Élaboration d'un référentiel terminologique au sein d'un réseau de recherche par l'analyse textuelle de données ouvertes

Dominique Desbois

Élaboration d'un référentiel terminologique au sein d'un réseau de recherche par l'analyse textuelle de données ouvertes

Dominique Desbois [1]

Résumé. L'objectif de cette contribution est d'illustrer l'élaboration d'un référentiel terminologique au sein d'un projet *Eranet* de recherche en réseau pour l'élaboration d'un consensus sur l'impact des services écosystémiques pour le développement durable. Cette élaboration est menée au moyen d'une analyse textuelle de correspondance simple de données ouvertes constituées par des descriptions de tâches au sein des ateliers du projet, grâce au logiciel libre *Dtm-Vic*. Les résultats de l'analyse textuelle des rapports de recherche montrent une distinction entre les termes techniques, de préférence utilisés pour décrire l'organisation du contexte opérationnel, et les termes conceptuels plus impliqués dans la description des résultats scientifiques obtenus.

Mots clés : terminologie, analyse textuelle, rapports de recherche, analyse de correspondance simple

La problématique d'élaboration d'un consensus

Depuis le sixième programme cadre de recherche et développement (PCRD), une part significative des recherches financées au plan européenne est désormais conduite en mode réseau au sein de programmes spécifiques appelés *Eranet* (*European Research Area Network*) d'après leur mode de financement. Ces projets permettent de mener des actions multilatérales de recherche et de développement réunissant plusieurs partenaires (5 au minimum) provenant de différents pays de l'espace européen de la recherche (au moins 5 pays différents). Dans le cadre du programme européen *Horizon 2020* (*H2020*), ces réseaux de la recherche ont vocation à capitaliser leurs expériences sur la base d'un mécanisme de pré-sélection (*Eranet-Plus*) initié dans le contexte du 7e PCRD leur permettant de se coordonner (*Eranet Cofund*) pour participer à des initiatives de programmation conjointe (*Joint Programmation Initiative - JPI*).

L'élaboration d'un consensus (*Consensus Building*) permettant de partager enjeux et résultats constitue donc une étape incontournable du cycle d"interactions et d'échanges au sein de la communauté de Recherche et Développement (*R&D*), non seulement entre scientifiques mais également avec les porteurs d'enjeux du projet. Cette recherche de consensus est indispensable à la capitalisation des expériences et constitue un levier facilitant

[1] *UMR Economie publique, Inra-AgroParisTech, Université Paris-Saclay, 16 rue Claude Bernard, F-75231 Paris Cedex 05, France ; dominique.desbois@inra.fr*

la coordination d'actions pour la diffusion des résultats, en particulier en direction de porteurs d'intérêts et la programmation de projets ultérieurs.

Le projet *Towards Rural Synergies and Trade-offs between Economic Development and Ecosystem Services* (*Trustee*) de l'*Eranet Ruragri* a inscrit parmi ses objectifs la recherche d'un consensus sur les scénarios durables de développement rural. La recherche de consensus porte sur les possibilités de fournir en Europe des biens publics et des services écosystémiques en même temps que de produire des biens privés agricoles et forestiers. Cette recherche de consensus suppose de solliciter l'opinion des parties prenantes sur les futurs scénarios de développement rural, en mettant l'accent sur la conception de scénarios durables impliquant l'utilisation des terres et la fourniture de services écosystémiques, ainsi que la mise en œuvre des instruments politiques correspondants.

Le comité de pilotage du projet *Trustee* a décidé d'initier l'activité de la tâche *Consensus Building* de l'atelier *Validation, training, and consensus building* par la recherche d'un référentiel terminologique (Rey-Debove,1998)[2] partageable par les participants au projet pour être utilisé dans la communication avec les porteurs d'intérêts. En effet, un tel référentiel terminologique pourra être réutilisé ultérieurement dans le cadre du projet H2020 *Pegasus* pour animer une conférence de consensus avec les porteurs d'intérêts professionnels et les usagers concernant les synergies à établir entre services écosystémiques et développement rural. Les rapports de résultats issus de projets de recherche européens ou nationaux se prêtent comme corpus à de telles problématiques de par leur caractère public et leur structure relativement normalisée.

