques générales du corpus étudié. La méthode de Reinert permet de classer les formes selon leur indépendance mesurée par un test de Chi². La valeur du Chi² exprime l'intensité du lien qui rattache la forme à une classe donnée. A chaque classe du dendrogramme sont associées : (1) la liste des formes qui la composent, (2) une indication de sa taille exprimée en pourcentage du corpus étudié. Tous les segments d'un corpus ne sont pas classés par Iramuteq : plus le pourcentage de segments classés est élevé, plus l'analyse est robuste. L'utilisateur peut faire varier la taille des segments de texte pour maximiser le pourcentage de segments classés. Iramuteq produit également des analyses factorielles des correspondances (AFC). L'AFC organise les classes dans un graphique à deux dimensions, montrant les relations lexicales entre elles en termes de ressemblance ou de dissemblance du vocabulaire qui les compose. Autrement dit, l'AFC permet d'identifier les classes au vocabulaire proche (qui se regroupent sur le graphique), et celles au vocabulaire opposé (éloignées les unes des autres sur le graphique).

Il existe deux modalités d'utilisation de la méthode Reinert : soit le texte est divisé en segments (méthode dite « simple sur segment de texte »), soit le texte est considéré en entier (méthode dite « simple sur texte »). Pour la classification « sur segment de texte », l'algorithme découpe le corpus en segments, puis observe la distribution des formes identifiées dans chaque segment. Il regroupe ensuite les segments en classes en fonction des formes qui les composent par un processus d'itération : plus le nombre de formes communes à deux segments donnés est élevé, plus ces deux segments sont considérés comme proches et susceptibles d'être regroupés dans la même classe. La classification « sur texte » utilise le même processus de classement mais en considérant les textes dans leur totalité. Nous avons privilégié la classification « sur segment de texte » plutôt que « sur texte » du fait d'un nombre de résumés relativement faible (une quarantaine). La taille des segments est de 40 formes par défaut ; nous avons également réalisé les calculs avec des tailles de 20 à 80 formes (Annexe 3.2), et conservé le résultat pour lequel le pourcentage de segments classés était le plus élevé (de l'ordre de 90 %).

12 Les éléments méthodologiques Iramuteq supplémentaires sont présentée dans l'Annexe 3

3.3.3. Analyse de similitudes

L'analyse de similitudes, ou analyse des co-occurrences, permet de dénombrer les fois où les formes apparaissent simultanément dans un segment. Le résultat de cette analyse est présenté sous forme de graphes de formes associées. L'objectif est d'étudier la proximité et les relations entre les formes d'un corpus. Sur le graphe, les formes associées sont regroupées en communautés qui se distinguent par un halo de couleur. La taille des formes est proportionnelle à leur fréquence dans le corpus. L'épaisseur du lien entre deux formes est proportionnelle à l'indice de co-occurrence entre ces deux formes ; celui-ci correspond au comptage du nombre de segments dans lesquels les deux formes sont associées.

3.4. Analyses menées avec CorTexT Manager[13]

3.4.1. Extraction terminologique

L'extraction terminologique des expressions (script Terms Extraction) peut être réalisée à l'échelle de la phrase ou à l'échelle du document, avec la mesure de Chi² ou de la fréquence, pour une taille donnée d'expression (N-gram où N est à définir). Sur la base d'une première analyse de la courbe de distribution des fréquences des expressions extraites, une liste d'un certain nombre d'expressions, de l'ordre de la centaine, est récupérée. Avec les précautions qui s'imposent pour porter l'analyse à l'échelle du document au regard du nombre assez faible de documents, nous avons privilégié, en cohérence avec ce que a été fait avec Iramuteq, l'extraction terminologique à l'échelle de la phrase.
Les expressions sont comptées autant de fois qu'elles apparaissent dans les phrases du document. L'extraction à l'échelle de la phrase permet d'identifier le vocabulaire spécifique à l'échelle des phrases, et qui peut être donc partagé entre plusieurs opérations.

Par ailleurs, la fonction d'extraction terminologique proposée dans l'application CorTexT réalise, outre une lemmatisation (regroupement de tous les mots d'une même forme), la quantification des associations d'expressions les plus fréquentes au niveau des phrases. Par exemple l'expression changement climatique est repérée comme une expression figée du fait de son importance quantitative : elle n'est pas simplement repérée à partir de la fréquence de la relation entre changement et climatique.

3.4.2. Co-occurrences des expressions

Une fois les expressions extraites retenues, l'application CorTexT établit un tableau de contingence des expressions pour calculer les associations entre ces expressions en fonction d'une métrique de similitude. Le graphe obtenu est ensuite l'objet d'un second calcul de partitionnement suivant la méthode dite de Louvain. Cette méthode permet de produire une visualisation des clusters optimisant ce partitionnement. Du point de vue de l'interprétation de la visualisation, chaque cluster est ainsi indicateur d'expressions mises en proximité dans le corpus (suivant la métrique choisie pour calculer la valeur de similitude de ces expressions). En conséquence, ces clusters d'expressions sont interprétables comme des univers de sens présents dans le corpus dans son ensemble. Il est ensuite possible de « projeter » sur cette cartographie des variables présentes dans le tableau initial de données, comme par exemple le type d'opération, de façon à établir et mesurer la « contribution » de telle ou telle opération au cluster.
L'utilisation de la fonction de « clusterisation » de CorTexT permet une analyse qui aborde la construction d'associations telles qu'elles apparaissent au niveau des textes et de faire porter

13 Les éléments méthodologiques CorTexT supplémentaires sont présentés dans l'Annexe 4

une analyse des similitudes qui tient compte non pas seulement d'une mesure brute terme à terme mais également de l'effet des relations de chaque expression dans son contexte propre. Enfin, des analyses complémentaires mettant en valeur la clusterisation peuvent être conduites comme la réalisation d'une carte de contingence avec une variable native du jeu de données, *cf.* ci-après.

3.4.3. Matrice de contingence

Outre la visualisation de clusters sous forme de cartographies, l'application CorTexT permet de produire des matrices de contingence classique fondées sur une mesure du Chi^2 d'un tableau (script *Contingency matrix*). Cela permet d'établir et de visualiser la corrélation entre deux variables quelconques provenant, soit du tableau de données (par exemple le type d'opération ou l'origine des experts), soit de la clusterisation (par exemple les identifiants des clusters calculés précédemment). De la sorte, une mesure de Chi^2 permet de mesurer la contribution spécifique de telle ou telle opération, et de classer ces valeurs sur une échelle bidirectionnelle avec un gradient (contribution positive ou négative). C'est un mode de lecture a *priori* pertinent pour repérer des tendances structurantes du corpus et ainsi compléter l'approche de clusterisation pour en éclairer la signification.

4. Analyse textuelle des résumés avec Iramuteq

4.1. Les mots utilisés

Il s'agit ici de décrire la façon dont sont rédigés les résumés des opérations EPE et de mettre en évidence les différences éventuelles de rédaction reflétant les spécificités de chaque type d'opération.

4.1.1. Nombre de formes et de mots

La longueur du texte de chaque résumé (c'est à dire sans graphique ni tableau) est de huit pages en moyenne, elle varie de 4 à 15 pages (*cf.* les étiquettes des *Figures 8-9*).

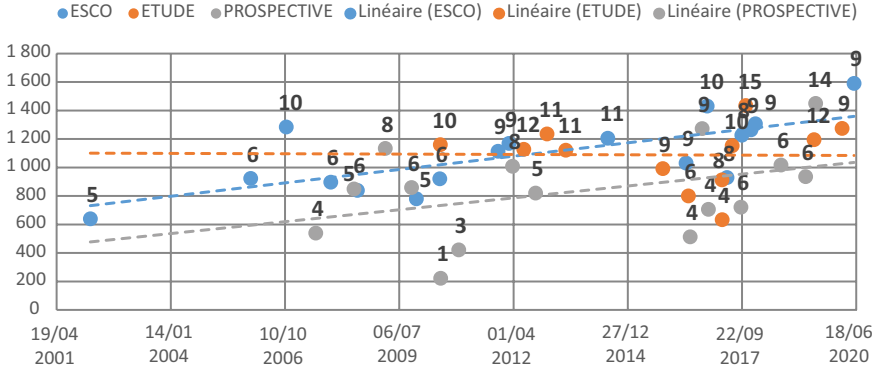


Figure 8 : Evolution temporelle du nombre de formes par résumé (avec en étiquette le nombre de pages de texte de chaque résumé)

Le nombre de formes par résumé est de l'ordre de 1 000 : 1 093, 1 088 et 833 en moyenne, respectivement pour les ESCo, les études et les prospectives. Il tend à augmenter dans le temps pour les ESCo et les prospectives, tandis que pour les études, on ne décèle aucune tendance d'évolution (*Figure 8*). Le nombre de mots par résumé est de l'ordre de 5 000 : 5 010, 6 243 et 3 559, respectivement pour les ESCo, les études et les prospectives (*Figure 9*). Les tendances d'évolution dans le temps sont les mêmes que celles décrites pour le nombre de formes.
Au final, on constate plutôt une augmentation de la longueur du texte des résumés qu'un enrichissement du vocabulaire.

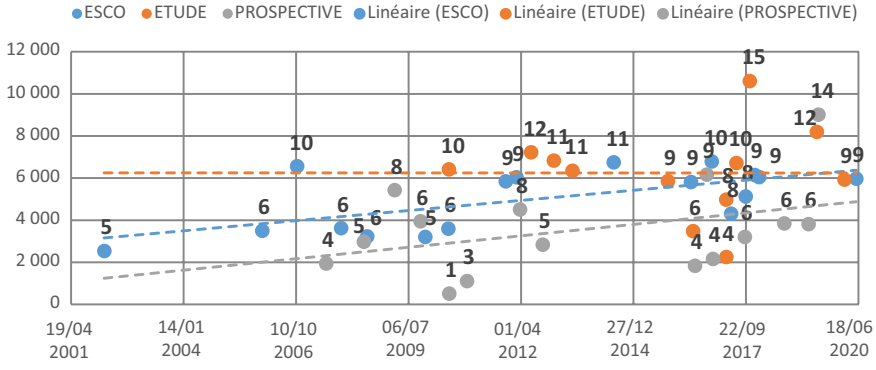


Figure 9 : Evolution temporelle du nombre de mots par résumé (avec en étiquette le nombre de pages de texte de chaque résumé)

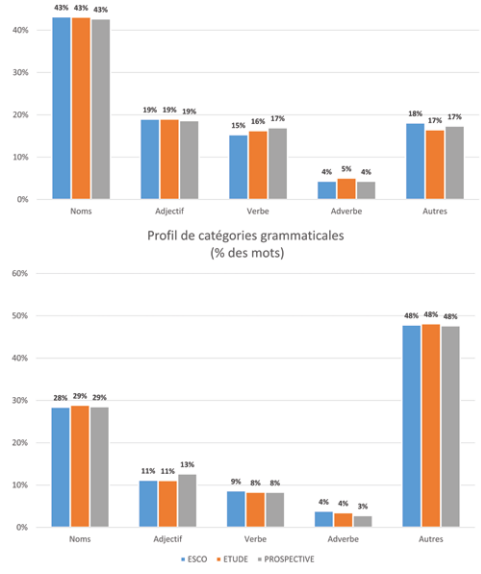


Figure 10 : Profil de catégorie grammaticale

4.1.2. Profil de catégories grammaticales

La répartition moyenne des catégories grammaticales présente des profils très proches entre les types d'opération, que ce soit en pourcentage de formes ou en pourcentage de mots (*Figure 10*). Cela atteste qu'il n'y a pas de différence notable dans l'écriture sur le plan grammatical. Cette homogénéité dans l'écriture, quel que soit le type d'opération, peut être interprétée comme le résultat du travail important des membres de la DEPE en terme de rédaction de leurs productions.

4.1.3. Place des hapax

L'hapax est une forme n'apparaissant qu'une seule fois dans chaque résumé. L'examen des hapax permet donc de mettre en évidence la spécificité de certaines formes et la richesse du vocabulaire. Les hapax correspondent à la fois à des formes très spécifiques comme éviscération, repeuplement… et à des formes parfois très génériques (issues souvent du langage courant) comme étoile, piocher… Les hapax représentent environ la moitié des formes et de 9% à 15% des mots (*Figure 11*), ce qui est un phénomène normal en lexicométrie[14]. Les prospectives ont un pourcentage d'hapax légèrement plus élevé, les études ont un pourcentage d'hapax un peu plus faible : la structure de l'écriture des prospectives apparaît un peu plus riche, mais les différences restent peu importantes.

4.1.4. Conclusion

Au final, on observe des mots utilisés très proches quel que soit le type d'opération, avec un vocabulaire très légèrement plus riche pour les prospectives.

14 On se réfère ici au classique Lebart, L., & Salem, A. (1988). *Analyse statistique des données textuelles: questions ouvertes et lexicométrie.* Paris: Dunod.

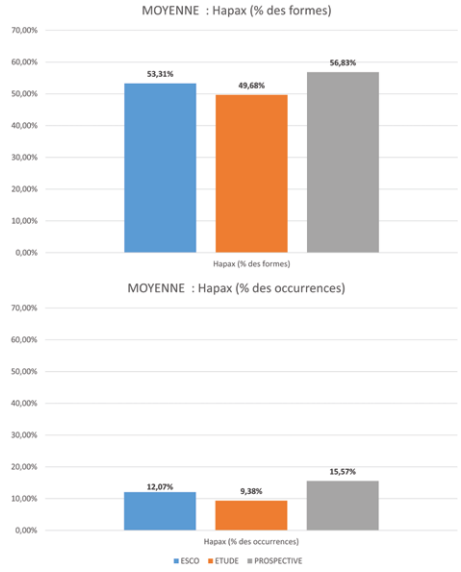


Figure 11 : Nombre d'hapax dans le corpus

4.2. Le mode d'écriture : analyse des verbes

Travailler sur les verbes permet de décrire la façon dont les thèmes sont traités sur le plan de la mise en action des Entités-Objets par des Entités-Sujets au sein de la phrase. Observer une possible hétérogénéité à ce niveau peut permettre de repérer des « façons d'écrire » et donc la qualité intrinsèque et contextuelle (textualité) des différentes opérations EPE du fait de leurs orientations propres[15].