Élaboration d'un référentiel terminologique et d'une carte conceptuelle

Matériel

Selon une norme *de facto* partagée par les projets de recherche européens, le projet Trustee est organisé en sept ateliers distincts (*WP - Working Package*), subdivisés en un nombre variable de tâches (*T – Task*) pour chaque atelier. La liste complète de ces ateliers et de ces tâches ainsi que leurs descriptifs sont consultables sur le site du projet *Trustee*[3]. À la fin du projet *Trustee* (mars 2017), l'ensemble des rapports de tâches a été collecté pour former la base du rapport d'activité du projet. Chaque rapport de tâche est structuré en quatre sections différentes selon un format standardisé. Pour des raisons de pédagogie de l'exposé, nous nous limitons à l'analyse de la section principale intitulée « *Work done and results obtained during the project* ». Ainsi, un ensemble de 21 textes décrivant les résultats de recherche obtenus au sein des tâches *Trustee*, a été soumis à l'analyse textuelle afin d'établir un

[2] Défini comme « *Ensemble de tout ce dont un locuteur peut parler dans une langue donnée (objets réels ou imaginaires, concrets ou abstraits, appelés référents)* ». Dans le contexte de recherche propre au projet Eranet *Trustee*, le référentiel terminologique est constitué par l'ensemble des termes lexicaux mobilisés au sein du projet *Trustee* par les différentes tâches du projet et les problématiques qui s'y rattachent.

[3] Pour une description de ce projet de recherche, cf. site du projet : https://www.trustee-project.eu/.

référentiel lexicographique du projet pour la communication avec un public de porteurs d'intérêt dans des échanges portant sur les liens entre développement rural et services écosystémiques.

Méthodes

Les analyses effectuées pour établir un référentiel terminologique à partir des correspondances lexicographiques sont basées sur le cadre méthodologique principalement développé par Lebart et Salem (1994) depuis les années 1970 sous l'égide de l'analyse des correspondances et des procédures associées (aides à l'interprétation, classification automatique) applicables aux données textuelles. Cette analyse automatique du discours (Pêcheux et al., 1982) confère une place décisive au lexique, approche justifiée dans la mesure où le discours scientifique est relativement normalisé et peu ambigu (Rinck, 2010). Derrière les développements des outils lexicométriques catégorisés par Jenny (1997), se profilent les deux paradigmes suivants : d'abord, celui de l'analyse du discours (Harris, 1969) ; d'autre part, celui de la linguistique textuelle (Benveniste 1966, 1974). Parmi ces outils, celui proposé précédemment par Lebart et al. (1998), basé sur des extensions textuelles de l'analyse de correspondance simple (ACS), de l'analyse des correspondances multiples (ACM) et des méthodes de classification automatique associées (Lerman, 1981), est adapté à l'analyse lexicale des morphèmes (graphie des mots du langage) pour établir un référentiel lexicographique, en raison des propriétés distributionnelles de la métrique du Khi-deux. Suivant cette démarche, nous utilisons la séquence d'analyses textuelles fournies par le logiciel *Dtm-Vic* (Lebart et Piron, 2012), après un certain prétraitement des textes. La séquence d'analyse textuelle – *VisuText* – au sein du logiciel *Dtm-Vic*[4] retenue pour cet exposé introductif a pour objectif d'effectuer une ACS basée sur une table lexicale obtenue après un filtrage des mots basé sur un seuil minimal d'occurrences.

Résultats

L'analyse textuelle de la correspondance des tâches repose sur la construction d'une table de correspondance lexicale entre les mots et les textes, produite par le dénombrement des mots au sein des textes décrivant les tâches et leurs résultats. En premier lieu, les contributions des mots aux axes principaux de l'ACS (cf. Figure 1) sont interprétées afin de construire un référentiel terminologique commun. Complémentairement, les projections graphiques des tâches de recherche du projet *Trustee* (cf. Figure 2) sont analysées afin de les relier aux thématiques et aux problématiques du projet.

[4] *Data and Text Mining: Visualisation, Inférence, Classification*, version 6.0, cf. http://www.dtmvic.com. Ce logiciel libre d'usage peut être implanté sous le système d'exploitation Windows.

Dominique Desbois

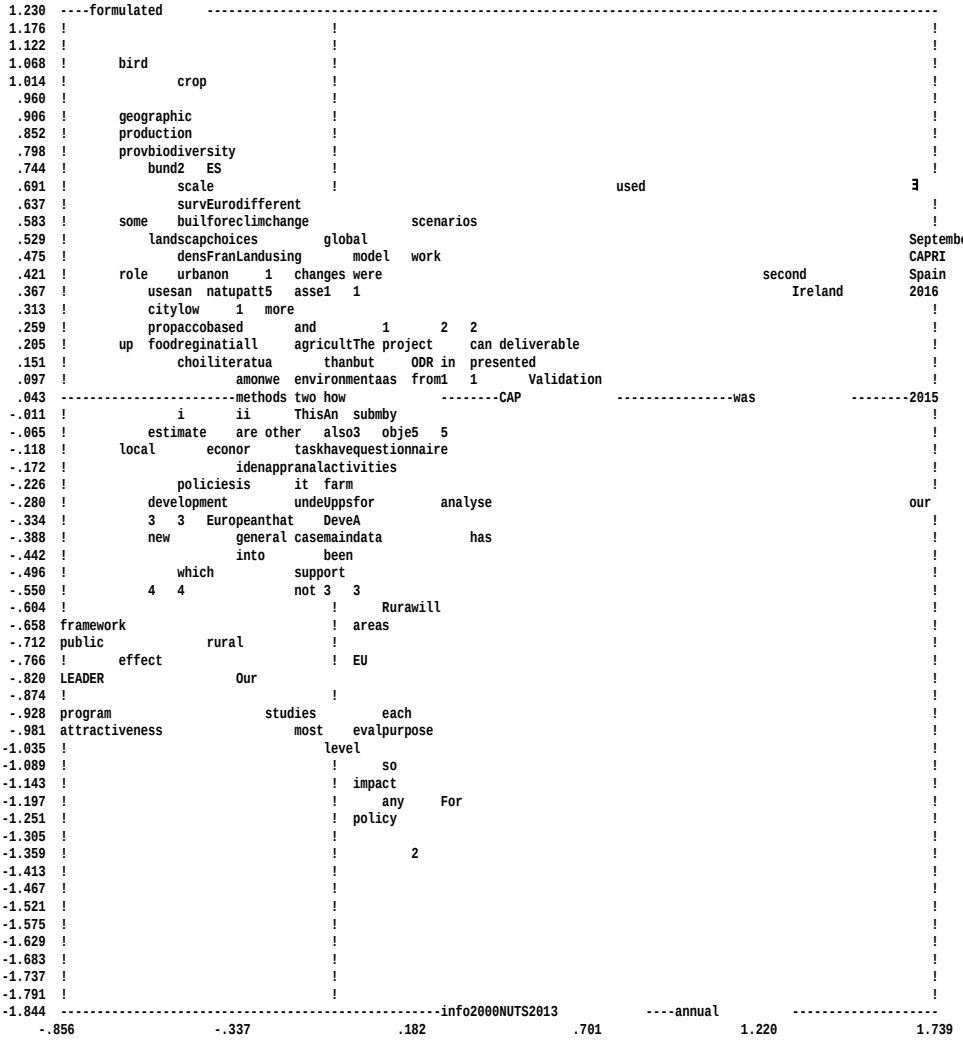
L'interprétation des dimensions du référentiel lexicographique