4.2.1. Les verbes les plus cités (nuage de mots)

Les analyses sont présentées par type d'opération ou sur le corpus total (*Figures 12, 13 et 14*). Les verbes *permettre*, *mettre* et *prendre* sont les plus utilisés pour les trois types d'opération : il s'agit de *permettre de dire* ou *permettre d'affirmer*, de *mettre en évidence*, de *prendre en compte*, expressions marquantes de la description de la façon dont les sujets disposent du réel ou l'actionnent en se référant à une autorité ou un contexte.

Les verbes *retenir, lier, utiliser, concerner, montrer, rester* sont très présents dans les ESCo : ce sont des verbes marqueurs des concepts et des analyses supposant le rapport au monde que tisse le texte.

Les verbes des études sont très proches de ceux des ESCo, avec également des verbes comme *couvrir, cultiver, produire,* qui traduisent une entrée plus directement sur les activités agricoles, qui est caractéristique de nombreuses études.

Les verbes des prospectives sont eux très différents : *développer, devenir, construire, intégrer, répondre,* ils traduisent une dynamique et un rapport au temps, une construction verbale très présente dans la description de différentes formes de rapport au futur.

15 On renvoie ici à Adam, J-.M. (2005). La linguistique textuelle. Introduction à l'analyse textuelle des discours. Paris : Armand Colin, collection « Cursus », ainsi qu'à Rastier, F. (1988). Sens et textualité. FeniXX.

Mots-clés	CorpusTotal	ESCo	Etudes	Prospectives
permettre	474	187	196	91
mettre	422	169	161	92
prendre	236	96	75	65
réduire	213	68	110	35
lier	202	86	80	36
considérer	178	60	99	19
demander	160	62	59	39
utiliser	151	75	53	23
retenir	147	118	19	10
produire	144	50	55	39
concerner	133	75	43	15
montrer	119	72	35	12
rester	141	66	49	26
cultiver	136	32	92	12
couvrir	101	20	71	10
développer	138	48	36	54
devenir	103	30	25	48
construire	65	14	13	38
intégrer	137	62	37	38
répondre	81	27	17	37

Figure 12 : Tableau comparatif des 10 verbes les plus cités par type d'opération (cases colorées)

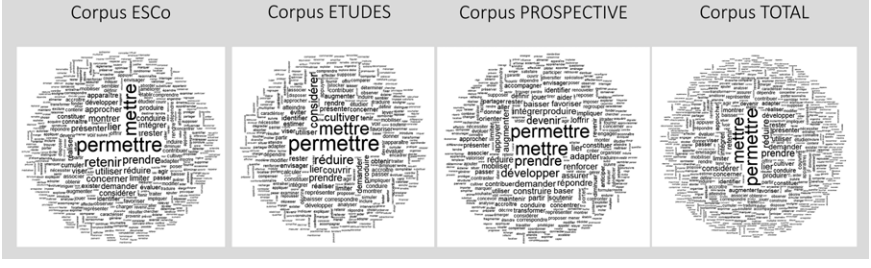


Figure 13 : Nuage de mots des verbes par type d'opération (tous les verbes)

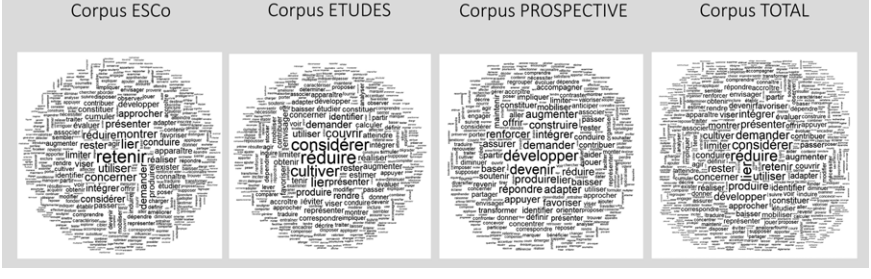


Figure 14 : Nuage de mots des verbes par type d'opération (sans verbes permettre, mettre et prendre)

4.2.2. Les classes de verbes (méthode Reinert)

L'application de la méthode Reinert sur les verbes aboutit à deux branches, la deuxième branche étant divisée en deux classes (*Figure 15*), avec des classes très distinctes (*Figure 16*). D'un côté (première branche), avec 48,9% des formes, nous trouvons une première classe de verbes décrivant globalement des « actions » comme favoriser, offrir, construire, concentrer, développer, viser, engager, intégrer, baser, orienter, devenir, aider, assurer, partager, créer.

De l'autre côté (deuxième branche), nous trouvons une deuxième classe, avec 33% des formes, avec des verbes décrivant globalement des « analyses » comme retenir, considérer, évaluer, estimer, observer, établir, comparer, lier, rendre, utiliser, montrer, correspondre, cumuler, réaliser, expliquer, étudier et une troisième classe, avec 18,1% des formes, avec des verbes décrivant globalement des « variations » (*augmenter, réduire, baisser, lever, repousser, modérer, compenser, multiplier*), des « connexions » (*représenter, situer, indiquer, dépendre, intervenir, éviter, provenir, libérer*) ou correspondant à une dimension « agricole » (*couvrir, stocker, semer, absorber*).

Ces trois classes se rattachent spécifiquement aux types d'opérations EPE : la classe des verbes des d'analyses est liée aux ESCo et aux études, celle des verbes de variation ou de connexion est liée aux études, celle des verbes d'action est liée aux prospectives (*Figure 17*).

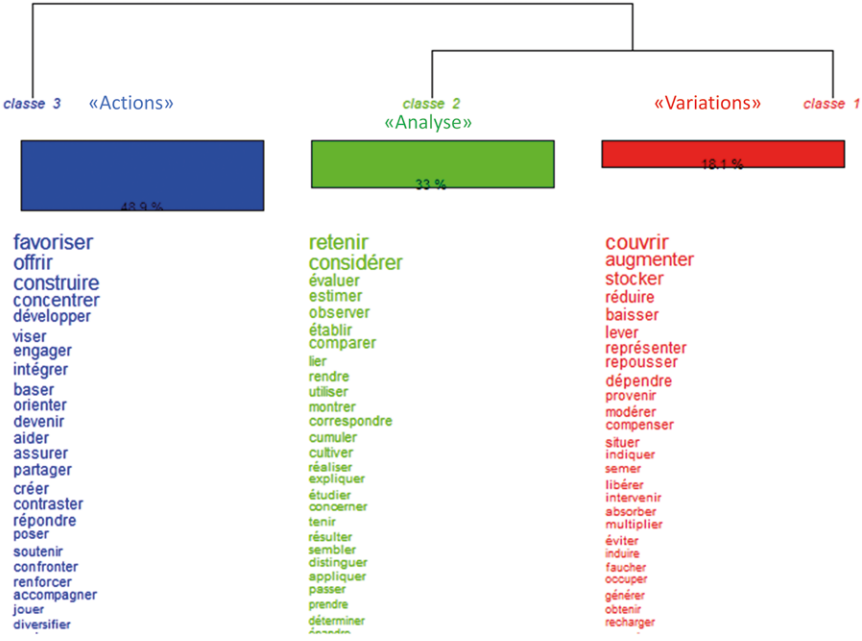


Figure 15 : Dendrogramme des classes de verbes obtenues par la méthode Reinert. Classification simple sur les segments de texte par Iramuteq. Corpus total. Nombre d'itérations à 10, taille de séparation de segments à 80, 98% de segments classés.

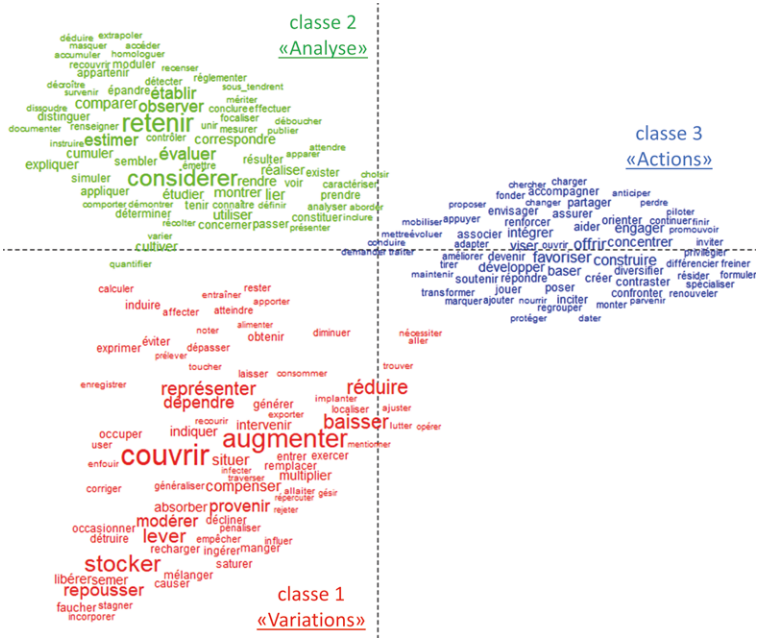


Figure 16 : Analyse Factorielle des Correspondances des classes de verbes obtenues par la méthode Reinert. Classification simple sur segments de texte des verbes par Iramuteq. Corpus total. Nombre d'itérations à 10, taille de séparation de segments à 80, 98% de segments classés.

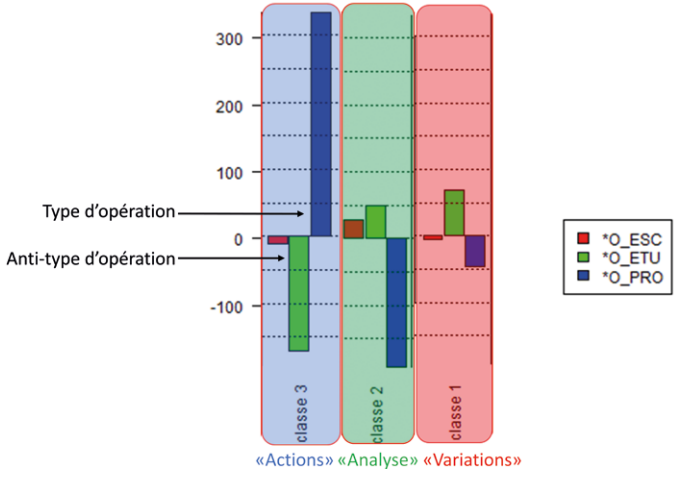


Figure 17 : Contribution des trois types d'opération aux classes de verbes obtenues par la méthode Reinert. Classification simple sur les segments de texte des verbes par Iramuteq. Corpus total. Nombre d'itérations à 10, taille de séparation de segments à 80, 98% de segments classés. Chi² Modalités de la variable « type d'opération ». Chi² comme l'unité de l'axe des ordonnées. Un Chi² très élevé/faible témoigne d'une forte/faible contribution d'un type d'opération à une classe.

4.2.3. Les verbes les plus souvent associés (analyses de similitudes)

Afin d'étudier la proximité et les relations entre les verbes, nous avons généré un arbre de similitudes (*Figure 18*). Deux communautés lexicales sont constituées autour des formes *mettre* et *permettre*. La première communauté construite autour de la forme *mettre* souligne son caractère plutôt actif avec des formes comme *développer, mobiliser, renforcer, favoriser, modifier, contribuer,* etc.

La deuxième communauté construite autour de la forme *permettre* renvoie au monde des concepts et des analyses notamment à travers les formes *considérer, étudier, analyser, identifier, caractériser, réaliser, traiter, estimer, décrire, évaluer, viser, appuyer, représenter, associer, assurer,* etc. Huit branches se développent à partir de la forme racine *permettre* avec des formes *réduire, lier, cultiver, demander, présenter, produire, prendre, retenir.*

On trouve ici deux plans grammaticaux qui identifient les discours du registre lexical des objets et ceux du registre des concepts.

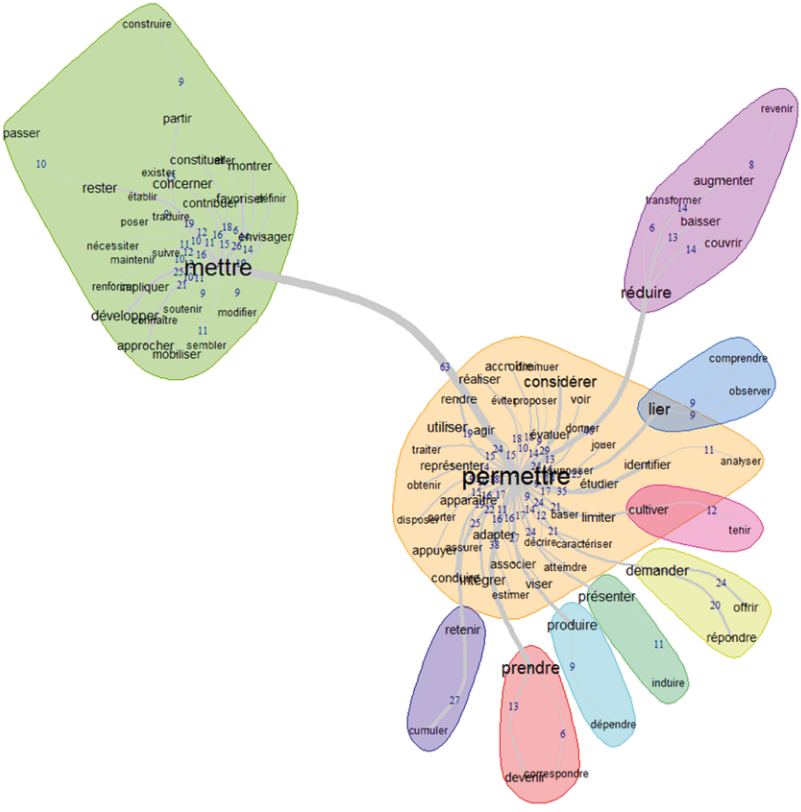


Figure 18 : Analyse des similitudes de verbes. Corpus total, taille de séparation de segments à 80, fréquence supérieure à 50

4.2.4. Conclusion

Les opérations EPE comportent à la fois un socle commun de verbes (*permettre*, *mettre* et *prendre*) qui décrivent un contexte ou qui font une référence à un débat public, des verbes d'action spécifiques aux prospectives, des verbes d'analyses spécifiques aux ESCo et des verbes de variations ou de connexions spécifiques aux études. Ces classes de verbes correspondent assez bien aux principaux attendus de chaque type d'opération : une prospective se projette dans l'avenir au travers de différentes actions, une ESCo correspond à une analyse de la littérature scientifique, une étude par le biais de simulations caractérise les variations du système concerné.