Espace des mots : la dimension organisationnelle et thématique du référentiel

L'orientation positive ([*0<F1*[) du premier axe factoriel (*F1*) de l'ACS concerne d'une part l'organisation opérationnelle du projet, avec des termes décrivant son organisation (*work*, *training*, *workshop*, *terms*), les périodes (*2015*, *2016*, *September*), la plate-forme commune de l'Observatoire du Développement Rural (*ODR*), les mots-outils adaptés à cette description (*for*, *from*, *during*, *other*, *second*, *following*), l'utilisation de participes passés (*developed*, *collected*, *used*) et du prétérit (*was*), d'autre part sa dimension évaluative avec des termes référentiels tels que *validation* et *baseline*, son objet (*CAP*) – la Politique agricole commune, le modèle d'évaluation d'impact utilisé (*CAPR*I) et certains des pays choisis pour les études de cas (*Spain*, *Ireland*). À l'opposé, l'orientation négative du premier axe factoriel de l'ACS]*F1<0*] concerne le contenu des tâches du projet avec des termes spécifiques plus orientés vers la description et la partage des objets scientifiques tels que *attractiveness*, *bird*, *buildings*, *bundles*, *city*, *cover*, *development*, *environmental*, *food*, *forest*, *governance*, *framework*, *public*, *services*, *sprawl*, *survey*, *territorial*, et certains mots-outils associés au discours programmatique et évaluatif tels que *choices*, *effect*, *identified*, *high*, *low*, *new* et *than*. Cette dimension peut être interprétée comme se référant à l'organisation générale et à l'articulation des thématiques scientifiques du projet.

Figure 1 page suivante

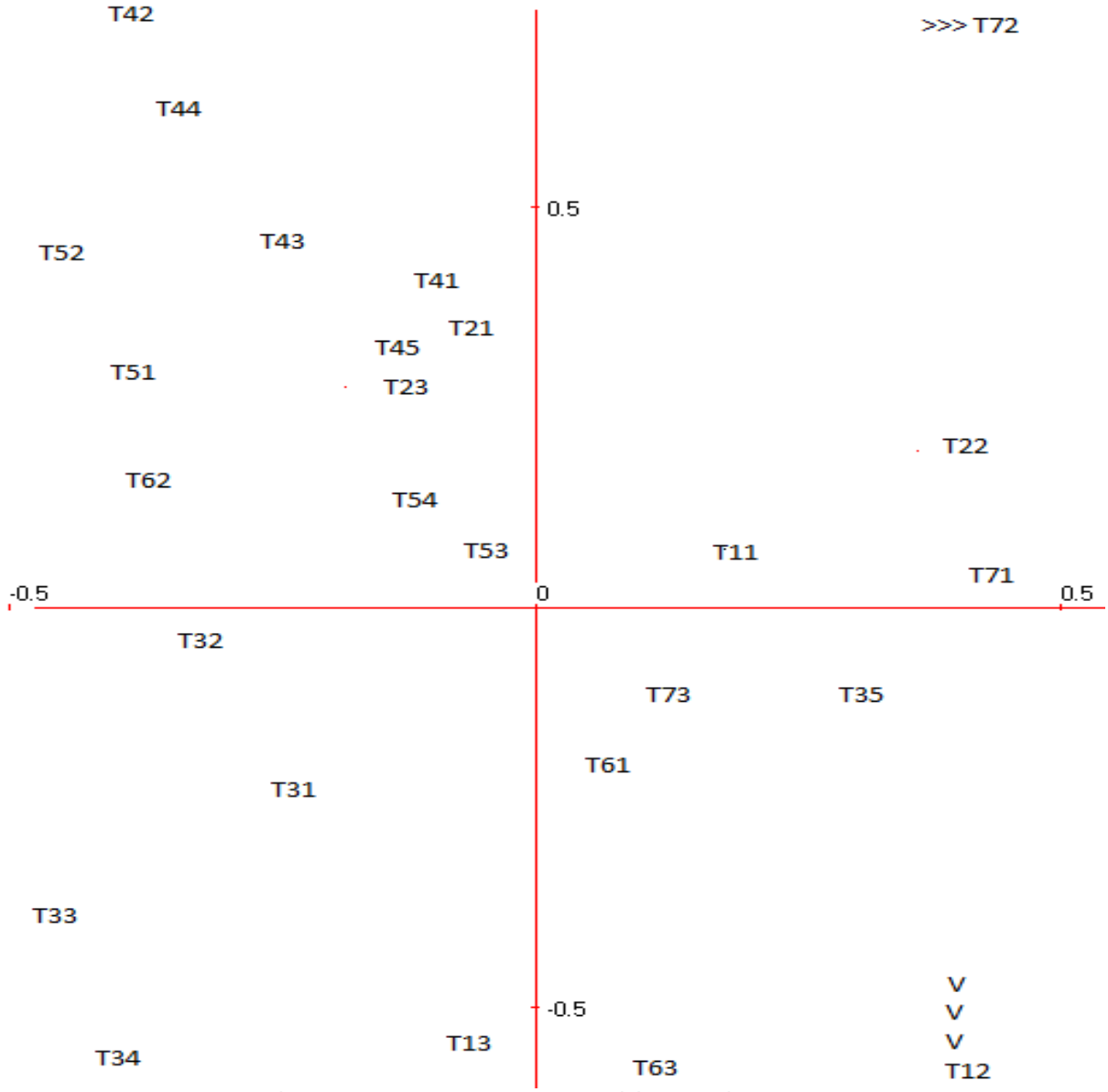
Figure 1 : les mots de la correspondance textuelle, plan factoriel F1 x F2.