L'analyse de similitude permet de distinguer les discours sur la base des différences de registres lexicaux (celui des objets versus celui des concepts).

4.3. Les thématiques abordées : analyse des noms

Travailler sur les noms permet de mettre en évidence les thèmes traités par l'Inra depuis 20 ans. C'est donc là un ensemble de résultats importants pour observer le déploiement de ces opérations d'expertise au sens large sur de nombreux sujets marquants de cette période où la recherche agronomique a annoncé, pour elle-même et pour ses contributions, de nombreux changements et des objectifs de durabilité puis de transition agroécologique.

4.3.1. Les noms les plus cités (nuage de mots)

Les analyses sont présentées par type d'opération et pour le corpus total (*Figures 19 et 20*). Contrairement aux verbes, les noms les plus fréquents sont spécifiques à chaque type d'opération, mis à part le nom *production* qui est commun aux trois types d'opération, renvoyant souvent à la notion de production agricole. Les corpus d'ESCo et d'études se caractérisent par les noms relevant des deux domaines Agriculture et Environnement : *agriculture, culture, élevage* d'un côté, et *sol, eau, carbone* et *stockage* de l'autre, avec en plus pour les études les noms *scénario, niveau, effet* qui caractérisent la place accordée à la modélisation et aux simulations d'effets de nouvelles pratiques. Le corpus des prospectives est assez différent : outre le nom scénario qui domine le nuage, la plupart des noms les plus fréquents intègre une dimension renvoyant à une action humaine, comme *filière, acteur, développement, évolution, territoire*.

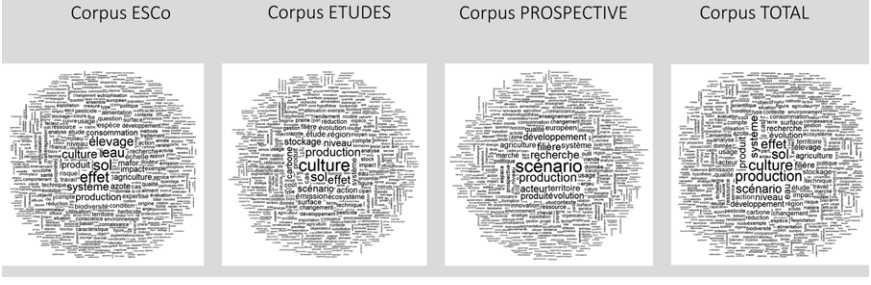


Figure 19 : Nuages de mots des noms par type d'opération

Mots-clés	CorpusTotal	ESCo	Etudes	Prospectives
culture	704	219	430	55
production	679	182	276	221
sol	626	290	313	23
effet	617	290	270	57
scénario	565	10	224	331
système	503	212	161	130
produit	451	165	140	146
niveau	423	125	222	76
eau	400	256	127	17
filière	372	56	138	178
élevage	333	243	31	59
azote	255	152	102	1
agriculture	341	146	69	126
stockage	294	65	217	12
carbone	262	49	202	11
région	262	35	176	51
émission	253	69	175	9
recherche	338	107	45	186
acteur	291	38	83	170
développement	349	93	97	159
évolution	365	75	142	148
territoire	270	97	36	137

Figure 20 : Tableau comparatif des 10 noms les plus cités par type d'opération (cases colorées)

4.3.2. Les classes de noms (méthode Reinert)

La classification Reinert aboutit à sept classes de noms. Le dendrogramme (*Figure 21*) indique une première branche de noms caractéristiques de l'eau et de l'azote, classe 2 dite « Eau et azote » (*eau, azote, mafor, apport, nitrate, épandage, phosphore*), de la culture, classe 1 dite « Culture » (*culture, herbicide, plante, cuivre, maïs, variété, résistance, vth, récolte, rendement*) et du carbone, classe 5 dite « Carbone » (*carbone, stockage, émission, stock, gaz à effet de serre, sol,* CO_2), trois classes relativement proches qui représentent une entrée agro-environnementale (*Figure 22*). La deuxième branche se subdivise en quatre sous-classes dont l'une comporte des noms caractéristiques des produits et des marchés agricoles, classe 7 dite « Produit et marché » (*produit, marché, consommateur, consommation, viande, qualité, aliment, propriété, filière*), l'autre contient des noms relevant essentiellement la recherche scientifique au sens large, classe 6 dite « Recherche scientifique » (*recherche, scientifique, connaissance, donnée, expertise, outil, Inra, question*) et les deux autres comportent des noms caractéristiques des scénarios d'un côté, classe 4 dite « Scénario » (*scénario, évolution, trajectoire, hypothèse, lande, horizon, gascogne, changement, famille, prospective, conséquence*) et des territoires de l'autre, classe 3 dite « Territoire » (*territoire, agriculteur, acteur, service, espace, écosystème, politique, ville, biodiversité, activité, action*).

L'Analyse Factorielle des Correspondances (*Figure 23*) met en évidence que chaque classe est spécifique à un type d'opération. Ainsi, la classe *carbone* est liée aux études, la classe *eau* et *azote* est liée aux ESCo, la classe *culture* est liée aux études et aux ESCo à la fois, les classes *produit* et *marché agricole* et *territoire* sont liées aux prospectives. Cette opposition reflète l'absence d'opérations EPE qui combinent au même niveau les questions de production agricole et de qualité des sols et des eaux au sens large.

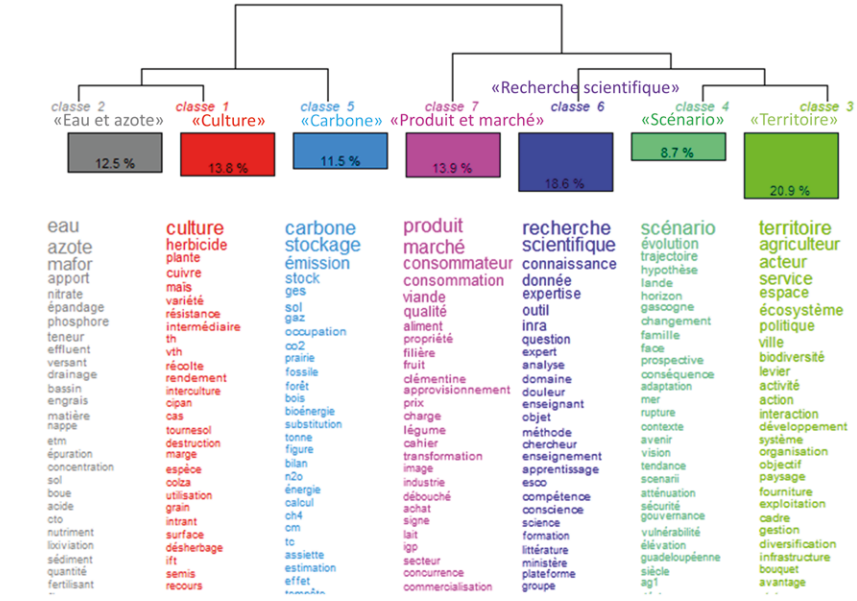


Figure 21 : Dendrogramme des classes de noms obtenues par la méthode Reinert. Classification simple sur les segments par Iramuteq. Corpus total. Nombre d'itérations de 14, taille de séparation de segments à 20, 93% de segments classés.

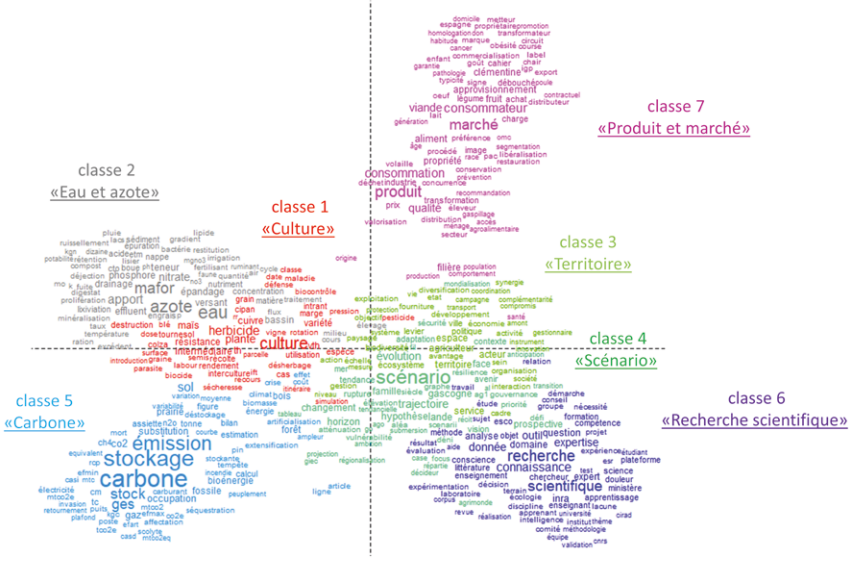


Figure 22 : Analyse Factorielle des Correspondances des classes de noms obtenues par la méthode Reinert. Classification simple sur les segments de texte des noms par Iramuteq. Corpus total. Nombre d'itérations de 14, taille de séparation de segments à 20, 93% de segments classés.

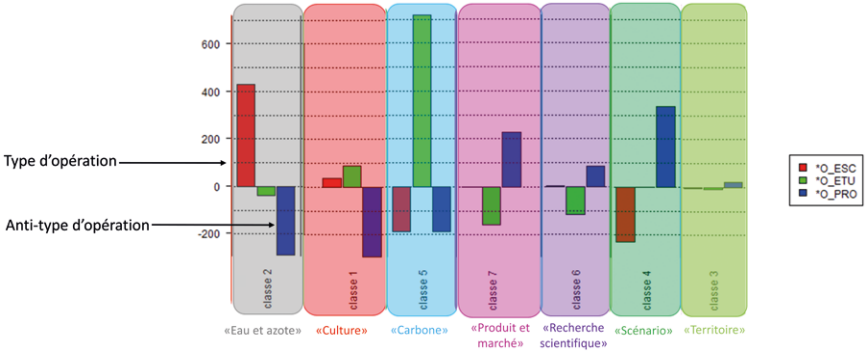


Figure 23 : Contribution des trois types d'opération aux classes de noms obtenues par la méthode Reinert. Classification simple sur les segments de texte des noms par Iramuteq. Corpus total. Nombre d'itérations de 14, taille de séparation de segments à 20, 93% de segments classés. Chi² Modalités de la variable « type d'opération ». Chi² comme l'unité de l'axe des ordonnées. Un Chi² très élevé/faible témoigne d'une forte/faible contribution d'un type d'opération à une classe.

La classe *scénario* est très liée aux prospectives, elle n'est pas du tout liée aux ESCo, en cohérence avec les caractéristiques de ces deux types d'opération. La classe *recherche scientifique* est principalement liée aux prospectives et elle n'est pas du tout liée aux études, alors même que ces notions sont présentes *a priori* dans les trois types d'opération : cela semble lié au poids de la prospective « Transition numérique et pratiques de recherche et d'enseignement supérieur en agronomie, environnement, alimentation et sciences vétérinaires à l'horizon 2040 » qui a concerné les conséquences du développement du numérique sur la recherche scientifique et ses pratiques (*Figure 24*).

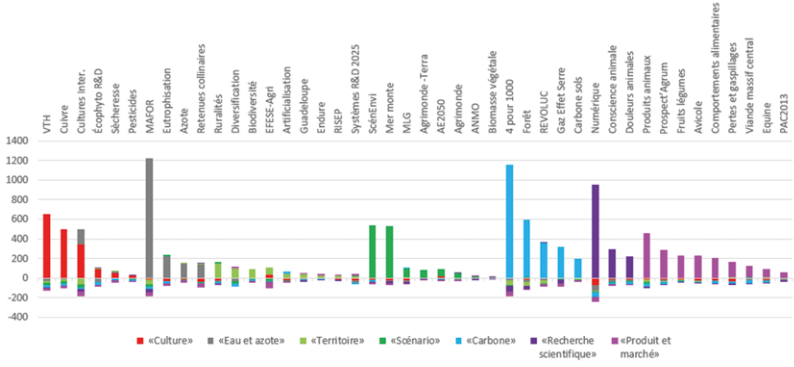


Figure 24 : Contribution de chaque opération aux classes de noms obtenues par la méthode Reinert. Classification simple sur les segments de texte des noms par Iramuteq. Corpus total. Nombre d'itérations de 14, taille de séparation de segments à 20, 93% de segments classés. Chi² Modalités de la variable « intitulé d'opération » (Annexe 1).

4.3.3. Les noms les plus souvent associés (analyses de similitudes)

L'analyse de similitudes nous permet d'affiner l'interprétation des résultats de la méthode Reinert. Pour obtenir un graphe lisible des analyses de similitudes, la fréquence est fixée à 150 pour être représentée. Les co-occurrences de formes associées sont regroupées au sein de trois communautés principales constituées autour des formes suivantes : *production, culture* et *sol* (*Figure 25*).
Ces trois communautés sont connectées au travers de la forme *culture*.