Source : traitements de l'auteur d'après les rapports de tâche Trustee (2017).

Figure 2 page suivante

Figure 2 : les tâches de la correspondance textuelle, plan factoriel F1 x F2.



Lecture : les tâches 1.2 (T12, ordonnée = -2) et 7.2 (T72, abscisse = 3) ont été relocalisées (>>>) aux bords du cadre.
Source : traitements de l'auteur d'après les rapports de tâche Trustee (2017).

Espace des mots : la dimension scientifique et technique du référentiel

L'orientation positive [*0<F2*[du second axe factoriel (F2) de l'ACS est déterminée par des mots tels que *information*, *land* ou *scale* qui relèvent de catégories techniques manipulées par le projet *Trustee*.

L'orientation négative]*F2<0*] du second axe factoriel concerne les données (*data*) et les dates de la période de référence pour les études entreprises (*2000*, *2013*), le référentiel européen utilisé pour les unités territoriales (*NUTS*), sa périodisation temporelle (*annual*), sa délimitation spatiale (*areas*), et plus généralement son vocabulaire technique : *crop*, *ecosystem*, *impact*, *funds*, *level*, *measures*, *payments*, *policy*, *production*, *provision*, *purpose*, *rural* et *services*, avec certains mots-outils associés tels que *for*, *on* et *per*.

Cette dimension peut être comprise comme spécifique aux catégories et références techniques mobilisées par un projet d'évaluation des services écosystémiques principalement centré sur la distribution territoriale de l'usage des terres (*land use*).