La première branche *production* est relative à l'agriculture ou à l'élevage, aux systèmes et aux scénarios d'évolution. La forme *scénario* dans la communauté principale production nous renvoie à la notion de prospective et la classe 4 dite « Scénario » de la méthode Reinert avec des formes comme *évolution, donnée, modèle, connaissances scientifiques*. Une branche secondaire se développe à partir de cette forme *production*, notamment avec des formes *produit* et *filière* qui retrouvent le lexique économique avec des formes comme *consommation, qualité, acteur, territoire, échelle,* etc. On retrouve ici la classe 7 dite « Produit et marché » et la classe 3 dite « Territoire ».

La branche *production* est liée à la deuxième branche *culture* qui constitue une communauté autour des formes *surface, espèce, agriculteur, zone, gestion, écosystème*. Les éléments minéraux sont représentés dans la branche *culture via* des formes *pesticide, utilisation* et *réduction des risques* qui nous renvoient à la classe 2 dite « Eau et azote » et à la classe 1 dite « Culture ». Dans la branche *culture* il existe une sous-branche *effet* qui porte sur des impacts environnementaux, sanitaires ou alimentaires, *émission, changement, évaluation, analyse,* et une autre sous-branche *recherche* qui porte sur le *développement* et l'*agriculture, alimentation, biodiversité, question.* Nous retrouvons donc la classe 6 dite « Recherche scientifique ».

La branche *culture* est liée à une autre communauté, celle du *sol*. Dans la branche *sol*, se retrouve la classe 5 dite « Carbone » de la méthode Reinert avec des formes associées comme *stockage de carbone, usage de terre* d'un côté et la classe 2 dite « Eau et azote » avec des formes associées comme *ressources, eau, type* d'un autre côté.

Cette mise à plat des proximités directes à l'échelle du corpus traduit d'une autre façon l'existence de grande polarité d'objets : 1) la *production* de commodités et des *systèmes* qui les portent dans un rapport à l'action humaine (*consommation, qualité, filière, acteurs* et *marché*) ; 2) les cultures, en lien avec le changement climatique et ses effets (*rendement, surface, émission, impact, protection*), avec la question de l'usage des pesticides, et avec la recherche scientifique (*recherche, enseignement*); 3) le *sol* pris comme substrat ou ressources avec l'importance de la carbonisation du sol et de la question de la circulation de l'eau. On retrouve ici bien les trois domaines qui décrivent l'Inra : agriculture, alimentation, environnement.

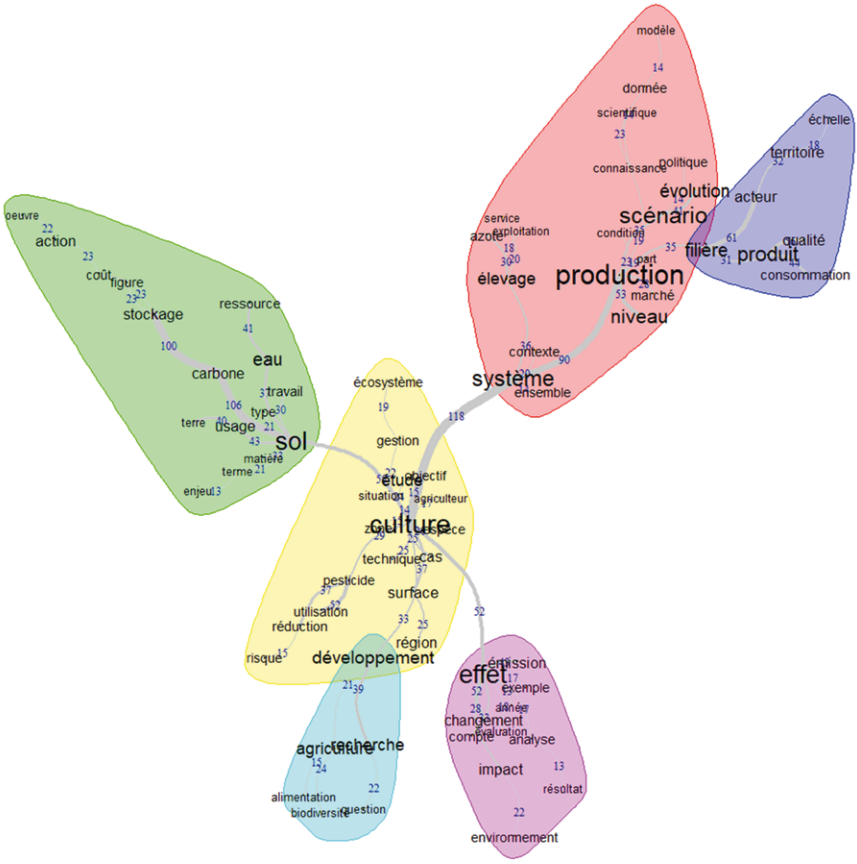


Figure 25 : Analyses de similitudes des noms . Corpus total, taille de séparation de segments à 20, fréquence supérieure à 150

4.3.4. Conclusion

L'analyse des noms des résumés montre une nette différenciation thématique entre les trois types d'opération : d'un côté, les thèmes plutôt liés à la qualité des sols et des eaux pour les études et ESCo, respectivement, de l'autre, les thèmes plutôt liés aux produits et marchés agricoles pour les prospectives. Cela peut traduire une difficulté à prendre en compte, dans une même opération, les dimensions productives et environnementales.

4.4. Conclusions des analyses réalisées avec Iramuteq

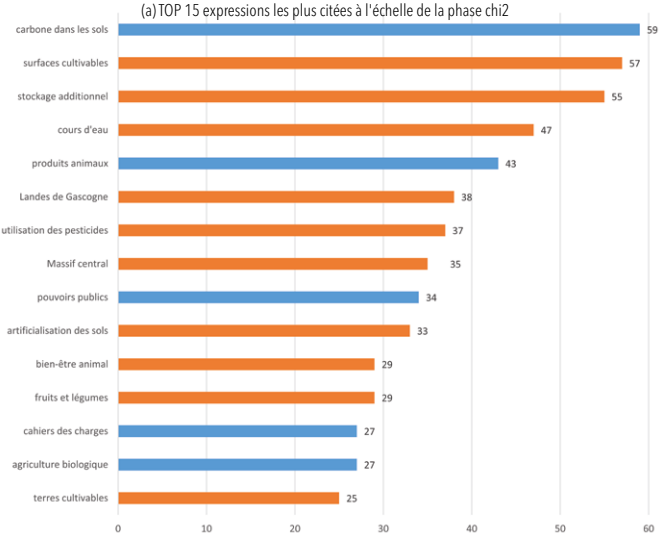
Les analyses conduites avec Iramuteq nous montrent à la fois :

- un socle commun dans la rédaction des résumés : les verbes les plus utilisés sont *permettre, mettre* et *prendre*, trois verbes très utilisés dans les écrits scientifiques en général,
- des verbes spécifiques aux trois types d'opération, en cohérence avec leurs caractéristiques : analyses pour les ESCo, variations pour les études, actions pour les prospectives,
- un vocabulaire un peu plus fin pour les prospectives, en lien probablement avec la façon de se projeter dans un avenir incertain,
- des thématiques relativement spécifiques aux trois types d'opération, traduisant la non-prise en compte, jusqu'à maintenant de problématiques systémiques intégrant les trois domaines Inra : agriculture, alimentation et environnement.

5. Analyse des expressions avec CorTexT Manager

5.1. Choix du nombre d'expressions à analyser

En utilisant une mesure du Chi^2 pour extraire les expressions de 2 ou 3 mots à l'échelle de la phrase (pas de monogramme), nous proposons ci-dessous une première approche des 15 expressions les plus présentes (TOP 15). Ces expressions sont triées soit selon le nombre de citations dans l'ensemble des documents, soit selon le nombre de documents dans lesquels elles sont citées. Nous obtenons ainsi les expressions les plus fréquentes (*Figure 26a*) ou les plus partagées (*Figure 26b*). Les expressions à la fois les plus citées et les plus partagées sont : *carbone dans les sols, produits animaux, pouvoirs publics, agriculture biologique et cahier des charges.*



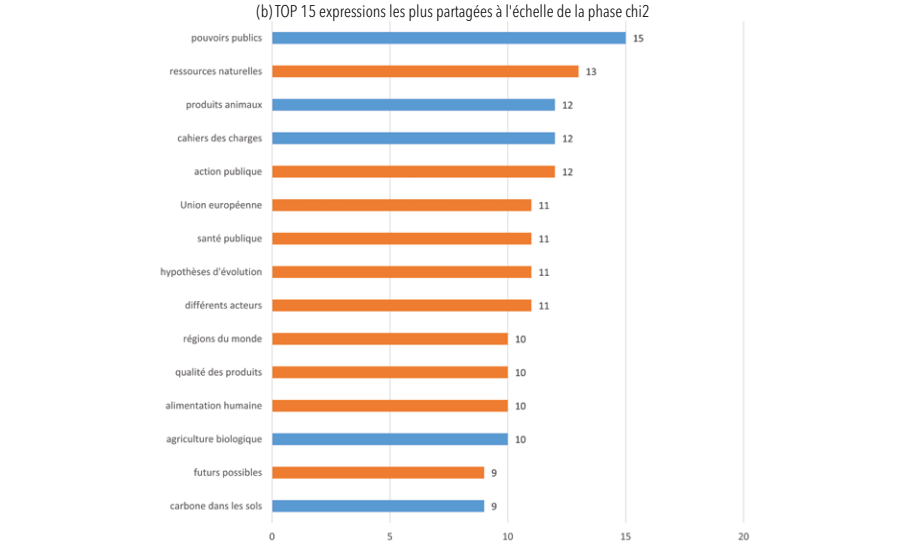
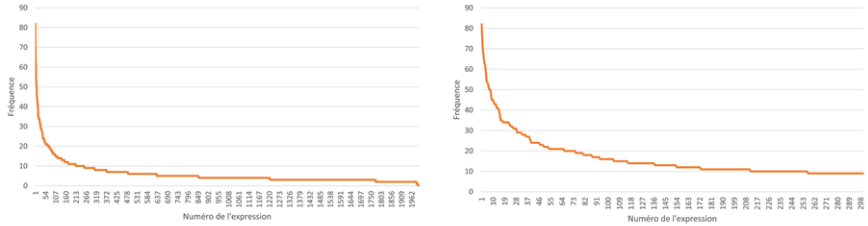


Figure 26 : Les 15 expressions extraites par la mesure de Chi² les plus citées (a) ou les plus partagées (b) (en rouge les expressions communes)

Pour choisir le nombre optimal des expressions extraites par CorTexT, nous avons produit la courbe de distribution des expressions (*Figure 27*). Aucun seuil dans la traîne de la courbe de distribution n'apparaît évident à définir, probablement à cause de la grande hétérogénéité thématique et du faible nombre de documents. Il nous faut donc trouver un compromis entre différents critères pour trouver le nombre optimal d'expressions exportées : la prise en compte de l'ensemble des 44 opérations EPE dans l'extraction terminologique et la lisibilité des cartes de clusters produites. Au final, nous avons privilégié l'extraction terminologique des 75 premières expressions extraites (TOP75 ce qui correspond à une fréquence minimale de 20).



(a) TOP2000 (b) TOP300

Figure 27 : Courbe de distribution des 2000 (a) ou 300 (b) expressions extraites par la mesure de Chi² les plus citées

5.2. Co-occurrence lexicale des expressions nominales

Nous présentons en *Figure 28*, la clusterisation des 75 premières expressions extraites au niveau de phrase.

La lecture du graphe consiste à observer le paysage des clusters, les relations entre les nœuds et la taille des nœuds. Pour chaque cluster les deux expressions en gras sont celles qui sont les plus reliées aux autres (elles ont la plus forte centralité dans ce cluster). La taille d'un nœud (triangle) est proportionnelle à la fréquence de l'expression. L'épaisseur du trait reliant deux expressions est proportionnelle au nombre de co-occurrences des deux expressions. La surface d'un cluster est proportionnelle au nombre de documents qui contribuent à créer ce cluster. Pour chaque cluster les opérations les plus spécifiques (au sens de Chi^2) sont positionnées en variable ajoutée.

Notons que le script *Network Analysis* de CorTexT produit des cartes en 2D sous forme de carte de clusters dont le positionnement à plat est tiré d'un calcul d'optimisation d'encombrement. Reste que parfois, deux clusters qui semblent superposés, peuvent en réalité être distants l'un de l'autre : il est alors essentiel de repérer les liens effectifs entre les nœuds de ces clusters pour savoir s'ils sont liés effectivement.

La clusterisation CorTexT met en évidence huit clusters : un premier cluster traite des terres cultivables et la demande alimentaire, un deuxième cluster assez proche est relatif à la qualité des produits animaux et au bien-être animal. A l'opposé, trois clusters traitent du stockage du carbone dans les sols ou de sa séquestration dans le bois, et des émissions de gaz à effet de serre azotés. Entre ces deux groupes, trois clusters traitent des acteurs au sens large, qu'ils soient de la société civile, publics ou économiques en lien avec les territoires. Le cluster relatif à la société civile, qui inclut les questions numériques, est uniquement relié à des opérations de prospective. Les clusters relatifs au carbone et aux émissions de gaz à effet de serre sont uniquement reliés à des études et à une opération d'expertise scientifique collective.

Le nombre de liens d'une expression avec toutes les autres détermine la taille du cercle qui est représenté. Les 10 expressions avec le poids le plus élevé sont *surfaces cultivables, carbone dans les sols, stockage additionnel, produits animaux, terres cultivables, outils numériques, demande alimentaire, dépendance aux importations, surplus de terres, intelligence artificielle*.

On retrouve les thématiques affichées par l'institution : le changement climatique, la filière, l'usage des terres, la politique agricole et le système de culture. En revanche, des thématiques plus émergentes comme la nutrition et la biodiversité n'apparaissent pas à ce seuil de calcul du partitionnement.

Il est intéressant de regarder plus précisément ce que sont les expressions qui relient les clusters entre eux. Et donc de considérer l'effet de la mesure de centralité dans le calcul des clusters réalisés sous CorTexT. Et pour cela on recourt à la visualisation proposée par le logiciel en ligne Gephi[16].

On a utilisé les scores de centralité (*Figure 29*) pour identifier les expressions qui engendrent la plus grande « circulation de l'information ». De ce point de vue, un nœud est central lorsqu'il est le point de

16 Gephi est un logiciel libre d'analyse et de visualisation de réseaux https://gephi.org/

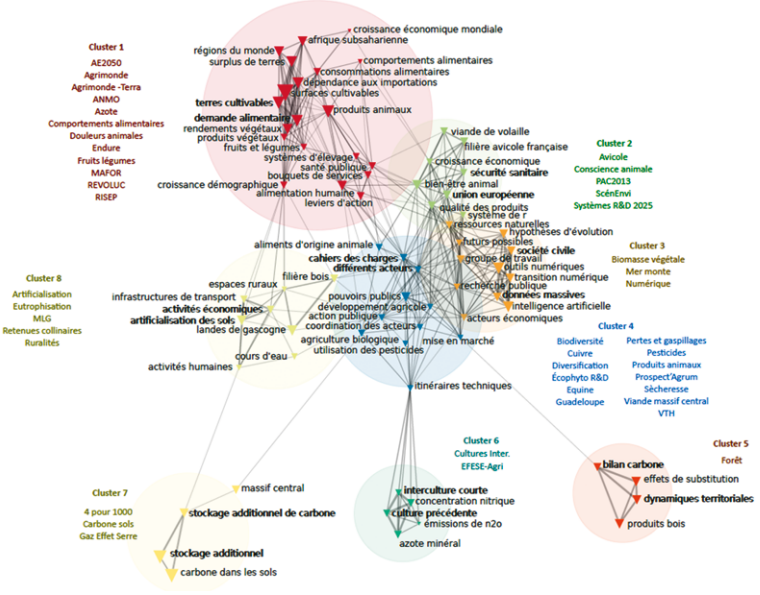


Figure 28 : Clusterisation (Louvain) lexicale des expressions nominales

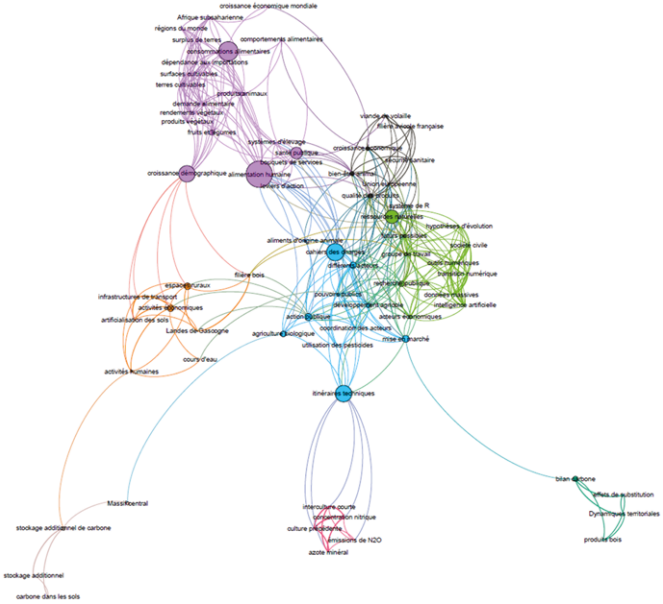


Figure 29 : Clusterisation (Louvain) lexicale des expressions nominales (la taille de nœuds en fonction de leur centralité d'intermédiarité (betweeness))

passage obligé que les plus courts chemins doivent emprunter pour se rejoindre. Les 10 expressions les plus centrales sont *alimentation humaine, consommations alimentaires, cahiers des charges, itinéraires techniques, croissance démographique, ressources naturelles, santé publique, action publique, mise en marché, espaces ruraux.*

Il est pertinent aussi d'observer la contribution des opérations à la création des clusters.

Pour cela, la *Figure 30* présente la matrice de contingence croisant les clusters et le type d'opération pour mesurer, par une valeur de Chi^2 (échelle à droite de la figure), la spécificité de contribution des opérations à tel ou tel cluster (spécificité forte contributive en rouge). Ainsi, la qualité des produits animaux et le bien-être animal en lien avec l'Union Européenne et les questions sanitaires (cluster 2), les acteurs de la société civile (cluster 3), sont des thèmes fortement présents dans les prospectives. Les clusters relatifs au carbone et aux émissions de gaz à effet de serre (clusters 5, 6 et 7) sont emblématiques des études. Les clusters relatifs aux terres cultivables et la demande alimentaire (cluster 1), aux acteurs publics (cluster 4) sont emblématiques des ESCo. Les acteurs économiques (cluster 8) sont un thème fortement présent dans les ESCo, et dans une moindre mesure dans les prospectives.

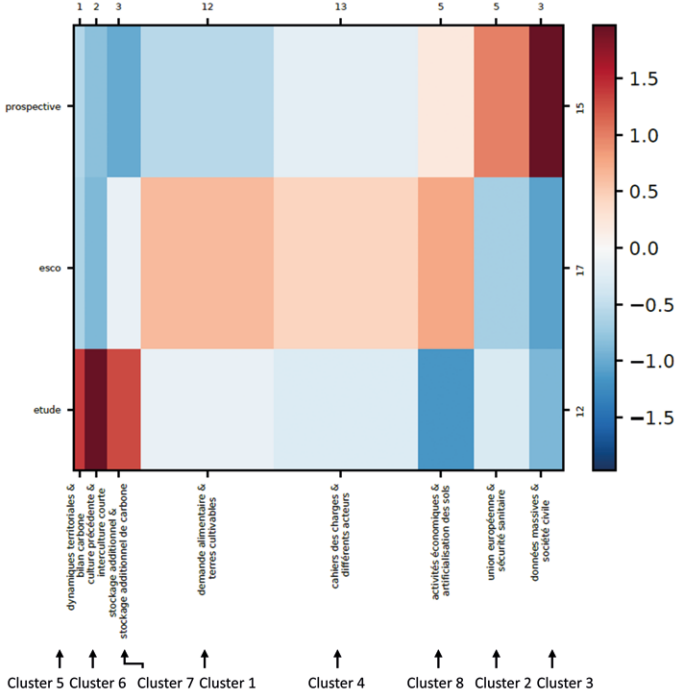


Figure 30 : Matrice de contingence entre le type d'opération et les thématiques des opérations EPE représentées par cluster (8 clusters)

5.3. Evolution temporelle des expressions nominales

Par ailleurs, nous nous sommes posé la question suivante : y-a-t-il une évolution temporelle des thématiques des opérations EPE (*Figure 31*). Pour trouver la répartition optimale des périodes nous avons lancé le script *Period Detector*, alors que CorTexT ne propose pas un découpage optimal de périodes.

Le script *Period Detector* ne renvoie qu'une seule période, chaque année est indépendante des autres d'un point de vue thématique. Cela signifie qu'il n'y a pas, entre 2002 et 2020, une période continue durant laquelle les études réalisées porteraient sur des thématiques assez proches, puis une période suivante où les études concerneraient une autre thématique.

Nous constatons donc l'absence de découpage temporel parmi les opérations EPE à cause de manque d'évolution continue sur la durée de 2002 à 2020.

Néanmoins, sur la *Figure 32* nous voyons le classement des expressions les plus occurrents à l'échelle de la phrase dans le temps.
Pour la plupart des expressions on ne voit pas la continuité durable pendant tous les 20 ans sauf des *surfaces cultivables*.

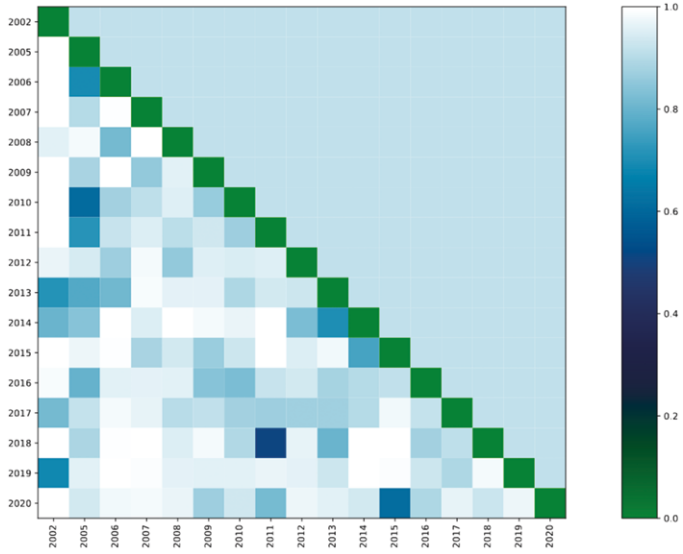


Figure 31 : Découpage de périodes des expressions à l'échelle de la phrase par CorTexT Manager

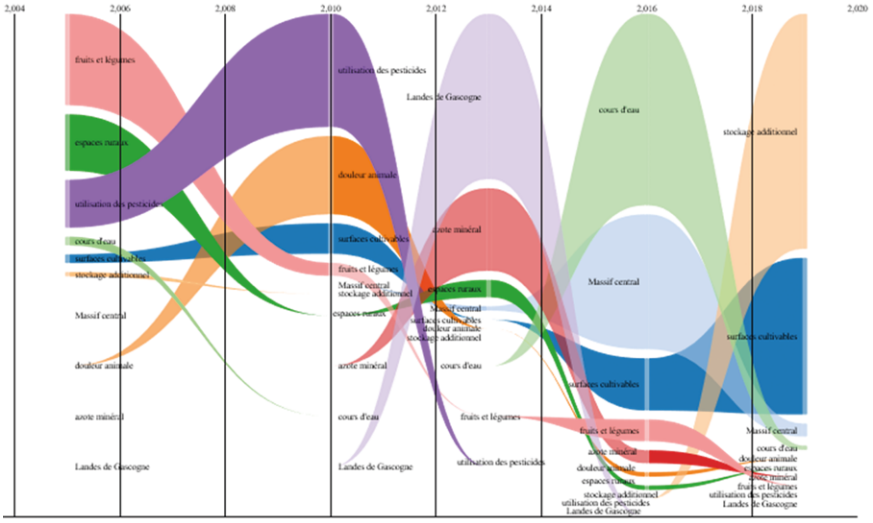


Figure 32 : Evolution dans le temps du classement des expressions les plus occurrentes (les 20 années sont découpées en cinq sous périodes par CorTexT Manager)

5.4. Conclusions des analyses réalisées avec CorTexT Manager

Les analyses réalisées avec CorTexT, à l'échelle du document ou de la phrase, mettent en évidence des thématiques traitées par les 44 opérations EPE qui relèvent principalement du carbone dans les sols, et des activités agricoles au sens large avec une spécificité liée au type d'opération : le carbone est un thème majeur des études, les acteurs économiques sont quasi-spécifiques aux prospectives. Par ailleurs, nous constatons l'absence de découpage temporel parmi les opérations EPE à cause du manque d'évolution continue sur la durée de 2002 à 2020.

6. Conclusions

6.1. Le mode opératoire de la Master Class

1- L'implication des agents de la DEPE dans la Master Class et dans la fabrication des données, étape toujours cruciale et délicate, a montré la capacité du service à se mobiliser sur des activités transversales à la fois réflexives et ouvertes à l'acquisition de nouvelles compétences. La connaissance des dossiers est également un atout pour le travail d'interprétation des résultats au sein d'un type d'analyse ou entre les analyses, travail qui invite à établir les conclusions sur une flexibilité interprétative préalable féconde.

2- L'intérêt de mobiliser deux méthodes d'analyse. Nous avons mobilisé deux approches totalement différentes qui sont à la base d'Iramuteq et de CorTexT. Iramuteq est basé sur le principe de la recherche d'opposition dans la construction des classes, alors que CorTexT est basé sur la notion de similitude des mots pour former les clusters, tout en ayant l'avantage de permettre une extraction terminologique.

6.2. Conclusions issues des statistiques descriptives des opérations

6.2.1. Façons d'écrire

Chaque type d'opération est associé à une façon d'écrire : on découvre l'existence d'une certaine typicité des rapports aux réels en fonction des opérations. Trois classes se rattachent spécifiquement aux trois types d'opérations. Ainsi, les ESCo se caractérisent plutôt par les verbes d'analyse, les études recourent plutôt à des verbes de variations et des problématiques agraires, tandis que les prospectives utilisent plutôt des verbes d'action.

Le mode d'écriture des résumés des opérations d'expertise scientifique collective, d'étude et de prospective, au travers des verbes utilisés, apparait relativement spécifique du type d'opération et cohérent avec les attendus de chaque type d'opération : cela témoigne du professionnalisme rédactionnel des chefs de projet et de la pertinence des outils d'analyse textuelle mobilisés.

6.2.2. Les résultats de l'analyse thématique

3- Trois grandes thématiques émergent de 20 années d'opération : (1) la demande et l'offre en produits agricoles et les systèmes de culture ou d'élevage, thématiques partagées par les trois types d'opération, (2) l'atténuation du changement climatique, thématique principalement partagée par les expertises scientifiques collectives et par les études, (3) les acteurs, les activités humaines et économiques, thématiques principalement présentes dans les prospectives.
Ce sont trois thématiques qui sont centrales pour INRAE : l'Inra et maintenant INRAE est fortement reconnu sur celles-ci. Cela laisse peut-être une moindre place à des thématiques plus émergentes comme la biodiversité, la bioéconomie ou la santé par exemple.
Une première thématique ressort de l'ensemble des analyses réalisées, c'est celle du carbone dans les sols ; elle est fortement liée aux études. C'est un type d'opération mis en place depuis une dizaine d'années et qui fait souvent appel à la modélisation du fonctionnement des agroécosystèmes, avec des modèles comme STICS ou PASIM pour lesquels les cycles biogéochimiques C-N et hydrologiques sont essentiels. On peut remarquer que c'est également le thème de la première ESCo réalisée par l'Inra au début des années 2000.