Les contributions des tâches aux dimensions du référentiel terminologique

Espace des tâches : l'organisation opérationnelle et la description technique

Dans le premier quadrant [*0<F1 et F2> 0*], nous trouvons {*T72*} - tâche 7.2 *Training for understanding and application of used and developed tools*, comme contributeur principal de l'axe F1, associée à {T71} - la tâche 7.1 *Validation of the baseline*, {*T22*} - tâche 2.2 *Models for the analysis of EU land changes*, et finalement {*T11*} - tâche 1.1 *Organisation of an integrated database system as a project backbone*. En effet, ces tâches sont liées dans l'organisation du projet : les données alimentent l'utilisation du modèle pour la validation et l'application. En référence à la précédente section concernant l'espace des mots du discours, ces tâches partagent un profil lexicographique basé sur des termes conceptuels généraux d'organisation opérationnelle identifiés par leurs projections positives selon la première dimension et sur les catégories techniques du projet projetées positivement selon la seconde dimension. Dans le second quadrant [*0<F1 et F2<0*], nous trouvons {*T12*} - tâche 1.2 *Stocktaking of policy measures that impact rural areas*, comme contributeur principal à l'axe F2, accompagnée par {*T63*} - tâche 6.3 *Evaluation of results and adaptation of the Regional Development Policies (RDPs)*. Ces tâches partagent un profil lexicographique de description opérationnelle et de termes techniques appartenant au langage spécifique de l'évaluation d'impact des politiques publiques.

Notons que toutes les tâches {*T7**} de l'atelier *Validation, training, and consensus building* (*WP7*) appartiennent au demi-plan [*0<F1*[, orienté vers l'organisation opérationnelle et la description technique du projet.

Espace des tâches : les services écosystémiques et les enjeux de développement rural

Dans le troisième quadrant [*0>F1 et F2 <0*], nous trouvons {*T34*} - tâche 3.4 *Public policies of rural development: Ex-post evaluation*, {*T33*} - tâche 3.3 *Governance of policy at a local level* et {*T13*} - tâche 1.3 *Case study choice*. Les autres tâches associées à ce quadrant sont plus proches de l'origine et donc un peu moins spécifiques : {*T31*} - tâche 3.1 *Analyse the driving factors that underlay local economic performance* et {*T32*} - tâche 3.2 *Best local practices that favor economic performance*. Par leurs profils de mots utilisés dans leur descriptif de résultats, ces tâches partagent le langage commun aux enjeux scientifiques du projet exprimé selon le demi-axe]*F1<0*] et les références techniques qui contribuent au demi-axe [*0<F2*[. Notons que presque toutes les tâches {*T3**} de l'atelier sur les politiques de développement rural et de cohésion (*WP3*) appartiennent à ce quadrant, à l'exception de la tâche 3.5 qui appartient au second quadrant. Ce sont, notamment, les tâches 3.3 *Governance of policy at a local level* et 3.4 *Public policies of rural development: Ex-post evaluation* qui contribuent le plus au demi-axe]*F2<0*] avec des références techniques spécifiques au domaine du développement rural.

Dans le quatrième quadrant [*0<F1 et F2 <0*], nous trouvons {*T42*} - tâche 4.2 *Causal factors of ecosystem service bundles* et {T44} - tâche 4.4 *Trade-offs and synergies among ecosystem services*, avec {*T52*} - tâche 5.2 *Can agricultural practices save ecosystem services?*, {*T43*} - tâche 4.3 *Common bird communities in areas with desirable and undesirable bundles of ecosystem services*, {*T41*} - tâche 4.1 *Mapping of ecosystem services and their interactions*, {*T21*} - tâche 2.1 *Global drivers of EU land use*, {*T45*} - tâche 4.5 *EU policies and ecosystem services change over time*, {*T51*} - tâche 5.1 *Urban sprawl and ecosystem service bundles*, et {*T23*} - tâche 2.3 *The rural-urban dynamics of land use within the EU* dans une position intermédiaire avec {*T62*} - tâche 6.2 *Urban and rural exchanges: Policy measures to favor the provision of ecosystem services*. [*T54*] - tâche 5.4 *Economic development and ecosystem service bundles*, et {*T53*} - tâche 5.3 *Regulating services and agriculture vs urban sprawl* sont trop faiblement corrélées pour être considérées comme des contributeurs spécifiques. Notons que toute les tâches de l'atelier *Composition and distribution of ecosystem services sets* » (*WP4*) et de l'atelier *Economic development and ecosystem services: some specific links services* (*WP5*) appartiennent à ce quadrant. En particulier, {*T44*} - tâche 4.4 *Trade-offs and synergies among ecosystem services* et {*T42*} - tâche 4.2 *Causal factors of ecosystem service bundles* contribuent pour la plus grande part au demi-axe]*F2> 0*], grâce à leurs profils de mots liés aux concepts généraux des services écosystémiques.

Notons que le semi-plan]*F1> 0*], constitué par des contributions moins dispersées en mots et en tâches, utilise le vocabulaire scientifique du projet, soit en termes de concepts généraux, soit en termes de références opérationnelles. En contraste, le semi-plan [*0<F1*[, constitué par des contributions plus dispersées, est consacré aux questions opérationnelles et techniques avec un ensemble plus restreint de vocabulaire standard, à la fois organisationnel et technique, dédié aux tâches situées à la périphérie du projet (tâches instrumentales principalement couvertes par les ateliers *WP1* et *WP7*).

Conclusion

L'exposé partiel car introductif de notre démarche d'analyse textuelle des rapports de recherche pour l'établissement d'un référentiel terminologique illustre les possibilités de mobilisation des données textuelles ouvertes pour l'organisation de conférences de consensus entre experts scientifiques et porteurs d'intérêts professionnels.

Ainsi, nous fournissons un référentiel terminologique des différentes réalisations au sein du réseau *Trustee* à travers l'analyse simple de correspondance textuelle des données ouvertes que sont les rapports de résultats pour les différentes tâches de ce projet *Eranet* centré sur la pertinence des services écosystémiques pour le développement rural. Ce référentiel terminologique documente les différents profils de vocabulaire utilisés entre, d'une part, les tâches opérationnelles et techniques, et d'autre part, les tâches dont l'objet est plus directement centré sur les problématiques scientifiques soulevées par l'évaluation d'impact des services écosystémiques. En correspondance avec cette dichotomie entre tâches opérationnelles et tâches scientifiques, nous mettons en évidence une ordination des termes utilisés dans les rapports d'activité *Trustee* allant des termes techniques, de

préférence utilisés pour décrire l'organisation du contexte opérationnel, aux termes conceptuels plus impliqués dans la description des résultats scientifiques obtenus.

Nous proposons d'étendre cette approche par l'analyse des correspondances multiples et les méthodes associées de classification automatique ou de discrimination à l'ensemble des productions écrites du projet *Trustee*, rapports et publications, afin de pouvoir rendre plus robuste et affiner ce référentiel terminologique avant son utilisation pour la communication et l'interaction avec des porteurs d'intérêts professionnels dans le cadre de futurs projets de développement rural.

Références bibliographiques

Benveniste E (1966, 1974*) Problèmes de linguistique générale : vol. I (1966), vol. II (1974)*. Gallimard, Paris.

Harris ZS (1969) Analyse du discours. *Langages 4(13*) : 8-45.

Jenny J (1997) Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. État des lieux et essai de classification. *Bull Méthodol Sociol (BMS) 54* : 64-112.

Lebart L, Piron M (2012) *Pratique de l'analyse des données numériques et textuelles avec Dtm-Vic.* L2C, Paris, 210 p.

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

Lebart L, Salem A, Berry L (1998) *Exploring Textual Data*. Kluwer, Boston.

Lerman I C (1981) *Classification et analyse ordinale des données*. Dunod, Paris.

Pêcheux M, Léon J, Bonnafous S, Marandin JM (1982) Présentation de l'analyse automatique du discours (AAD69) : théories, procédures, résultats, perspectives. *Mots 4* : 95-123.

Rey-Debove J (1998) *La Linguistique du signe. Une approche sémiotique du langage*. Armand Colin.

Rinck F (2010) L'analyse linguistique des enjeux de connaissance dans le discours scientifique. Un état des lieux. *Rev Anthropol Connaiss 4(3)* : 427-450.