La deuxième thématique qui ressort est relative aux acteurs économiques, aux activités et filières agricoles, elle est fortement liée aux prospectives.

Dans les deux cas, les thématiques sont relativement spécifiques : pas/peu de carbone et de climat dans les prospectives, pas/peu d'acteurs et de produits agricoles dans les études. Il y a un premier enjeu très important pour l'avenir d'intégrer ces deux thématiques pour couvrir les trois domaines Inra.
Certains écosystèmes ont été peu travaillés, c'est le cas des forêts ou des prairies qui ressortent très peu des analyses réalisées. C'est également le cas pour des thèmes comme la biodiversité, la santé, la ville, la bioéconomie. Il y a un troisième enjeu qui est à la fois de traiter des sujets venant alimenter la transition agroécologique des systèmes agricoles et des systèmes alimentaires, et de se faire reconnaître auprès de nos commanditaires comme un établissement porteur d'expertise sur des sujets émergents.

4- La spécificité thématique par type d'opération témoigne de l'absence d'opérations embrassant l'ensemble du périmètre Inra, en ayant ici à l'esprit qu'il s'agit d'opérations qui résultent d'une commande. Globalement, (1) les questions de changement climatique et d'environnement *s.l.* sont peu abordées dans les opérations de prospective, (2) les questions économiques et humaines *s.l.* sont peu abordées dans les opérations d'expertise scientifique collective et d'étude.

6.2.3. Des surprises et des constats

5- Les questions de productivité et d'efficacité des systèmes agricoles et celles liées à la composition du rendement sont peu visibles dans les analyses, et cela constitue une surprise. Cela pourrait être lié à la priorité donnée au traitement des questionnements sociétaux (environnement, sanitaire, condition animale) qui ont marqué le renouvellement du positionnement de l'Inra face à la demande politique depuis le début des années 2000. Ainsi on notera le nombre relativement élevé d'opérations rattachées au domaine de l'environnement au détriment des questions plus classiques concernant les performances productives des systèmes de culture ou d'élevage.

6- La place des sciences humaines et sociales interroge. Elles sont essentiellement présentes au travers des déterminants économiques des activités agricoles (*e.g.* analyse de la demande, certification des productions) ou de la prise en compte des jeux d'acteurs dans les territoires. On note une faible place donnée à l'économie de l'environnement, ainsi qu'aux approches sociologiques, juridiques et politiques sur les grands enjeux contemporains. Les raisons qui conduisent à ce déficit d'interdisciplinarité dans la production de connaissances finalisées par les demandes faites à la DEPE mériteraient d'être approfondies.

6.3. Recommandations

7- Au final, il nous semble que ces résultats appuient la stratégie INRAE2030 proposant de développer des approches plus « holistiques » et d'afficher des priorités scientifiques qui élargissent et renouvellent son champ d'activité et d'attractivité. Ils interrogent également la place de certaines disciplines ou de travaux scientifiques fonciers et classiques au sein de l'institut dont la contribution aux opérations EPE ne ressortent pas de façon forte. Il y a donc un véritable enjeu qui est à la fois de traiter des sujets venant alimenter les transitions agroécologiques, climatiques, énergétiques des systèmes alimentaires, et de se faire reconnaître auprès de nos commanditaires comme un établissement porteur d'expertise renouvelée, en capacité à également remobiliser des compétences foncières et à mieux porter l'interdisciplinarité dans le travail d'expertise scientifique.

Il faudrait envisager la mobilisation des nouveaux partenaires pour les futures opérations EPE afin de diversifier les thématiques déjà abordées depuis 20 ans.

Il faudrait mobiliser davantage d'experts étrangers et systématiser leur présence dans les comités d'experts.

Les nouvelles opérations EPE en cours, notamment l'étude « TempAG » et la Prospective « Agriculture européenne sans pesticides », sont réalisées par des comités d'experts français et étrangers anglophones. Cela avait déjà été le cas pour la prospective Agrimonde-Terra. C'est un axe d'évolution important pour la DEPE, qui passe sans doute par des partenariats avec des établissements européens ou des organismes internationaux.

ANNEXES

Annexe 1. Fiche de description des opérations

Type d'opération	Intitulé abrégé	Intitulé complet	Année	Commanditaires
ESCO	Carbone sols	Contribution à la lutte contre l'effet de serre - Stocker du carbone dans les sols agricoles de France ?	2002	Ministère en charge de l'Environnement
ESCO	Pesticides	Pesticides, agriculture et environnement, Réduire l'utilisation des pesticides et en limiter les impacts	2005	Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ESCO	Sécheresse	Sécheresse et agriculture - Réduire la vulnérabilité de l'agriculture à un risque accru de manque d'eau	2006	Ministère en charge de l'Agriculture
ESCO	Fruits légumes	Les fruits et légumes dans l'alimentation, Enjeux et déterminants de la consommation	2007	Ministère en charge de l'Alimentation
ESCO	Biodiversité	Agriculture et biodiversité	2008	Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ESCO	Douleurs animales	Douleurs animales	2009	Ministère en charge de l'Agriculture, Ministère en charge de la Recherche
ESCO	Comportements alimentaires	Les comportements alimentaires	2010	Ministère en charge de l'Alimentation
ESCO	VTH	Variétés végétales tolérantes aux herbicides	2011	Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ESCO	Azote	Les flux d'azote liés aux élevages	2012	Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ESCO	MAFOR	Mafor Valorisation agricole des effluents, boues et déchets organiques	2014	Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ESCO	Retenues collinaires	Impact cumulé des retenues d'eau sur le milieu aquatique	2016	Ministère en charge de l'Environnement
ESCO	RISEP	Rôles, impacts environnementaux, économiques et sociaux, et services rendus par les élevages	2016	Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture, Ademe
ESCO	Conscience animale	La conscience animale	2017	EFSA
ESCO	Eutrophisation	Eutrophisation : causes, mécanismes, conséquences Eutrophisation : causes, mécanismes, conséquences et prédictibilité	2017	Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ESCO	Artificialisation	Artificialisation des sols Déterminants, impacts et leviers d'action	2017	Ademe, Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ESCO	Cuivre	Peut-on se passer du cuivre en protection des cultures biologiques ?	2018	ITAB, INRA-métaprogramme SMaCH
ESCO	Produits animaux	Qualité des aliments d'origine animale	2020	Ministère en charge de l'Alimentation
ETUDE	Écophyto R&D	Écophyto R&D -	2010	Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ETUDE	Cultures Inter.	Réduire les fuites de nitrate au moyen de cultures intermédiaires	2012	Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ETUDE	Diversification	Diversification des cultures	2013	Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ETUDE	Gaz Effet Serre	Réduire les émissions de gaz à effet de serre du secteur agricole en France	2013	Ademe, Ministère en charge de l'Environnement, Ministère en charge de l'Agriculture
ETUDE	ANMO	Système alimentaire Afrique du Nord/Moyen Orient à l'horizon 2050, vers une dépendance accrue aux	2015	Pluriagri
ETUDE	Pertes et gaspillages	Réduire les pertes et gaspillages d'aliments dans un monde de plus en plus urbanisé	2016	INRA-Direction Scientifique Alimentation-Bioéconomie
ETUDE	REVOLUC	CAS : Changements 2017 des sols et évaluation environnementale	2017	Ademe, Ministère en charge de l'Agriculture
ETUDE	Forêt	Forêts : Quel rôle pour les forêts et filières forêt-bois françaises dans l'atténuation du changement	2017	Ministère en charge de l'Agriculture
ETUDE	EFESE-Agri	EFESE-Agroécosystèmes Évaluation des services écosystémiques rendus par les agroécosystèmes	2017	Ministère en charge de l'Environnement
ETUDE	4 pour 1000	4 pour mille	2019	Ademe, Ministère en charge de l'Agriculture
ETUDE	AE2050	Agricultures européennes à l'horizon 2050	2020	PluriAgri
PROSPECTIVE	PAC2013	Agriculture 2013	2007	INRA, Crédit agricole, Groupama
PROSPECTIVE	Ruralités	Les nouvelles ruralités à l'horizon 2030	2008	INRA
PROSPECTIVE	Agrimonde	Agrimonde - Scénarios et défis pour nourrir le monde en 2050	2009	INRA, CIRAD
PROSPECTIVE	Avicole	La filière avicole française à l'horizon 2025	2009	INRA, ITAVI
PROSPECTIVE	Biomasse végétale	Usages non alimentaires de la biomasse végétale à l'horizon 2050	2010	ANR
PROSPECTIVE	Endure	Endure : la protection des cultures en Europe à l'horizon 2030	2010	Reseau Endure
PROSPECTIVE	MLG	Le massif des Landes de Gascogne à l'horizon 2050	2012	Conseil régional d'Aquitaine
PROSPECTIVE	Equine	La -filière équine française à l'horizon 2030	2012	IFCE
PROSPECTIVE	Agrimonde -Terra	Agrimonde-Terra	2016	INRA, CIRAD
PROSPECTIVE	Viande massif central	Filières viande de ruminants dans le Massif Central	2016	CGET
PROSPECTIVE	Guadeloupe	Agriculture guadeloupéenne à l'horizon 2040	2016	Chambre d'Agriculture de la Guadeloupe
PROSPECTIVE	ScénEnvi	ScénEnvi : Visions du futur et environnement	2017	Allenvi-GT Prospective
PROSPECTIVE	Systèmes R&D 2025	Recherche, innovation et développement en agriculture : quel(s) avenir(s) ?	2017	Ministère en charge de l'Agriculture
PROSPECTIVE	Prospect'Agrum	Prospect'Agrum - Des visions d'avenir sur la filière agrumicole corse	2018	Ministère en charge de l'Agriculture
PROSPECTIVE	Mer monte	La mer monte	2019	Allenvi-GT Prospective
PROSPECTIVE	Numérique	La transition numérique	2019	INRA, Agreenium

Figure 33 : Fiche de description des opérations

Annexe 2. Préparation des données

Les analyses avec Iramuteq et sous CorTexT des documents des opérations EPE (résumé, synthèse, rapport) nécessitent de les transformer de manière homogène en format (.txt) à partir de format Word (.docx) ou de format (.pdf), à transformer au préalable en format Word (.docx). La procédure comporte quatre étapes.

1. Homogénéisation des fichiers

Chaque fichier est à renommer selon les règles de dénomination suivantes : DOCUMENT_OPERATION_TITRE_ANNEE_NUMERO-ABSOLU_NUMERO-RELATIF.docx

- DOCUMENT: pour le type de document, soit résumé, soit synthèse, soit rapport scientifique
- OPERATION : pour le type d'opération, soit ESCo, soit ETUDE, soit PROSPECTIVE
- TITRE : titre abrégé de l'opération en une quinzaine de caractères maximum, par exemple « MER_MONTE »
- ANNEE : année de publication des documents (année du colloque de restitution dans la plupart des cas)
- NUMERO-ABSOLU : Rang de réalisation de l'opération parmi toutes les opérations EPE depuis le début des années 2000
- NUMERO-RELATIF : Rang de réalisation de l'opération parmi le type d'opération considéré
- Tout en majuscule / Pas de minuscule
- Pas d'espace, séparer les termes par des tirets bas (« _ »)

A titre d'exemple : Résumé_PROSPECTIVE_MER_MONTE_2019_40_14(.docx), correspond à un résumé d'une prospective, dont l'intitulé était « La mer monte », qui a été conclu en 2019, notamment la 40e opération parmi toutes les opérations EPE ou la 14e prospective parmi les prospectives menées à la DEPE.
Avant d'enregistrer des fichiers en format (.txt) il faut parcourir le fichier Word afin de bien vérifier et corriger éventuellement les trois points suivants :

Point 1 = vérifier l'absence

- D'une première lettre du paragraphe reconnue comme un symbole.
- D'un saut de ligne mal placée au milieu de la phrase ou de l'intitulé (sinon une phrase ne sera pas reconnue comme une phrase unique).
- D'un trait d'union au milieu du mot.
- D'une liste à puces.
- D'un point à la fin de phrase.
- D'un titre et d'un sous-titre de l'opération en début de texte.

Point 2 = supprimer

- Les sources (Sources : FAOStat & GlobAgri-Pluriagri).
- Les dates (Mars 2017).
- La version (Version du 20/06/2017).
- Le type d'opération (Résumé de l'étude).
- Les encadrés de type « Pour en savoir plus ».
- Les encadrés de type « Organisation de l'étude » ou « Méthode DEPE ».
- Les notes de bas de page.
- Les tableaux.

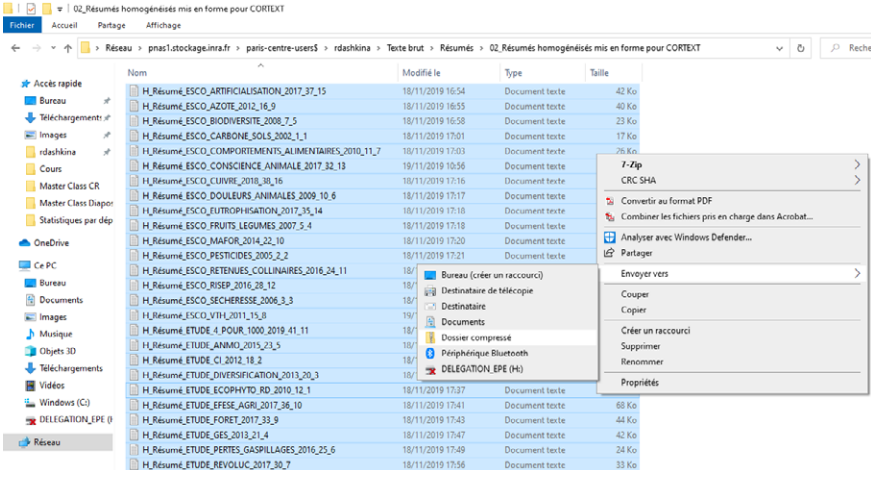
Point 3 = conserver

- Les intertitres présents dans le corps du document, mais supprimer les numéros qui précèdent les titres dans certains documents.
- Les titres des figures, mais supprimer le mot « Figure 1… » et les éléments associés (la clef de lecture, la source).
- Les autres encadrés, y compris les encadrés « méthode » relatifs spécifiquement au travail mis en oeuvre dans l'opération (par exemple un encadré sur un modèle employé spécifiquement dans telle opération), car ils portent sur le contenu spécifique de l'opération.

2. Mise en forme du texte brut homogénéisé en format (.txt)

Les fichiers en format Word sont à enregistrer en format (.txt) par les étapes suivantes :

1. Choisir « Enregistrer sous » dans le menu de Word.
2. Ajouter « H_ » dans le début du nom de fichier (H : pour document « Homogénéisé »).
   A titre d'exemple : H_Résumé_PROSPECTIVE_MER_MONTE_2019_40_14(.docx)
3. Choisir « texte brut » dans la liste des formats d'enregistrement (champs « type »).
4. Choisir « autre codage » parmi des codages de texte et puis sélectionner « Unicode (UTF-8) ».



3. Mise en forme des documents homogénéisés (.txt) pour CorTexT Manager

Un dossier compressé de tous les fichiers (.txt) est créé afin de charger le corpus de documents sur CorTexT Manager.

4. Mise en forme des documents homogénéisés (.txt) pour IRAMUTEQ

A partir des fichiers (.txt) utilisés pour CorTexT Manager, les documents (.txt) sont enregistrés pour IRAMUTEQ par les étapes suivantes :

1. Copier tous les fichiers (.txt) préparés pour CorTexT Manager dans le nouveau dossier dédié à l'IRAMUTEQ.

2. Ajouter « I_ » dans le nom de fichier (I : pour document préparé pour Iramuteq)
A titre d'exemple : H_I_Résumé_PROSPECTIVE_MER_MONTE_2019_40_14(.docx)
3. Insérer en début de chaque document dans le corps du texte l'intitulé suivant afin d'IRAMUTEQ puisse reconnaitre le début de chaque document :

**** *N_N°absolu *O_Operation *A_1234 *T_Titre

Il existe un format pour un couple variable et une modalité *Variable_Modalité qui vont être analysé par la Méthode Reinert (Classification simple sur les segments de texte). Attention il convient de mettre un espace avant chaque nouvelle étoile « * », et de séparer la variable de sa modalité par un tiret bas.

- **** : 4 stars pour initier la reconnaissance par le logiciel
- *N_N°absolu : Rang de réalisation de l'opération parmi toutes les opérations EPE
- *O_Opération : Type d'opération en 3 lettres, soit ESC (pour ESCo), soit ETU (pour ETUDE), soit PRO (pour PROSPECTIVE)
- *A_2014 : Année de réalisation du colloque
- *T_Titre : Titre abrégé

A titre d'exemple : **** *N_40 *O_PRO *A_2019 *T_LMM correspond à la 40e opération menée par la DEPE, il s'agit d'une prospective qui a été conclue en 2019, avec le titre « La mer monte ».

4. Vérifier qu'avant et après l'intitulé il y a bien un saut de ligne.
5. Concaténer des fichiers en 4 corpus afin de réaliser des analyses par type d'opération :

- corpus total
- sous-corpus ESCo
- sous-corpus études
- sous-corpus prospectives

Il est possible de concaténer automatiquement des fichiers en utilisant un protocole Python ci-joint (avec l'installation d'une plateforme de Python - Anaconda Navigator https://www.anaconda.com/distribution/)

Annexe 3. Éléments méthodologiques Iramuteq supplémentaires

1. Evolution des classes après l'enrichissement du dictionnaire

Après le lancement des analyses statistiques nous avons remarqué un grand nombre des formes nr correspondant aux formes absentes du dictionnaire d'Iramuteq qui ne peut reconnaitre la classe grammaticale. Il peut éventuellement s'agir de faute de frappe, mais c'est peu fréquent car les résumés ont été soumis à de nombreuses relectures.

Néanmoins, 2 200 formes non reconnus ont été identifiés pour le corpus total. Il existe deux cas, soit des formes ordinaires qui devraient être actualisées dans le dictionnaire selon le genre, le nombre ou la conjugaison, soit des acronymes spécifiques aux opérations. Comme les formes nr peuvent biaiser les résultats et induire un doute permanent quant à ces analyses, nous avons décidé d'enrichir le dictionnaire d'Iramuteq avec les formes ayant une fréquence supérieure à 3. Notamment nous avons lemmatisé 718 formes, soit 5 676 mots.

Puis nous avons regardé l'évolution des classes après l'enrichissement du dictionnaire. Les analyses ont été réalisées sur 4 corpus (ESCo, études, prospectives et Corpus total) avant et après l'enrichissement du dictionnaire. A titre d'exemple, voyons ci-dessous l'effet d'amélioration du dictionnaire d'Iramuteq pour le corpus total.

Ainsi, le score pour les segments classés par Iramuteq est amélioré de 10% pour le corpus total. Nous constatons également la réduction de nombre des classes de 4 classes à 3 classes (*Figure 34*). Nous faisons l'hypothèse que l'augmentation des segments classés a pour cause l'effet d'association. C'est-à-dire, la diversité du vocabulaire augmente et cela renforce la co-occurrence des mots-clés. Alors nous pouvons estimer l'augmentation de pertinence du nouveau dictionnaire.

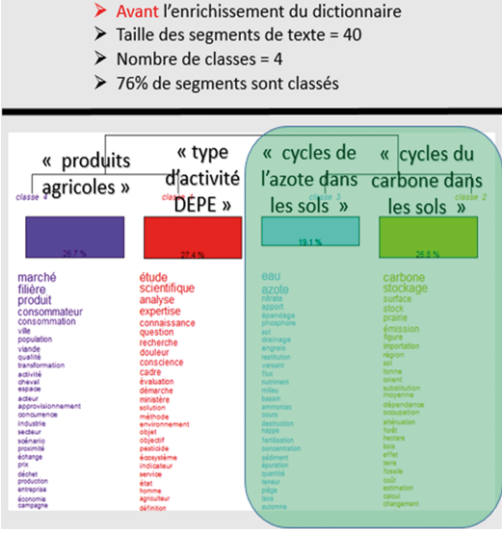


Figure 34a : Evolution des classes après l'enrichissement du dictionnaire. Méthode Reinert. Classification simple sur les segments de texte des noms par Iramuteq. Corpus total. Nombre d'itérations de 10, taille de séparation de segments à 40, 76% de segments classés avant l'enrichissement du dictionnaire respectivement

Figure 34b : Evolution des classes après l'enrichissement du dictionnaire. Méthode Reinert. Classification simple sur les segments de texte des noms par Iramuteq. Corpus total. Nombre d'itérations de 10, taille de séparation de segments à 40, 86% de segments classés après l'enrichissement du dictionnaire respectivement

Si nous comparons les deux classifications, avant et après l'enrichissement du dictionnaire pour le corpus total, nous notons que les classes sont plutôt conservées, elles ne sont pas très différenciées. Notamment, les classes dit « alimentation » et « type d'activité DEPE » ne bougent pas alors que les classes dites « azote dans les sols » et « carbone dans les sols » se regroupent dans une seule classe dite « carbone et azote dans les sols ».

2. Effet de la taille des segments de texte 20 / 40 / 60

Par ailleurs, nous avons testé l'évolution des classes avec des segments de texte soit plus longs, soit plus courts (à 20 ou à 60) que propose Iramuteq par défaut (à 40) pour voir si la classe dite « type d'activité DEPE » disparait ou non.

Sous Iramuteq les segments de texte sont construits à partir d'un critère de taille et de ponctuation. Iramuteq cherche le meilleur ratio entre la taille et la ponctuation. L'objectif est d'avoir des segments de tailles homogènes en respectant le plus possible la structure du langage.

Les analyses suivantes sont produites pour le corpus total avec l'occurrence en tant que mode de construction des segments de texte et la taille des segments de texte égale à 20 / 40 / 60 (*Figure 35*). Pour faciliter la dénomination des classes produites par Iramuteq nous avons ajouté des valeurs négatives dans l'analyse d'évolution de taille de segments (dit anti-classes ou des valeurs très peu spécifiques pour une classe).

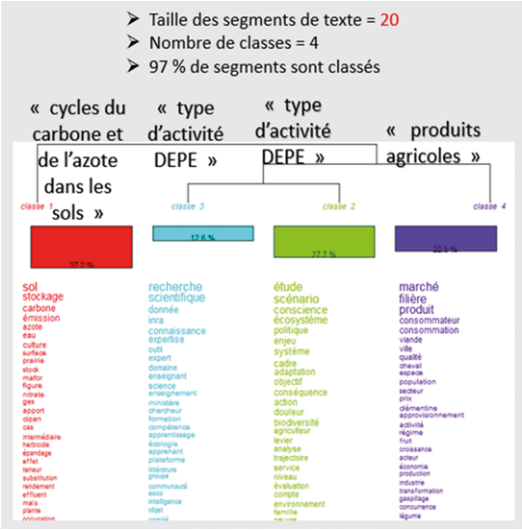


Figure 35a : Effet de la taille des segments de texte 20 / 40 / 60. Méthode Reinert. Classification simple sur les segments de texte des noms par Iramuteq. Corpus total. Nombre d'itérations de 10, taille de séparation de segments à 20, 97% de segments classés

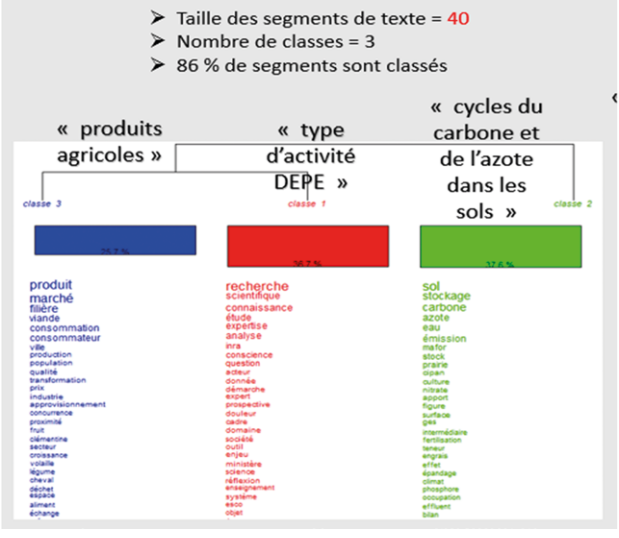


Figure 35b : Effet de la taille des segments de texte 20 / 40 / 60. Méthode Reinert. Classification simple sur les segments de texte des noms par Iramuteq. Corpus total. Nombre d'itérations de 10, taille de séparation de segments à 40, 86% de segments classés

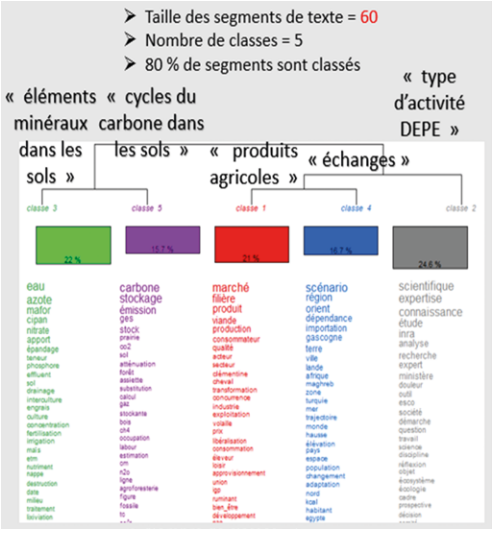


Figure 35c : Effet de la taille des segments de texte 20 / 40 / 60. Méthode Reinert. Classification simple sur les segments de texte des noms par Iramuteq. Corpus total. Nombre d'itération de 10, taille de séparation de segments à 60, 80% de segments classés

Nous constatons la réduction maximale du nombre de classes avec la taille des segments de texte égale à 40 ainsi que le meilleur score avec la taille des segments de texte égale à 20. Ainsi pour le minimum de classes il faut choisir la taille à 40 et pour le maximum de segments classés il faut choisir la taille à 20. En ce qui concerne les thématiques de classes, elles sont encore une fois plutôt conservées car les thématiques de classes ne changent pas beaucoup. Par ailleurs, grâce à l'analyse avec 3 tailles de séparation de segments 20 / 40 / 60 nous reconfirmons 3 grandes thématiques (des classes dites « cycles du carbone et de l'azote dans les sols », « type d'activité DEPE » et « produits agricoles »). Il y a des classes intermédiaires qui apparaissent ou disparaissent selon le réglage de taille de séparation de segments à 20 / 40 / 60 (par exemple, « type d'activité DEPE », « éléments minéraux dans les sols » et « échanges »). Par contre, la nature des classes ne change pas beaucoup, cela confirme donc la robustesse des résultats des analyses. Par exemple, pour la taille de séparation de segments à 60 la classe dite « cycles du carbone et de l'azote dans les sols » se distingue en « éléments minéraux dans les sols » et « cycles du carbone dans les sols ».

Annexe 4. Éléments méthodologiques CorTexT Manager supplémentaires

1. Evolution de nombre des clusters en fonction du nombre des expressions à extraire

Le nombre de clusters augmente toujours avec la hausse de nombre des expressions à extraire. Nous gagnons en précision en augmentant le nombre des expressions à extraire.
Pour les 75 expressions on a 7 clusters suivants : « Alimentation », « Eau / ressources », « Forêt / sol », « Numérique », « Pesticides », « Territoires ». Pour les 150 expressions on a 10 clusters qui représentent les mêmes thématiques que l'extraction terminologique de 75 expressions. Pour les 300 expressions on a 15 clusters qui ne se distinguent pas beaucoup des cartes à 75 ou 150 expressions. En revanche, il y a des nouvelles classes qui apparaissent comme « Bioénergie », « Diversification », « Environnement / biodiversité », « Environnement / eau », « Nexus ». Trois classes sont difficiles à nommer à cause du mélange de sujets.
Nous nous posons la question quel est le nombre des expressions optimal sur CorTexT. Qu'est-ce que nous gagnons d'avoir des expressions en plus ?

2. Le nombre des expressions optimal à extraire

Afin de choisir un nombre d'expressions optimal pour les analyses à venir nous avons fait plusieurs tests avec le nombre d'expressions à extraire différent pour le corpus total.
D'abord, nous avons récupéré des listes de TOP 75 / 150 / 300 expressions. Puis, nous avons extrait 1000 expressions afin de regarder la courbe de distribution des expressions par la fréquence et la loi de puissance (Scripts *List Builder*).
La définition du seuil dans la traîne de la courbe de distribution est assez empirique. Il n'existe pas une règle générale en lexicométrie, le seuil dépend beaucoup du corpus. Notamment, pour les opérations EPE on retrouve une grande hétérogénéité thématique sur un petit nombre de document : chaque opération a son propre vocabulaire.
Il faut donc trouver un compromis entre les critères suivants pour trouver le nombre optimal de expressions exportées:

- Toutes les 42 opérations EPE sont représentées dans l'extraction terminologique. Pour vérifier cela nous pouvons utiliser le script Corpus Explorer en cherchant s'il y a des lignes vides dans le champs « Terms » parmi 42 opérations.
- Le nombre de clusters produits reste interprétable en prenant en compte la lisibilité des cartes.
- La significativité des analyses produites : est-ce que des cartes représentent bien le coeur de l'activité DEPE ?

En prenant en considération les critères ci-dessus nous avons donc privilégié l'extraction terminologique des TOP 75 expressions.

3. L'inclusivité des listes de 75 / 150 / 300 expressions

Les listes de 75 / 150 / 300 expressions ne sont pas inclusives alors qu'elles sont extraites avec la même mesure de proximité. Parmi les TOP300 il y a 118 de TOP150 et 42 de TOP75. C'est à cause de plusieurs métriques à considérer par le logiciel pendant le calcul qui dépend de la mesure de pertinence choisie.
Si les expressions sont ordonnées par fréquence, les listes devraient être redondantes, dans les autres cas, c'est plus compliqué car les mesures de spécificité sont calculées relativement à un pool d'expressions candidats dont la taille dépend du choix de longueur de la liste.

4. Les monogrammes

Dans les paramétrages, il est possible de choisir le nombre de mots considérés dans l'expression (N-gram). Nous avons considéré la possibilité d'extraction terminologique avec les monogrammes et segment à 3 ou 4-gram.

CorTexT permet d'éliminer certaines expressions. Exclure des monogrammes est un problème car certains peuvent être significatifs. Par contre, l'intérêt des cartes excluant les monogrammes est la lisibilité, et aussi cela permet de faire apparaitre des thématiques par cluster.

Nous avons voulu extraire des monogrammes sur le corpus et les ajouter dans la liste des multi termes (script Term indexer) afin de ne pas perdre les expressions assez courantes dans le corpus. Ainsi, le multi terme « expertise scientifique collective » apparait parmi les expressions les plus partagées alors qu'il n'y a pas de mono terme « prospective » qui est un synonyme.

L'autre piste d'analyse est de comparer les mono termes issus d'Iramuteq et de CorTexT. D'une part, il faut vérifier que ces deux logiciels utilisent des méthodes de calcul identiques (mesure de proximité Chi^2 sur Iramuteq versus fréquence / Chi^2 sur CorTexT). D'autre part, il convient d'utiliser la même méthode que pour l'extraction des multi termes sur CorTexT afin de constituer la liste générale des expressions courantes.

Nous privilégions donc des multi-termes composés de 3 formes (N-gram de valeurs 3) pour l'analyse textuelle car il n'y a pas de critère pertinent de choix des monogrammes. En outre, la plupart des monogrammes ne semble pas significatif.

5. Le travail sur le résultat de l'extraction terminologique

Il est possible de retravailler les résultats de l'extraction terminologique en regroupant des expressions sémantiquement proches ou synonymes : cela permet de les faire monter dans les occurrences, et donc de les exclure des expressions au-dessous du seuil de sélection.

Il existe donc deux possibilités pour retravailler le vocabulaire sur le résultat de l'extraction terminologique pour un nombre des expressions optimal :

- Supprimer une ligne présentant une expression ou une liste d'expressions non pertinentes ;
- Regrouper plusieurs lignes pour rapprocher par exemple des concepts ou des notions proches.

Par contre, cela peut être un faux bruit, il est donc mieux de laisser le vocabulaire sans le rectifier sauf le vocabulaire basique à nettoyer ou des cas à caractère technique à supprimer qui n'appartiennent pas aux trois domaines Inra (par ex. « autre part », « mise en oeuvre », « mise en place ») ou des cas d'expressions sémantiquement liées à regrouper (par ex. « acteurs de la filière », « acteurs des filières »).

6. Mesures à choisir sur CorTexT lors le script Network Mapping

Il est possible de choisir plusieurs mesures de proximité (*Proximity Measure*) pour lancer le script Network Mapping:

- *Raw* : mesure de co-occurrence brute (fréquences réelles des co-occurrences). Elle calcule la proximité brute entre deux entités et repose sur l'hypothèse qu'un lien correspond à une interaction effective. Elle correspond tout simplement au nombre d'occurrences ou de co-occurrences réel.

Au centre d'un graphe, les expressions très fréquentes ; en périphérie, les moins fréquentes. Raw montre les valeurs fortes ou des évidences et elle peut diminuer le caractère informatif de la carte dans le cas où une valeur est présente dans tous les documents donc cette valeur sera aussi omniprésente dans la carte (« l'effet de taille »).

- *Chi2* : mesure de spécificité. Elle mesure l'intensité du lien entre deux expressions en appréciant l'écart par rapport à la valeur attendue. C'est un calcul avancé pour la distance permettant à la fois de mieux comprendre les valeurs spécifiques d'une variable et de permettre de diminuer « l'effet de taille ».
- *Distributional* : mesure de nombre de documents. Elle mesure la similarité des contextes d'apparition de ces expressions. Cela permet de détecter des synonymes, c'est-à-dire des expressions qui ne cooccurrent pas forcement mais qui ont des contextes d'apparition identiques.
- *dot_product_het* : mesure de similarité. Elle permet de trouver les documents les plus similaires dans le réseau hétérogène. Cette mesure peut être utile pour l'analyse des opérations.

7. La projection de la 3e variable

Nous avons retrouvé plusieurs fois le problème de projection des intitulés d'opérations EPE qui ne correspondaient pas aux mots-clés dans des clusters.

L'ordonnance est faite par apport de la mesure choisie : *raw* (fréquence) ou Chi^2 (les expressions les plus associés où le nombre de liens est le plus élevé). Néanmoins, le calcul pour projeter la 3e variable est assez sophistiqué sur CorTexT, il dépend:

- de la centralité du mot dans un cluster
- de la fréquence d'usage d'un mot dans un projet
- de la dispersion des expressions de chaque cluster
- de la fréquence des dits mots, etc.

Par contre, dans notre cas avec peu de documents les étiquettes de la 3e variable sont les mêmes pour un calcul du lien Chi^2 ou *raw* avec le champ « Titre » car la distribution statistique est assez faible et pas du tout suffisante pour faire jouer les mesures (le nombre de valeurs est égal au nombre de documents). Par contre, en changeant le champ à « Type » (3 valeurs catégoriales : ESCo, études, prospectives) les étiquettes ne sont plus les mêmes.

Finalement, nous avons extrait tous les intitulés des opérations EPE qui ont contribué au cluster et nous avons retravaillé des cartes *via* l'application de design *Inkscape*[17] afin d'avoir tous les titres d'opérations EPE à la fois et pas seulement les TOP 3.

8. Evolution des thématiques dans le temps

Sous CorTexT, la dimension temporelle ne peut pas être analysée pour les résumés. L'année d'une opération est attribuée selon la date de la fin d'opération et notamment de la sortie du rapport. Nous avons donc calculé le nombre d'opérations EPE par an vu que leur nombre est assez faible (42 projets). Ainsi le nombre d'opérations EPE par an varie très fortement d'une opération jusqu'à 8 opérations EPE en 2017.

D'abord, nous avons produit des cartes temporelles des 75 / 150 / 300 expressions les plus fréquentes sur le corpus total avec le découpage en 3 et 4 périodes. Une constante « agronomique » dominante ressort, elle traverse toutes les opérations EPE depuis 20 ans.

Par ailleurs, nous avons mis en question la signification des couleurs sur l'évolution temporelle (Script Network *Mapping, Dynamic*). Notamment, pour le découpage en 3 périodes la notion de stockage additionnel de carbone est présente en vert (1 période) et en bleu vert (3 période). Pareil pour le

17 https://inkscape.org/release/inkscape-1.1/

découpage en 4 périodes le stockage additionnel de carbone est présent en bleu (1 période) et en vert (4 période). Le script *Network Mapping* ne relie pas ces expressions car il y a discontinuité temporelle : le cluster disparaît sur une période intermédiaire, et quand il réapparait ensuite il n'est pas identifié comme étant le même cluster mais comme un nouveau cluster.

9. La méthode Iramuteq versus la méthode CorTexT Manager

Nous voyons bien deux approches totalement différentes qui sont à la base d'Iramuteq et de CorTexT. Iramuteq est basé sur le principe d'opposition alors que CorTexT est basé sur la similitude des mots. CorTexT cherche des mots assez équivalents ou subsidiaires. C'est probablement pour cette raison que nous retrouvons 2 grandes classes parmi 3 produites par Iramuteq dans CorTexT (dites « cycles du carbone et de l'azote dans les sols » et « produits agricoles ») ; en revanche, la classe dite « type d'activité DEPE » n'apparaît pas et des nouvelles classes apparaissent (dites « pesticides », « forêt », « numérique », etc.).

10. Les limites

- Pendant le travail d'interprétation il est important de ne pas se limiter aux points déjà connus sur les opérations EPE mais de les compléter également par les découvertes et étonnements apportées par l'analyse textuelle.
- Nous ne pouvons pas voir, sur CorTexT, les expressions les moins citées, le logiciel permet de voir seulement les mots-clés les plus cités. Par contre, sur Iramuteq il est possible d'analyser les hapax : une expression n'apparaissant qu'une seule fois dans le corpus.
- Par ailleurs, nous remarquons quelques soucis de résultats statistiques selon le choix de réglage entre sélection de l'extraction à la phrase ou au document (fichier multiterms_statistics_expanded.csv dans le script *Terms Extraction*). Notamment pour l'expression « impacts environnementaux » : avec l'extraction par fréquence au niveau du document, C-value est égale à 16 alors que le fichier statistique indique la présence de cette expression dans 17 documents. Avec l'extraction par fréquence au niveau de la phrase, C-value est égale à 43 alors que la fréquence selon les statistiques est de 44. Mieux vaut donc se référer aux fichiers livrés dans le répertoire « indexed list ».
- Nous avons reproduit une carte de co-occurrence des expressions en enlevant la projection de la 3e variable (*Network mapping*). Par contre, les clusters construits par CorTexT ne sont pas les même sur les deux cartes. Les expressions ont été extraites de la même façon : TOP150 à l'échelle de la phrase et par Chi^2.

  Ci-dessous des liens pour ces scripts-là :
  https://managerv2.cortext.net/logs/158788
  https://managerv2.cortext.net/logs/172652
- Dans le champs Projection_ Cluster_x_x deux expressions sont présents avec le poids le plus élevé pour un cluster. Par contre, sur la carte une taille des noeuds ne correspond pas à ces poids. La taille des noeuds est proportionnelle à leur mesure de degré à savoir leur occurrence ; les tailles de liens sont proportionnelles aux nombres de liens ; la taille des clusters est proportionnelle par sa surface au nombre de documents qui contribuent à créer le cluster

11. Les suites

- L'analyse des synthèses va permettre d'apparaître les différences de type d'écriture et mettre en évidence des types de productions des connaissances par Institution.
- Il pourrait être envisagé de situer notre corpus dans le panorama des connaissances, notamment positionner les opérations EPE par apport à Wikipédia, la base de connaissances avec des catégories déjà définies, ou une orientation des déclarations scientifiques. A ce stade-là nous avons envisagé le travail de positionnement des opérations EPE par apport aux documents d'orientations INRAE.
- Nous pouvons repartir le Corpus Total en sous-corpus selon les catégories de thématiques abordées quel que soit le type d'opération (carbone, pesticides, etc.). L'intérêt de choisir un thème est de voir comment les questions internes à ce sujet évoluent.
- Il possible de voir l'évolution de contexte autour des expressions les plus fréquents ou les plus partagés (le script *Distinct reading*), cela permet d'étudier des connaissances locales des expressions significatives de notre corpus.

Centre-siège Paris-Antony
Direction de l'expertise scientifique collective,
de la prospective et des études
147 rue de l'Université – 75338 Paris cedex 07
Tél. +33 1 (0)1 42 75 94 90

Rejoignez-nous sur :

Inrae.fr/collaborer/expertise-appui-aux-politiques-publiques

Institut national de recherche pour l'agriculture, l'alimentation et l'environnement

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE